

CENTER FOR APPLIED OPTIMIZATION  
Modeling and Computation for Engineering, Science and Industry  
University of Florida, Gainesville FL 32611  
<http://www.ise.ufl.edu/cao/>

*Engineering Address:*  
371 Weil Hall  
Phone: 352-392-9959  
Fax: 352-392-3537  
Email: [center@cao.ise.ufl.edu](mailto:center@cao.ise.ufl.edu)

*Mathematics Address:*  
201 Walker Hall  
Phone: 352-392-0286  
Fax: 352-392-6254  
Email: [center@math.ufl.edu](mailto:center@math.ufl.edu)

The Center for Applied Optimization at the University of Florida is an interdisciplinary center which encourages joint research and applied projects among faculty from engineering, mathematics and business. It also encourages increased awareness of the rapidly growing field of optimization through publications, conferences, joint research and student exchange. It was founded in September 1992. The co-directors are Dr. Donald Hearn and Dr. Panos Pardalos of Industrial and Systems Engineering and Dr. William Hager, from the Mathematics Department.

Center affiliates include several faculty from ISE, Civil Engineering, Aerospace Engineering, Chemical Engineering, Computer and Information Science, Food and Research Economics, Mechanics & Engineering Science, Electrical Engineering, Mathematics, and Decision and Information Sciences.

Individual and joint research includes global and discrete optimization, acceleration of decomposition methods, new dynamic programming techniques for lot sizing models, network optimization methods, optimal control problems, optimization of elastic materials, inverse problems, and multi-criteria optimization. Current applied projects include randomization in algorithm design, multidimensional assignment problems, molecular conformation problems, supply chain and e-commerce, telecommunications, financial engineering, modeling and solutions of water management problems, evacuation modeling, and warehouse location problems. Sponsors include the National Science Foundation, AirForce, the Army Research Office and Florida Water Management Districts.

The Center is interested in promoting collaboration with researchers at other universities through visitors and student exchange. It administers a program for visiting students from the Royal Institute of Technology (KTH), Stockholm.

The Center for Applied Optimization has hosted several recent conferences on "*Large-scale optimization: State of the art*," "*Network optimization*," "*Optimal Control: Theory, Algorithms, and Applications*," "*Complexity and Approximation*," "*Stochastic Optimization: Algorithms and Applications*," "*Biocomputing*," and "*Cooperative Control and Optimization*". In addition, the Center has hosted several international visitors from Brazil, China, Italy, India, Japan, Korea, Ukraine, and Sweden.

CONFERENCE PROGRAM FOR FEBRUARY 16, 2004 MONDAY (282 Reitz Union)		
8:30	Registration	
9:00 - 9:15	<b>Welcome Session</b>	
	<b>Coordinator:</b> Panos Pardalos	
	Win Phillips	University of Florida Vice President for Research
	Don Hearn	Chair, ISE Dept., University of Florida
	William Ditto	Chair, Biomedical Engineering Dept., University of Florida
9:15 - 10:15	<b>Plenary Speaker:</b> Lotfi A. Zadeh, University of California, Berkeley Computing with Words and its Application to Deduction, Definition and Search	
	<b>Coordinator:</b> Panos Pardalos	
10:15 - 10:30	<b>Coffee Break</b>	
10:30 - 12:00	<b>Session M1</b>	
	<b>Coordinator:</b> Alkis Vazacopoulos	
	Theodore Trafalis	University of Oklahoma
		Knowledge Base-Clustering of Multi-Class SVM for Genes Expression Analysis
	George Karypis	University of Minnesota
		Sub-structure Based Approaches for Hit-Compound Discovery in Drug Design
12:00 - 1:30	<b>Lunch</b>	
1:30 - 3:00	<b>Session M2</b>	
	<b>Coordinator:</b> Jian Pei	
	James Keller	University of Missouri
	Jianbo Gao	University of Florida
		Multifractal and Recurrence Time Based Methods for Deciphering the Structures of Genomic DNA Sequences
	Giuseppe Lancia	University of Udine, Italy
		Combinatorial Optimization Problems in the Study of Human Polymorphisms
3:00 - 3:30	<b>Coffee Break (349 Reitz Union)</b>	
3:30 - 5:00	<b>Session M3 (349 Reitz Union)</b>	
	<b>Coordinator:</b> Theodore Trafalis	
	Stanislav Busygin	University of Florida
	Jian Pei	SUNY Buffalo
		The Pattern-based Approaches to Mining Microarray Data
	Halima Bensmail	University of Tennessee
		A Novel approach for Clustering Proteomics Data using Bayesian Fast Fourier Transform
6:30	<b>Dinner (Hilton Hotel)</b>	

CONFERENCE PROGRAM FOR FEBRUARY 17, 2004 TUESDAY (282 Reitz Union)			
8:30	<b>Registration</b>		
8:30 - 10:30	<b>Session T1</b> <b>Coordinator:</b> Vladimir Boginski		
	Ding-Zhu Du	National Science Foundation	DNA Library Screening and Pooling Designs
	Liang Chen	University of Northern British Columbia, Canada	Medical Data Classification by Differential Latent Semantic Indexing Approach
	Sanjay Ranka	University of Florida	Data Mining For Fast Detection of Emerging Pathogens and Bioterrorism Attacks
	Pando Georgiev	Brain Science Institute, RIKEN, Wako-shi, Saitama, Japan	Sparse Component Analysis: a New Tool for Data Mining
10:30 - 10:45	<b>Coffee Break</b>		
10:45 - 12:00	<b>Session T2</b> <b>Coordinator:</b> Wanpracha Chaovaitwongse		
	Peter L. Hammer	Rutgers University	Logical Analysis of Data and Biomedical Applications
	Alkis Vazacopoulos	Dash Optimization, Inc.	Using Xpress-Mosel for Modeling and Solving Data Mining Problems
12:00 - 1:30	<b>Lunch</b>		
1:30 - 3:00	<b>Session T3</b> <b>Coordinator:</b> Carlos Oliveira		
	Anand Rangarajan	University of Florida	Entropy estimation in high dimensional feature spaces with applications to image registration
	Jose Principe	University of Florida	Applying data mining concepts to multidimensional time series analysis
	Su-Shing Chen	University of Florida	Ontology Search and Text Mining of MEDLINE Database
3:00 - 3:30	<b>Coffee Break</b>		
3:30 - 5:00	<b>Session T4</b> <b>Coordinator:</b> Alkis Vazacopoulos		
	Anthony Okafor	University of Florida	Data Clustering Via Entropy Minimization and Evidence Gathering
	Carlos Oliveira	University of Florida	A Branch-and-Bound Algorithm for the Closest String Problem
	Christian Cardenas-Lailhacar	University of Florida	Capacitance, Hardness and the Potential Energy Function: Chemical Reactivity Prediction
6:30	<b>Dinner (Hilton Hotel)</b>		

CONFERENCE PROGRAM FOR FEBRUARY 18, 2004 WEDNESDAY (282 Reitz Union)

8:30	<b>Registration</b>		
9:00 - 10:30	<b>Session W1</b>		
	<b>Coordinator:</b> Panos Pardalos		
	J. Chris Sackellares	University of Florida	Dynamical Entrainment among Epileptic Brain Areas
	Wanpracha Chaovalitwongse	Rutgers University	Data Mining In EEG: Application to Epilepsy
	Su-Shing Chen	University of Florida	A Degradomic Data Mining Algorithm for Brain Diseases
10:30 - 10:45	<b>Coffee Break</b>		
10:45 - 12:45	<b>Session W2</b>		
	<b>Coordinator:</b> Stanislav Busygin		
	Sergiy Butenko	Texas A&M University	Clique-detection approaches in biochemistry and genomics
	Hans van Oostrom	University of Florida	Quality of Data in Biomedicine
	Xiuzhen Cheng	George Washington University	An Ensemble Method of Discovering Sample Classes Using Gene Expression Profiling
	Dechang Chen	Uniformed Services University	CpG Island Identification with Higher Order and Variable Order Markov Models
12:45 - 2:00	<b>Lunch</b>		
2:00 - 4:00	<b>Session W3</b>		
	<b>Coordinator:</b> Vladimir Boginski		
	Bala Krishnamoorthy	University of North Carolina	A Topology-Based Characterization of Protein Structure
	Paul Carney	University of Florida	A Quantitative EEG Method for "Real-Time" Detection of Neonatal Seizures in the Neonatal Intensive Care Unit
	Deng-Shan Shiau	University of Florida	Real-Time Prospective Seizure Prediction and Statistical Assessment
	Oleg Prokopyev	University of Florida	Network-Based Techniques in EEG Data Analysis and Epileptic Brain Modeling

# *Computing with Words and its Application to Deduction, Definition and Search*

**Lotfi A. Zadeh**

Department of Electrical Engineering and Computer Sciences  
University of California at Berkeley

`zadeh@cs.berkeley.edu`

## **Abstract**

Computing with words and perceptions, or CWP for short, is a mode of computing in which the objects of computation are words, propositions and perceptions described in a natural language.

Perceptions play a key role in human cognition. Humans-but not machines-have a remarkable capability to perform a wide variety of physical and mental tasks without any measurements and any computations. Everyday examples of such tasks are driving a car in city traffic, playing tennis and summarizing a book.

One of the major aims of CWP is to serve as a basis for equipping machines with a capability to operate on perception-based information. A key idea in CWP is that of dealing with perceptions through their descriptions in a natural language. In this way, computing and reasoning with perceptions is reduced to operating on propositions drawn from a natural language. In CWP, what is employed for this purpose is PNL (Precisiated Natural Language.) In PNL, a proposition,  $p$ , drawn from a natural language, NL, is represented as a generalized constraint, with the language of generalized constraints, GCL, serving as a precisiation language for computation and reasoning, PNL is equipped with two dictionaries and a modular multiagent deduction database. The rules of deduction are expressed in what is referred to as the Protoform Language (PFL).

Any measurement-based theory,  $T$ , may be generalized to a perception-based theory,  $T_p$ , by adding to  $T$  the capability to operate on perception-based information. Two generalizations that are of particular importance involve probability theory,  $PT$ , and decision analysis,  $DA$ . Conceptually, computationally and mathematically,  $PT_p$  and  $DA_p$  are significantly more complex than their measurement-based versions. In this instance, as in many others, complexity is the price that has to be paid to reduce the gap between theory and reality.

# *Knowledge Base-Clustering of Multi-Class SVM for Genes Expression Analysis*

Theodore B. Trafalis<sup>1</sup> , Budi Santosa<sup>1</sup> and Tyrrell Conway<sup>2</sup>

<sup>1</sup>School of Industrial Engineering, University of Oklahoma

<sup>2</sup>Department of Botany and Microbiology, University of Oklahoma

ttrafal@ou.edu

## **Abstract**

This paper utilizes Support Vector Machines (SVM) for multi-class classification of a real data set with more than two classes. The data is a set of E. coli whole-genome gene expression profiles. The problem is how to classify these genes based on their behavior in response to changing pH of the growth medium and mutation of the acid tolerance response gene regulator GadX. The labels indicate the response of genes to the experimental variables: 1-unchanged, 2-decreased expression level and 3-increased expression level. To label the genes, an unsupervised K-Means clustering technique is applied in a multi-level scheme. Two other methods, Learning Vector Quantization (LVQ) network and Linear Discriminant Analysis (LDA) are implemented. Multi-class SVM is used for one-against-one method and one-against-all method. The results show that SVM outperforms LVQ and LDA. SVM experiments are performed using Matlab codes. For LVQ and LDA, experiments are performed using Neural Networks and Statistics Toolbox in Matlab.

# *Sub-structure Based Approaches for Hit-Compound Discovery in Drug Design*

**George Karypis**

Computer Science & Engineering Dept.  
University of Minnesota

karypis@cs.umn.edu

## **Abstract**

Discovering new drugs is an expensive and challenging process. Any new drug should not only produce the desired response to the disease but should do so with minimal side effects and be superior to the existing drugs in the market. One of the key steps in the drug design process is to identify the chemical compounds (widely referred to as “hit” compounds) that display the desired and reproducible behavior against the disease in a biological experiment. The standard technique to discover such compounds is to evaluate them with a biological experiment, known as an assay. The 1990s saw the widespread adoption of high-throughput screening (HTS), which uses highly automated techniques to conduct the biological assays and can be used to screen a large number of compounds. Though in principle, HTS techniques can be used to test each compound against every biological assay, it is never practically feasible for the following reasons. First, the number of chemical compounds that have been synthesized or can be synthesized using combinatorial chemistry techniques is extremely large. Evaluating this large set of compounds using HTS can be prohibitively expensive. Second, not all biological assays can be converted to high throughput format. Third, in most cases it is hard to find all the desirable properties in a single compound and chemists are interested in not just identifying the hits but studying what part of the chemical compound leads to desirable behavior, so that new compounds can be rationally synthesized.

In this talk we present our research on developing computational techniques based on classification that can be used to identify the hit compounds. These computational techniques can be used to replace or supplement the biological assay techniques. One of the key challenges in developing classification techniques for chemical compounds stems from the fact that their properties are strongly related to their chemical structure. However, traditional machine learning techniques are suited to handle datasets represented by multidimensional vectors or sequences, and cannot handle the structural nature of the chemical structures. We present a sub-structure-based classification algorithm that decouples the sub-structure discovery process from the classification model construction and uses frequent subgraph discovery algorithms to find all topological and geometric sub-structures present in the dataset. The advantage of our approach is that during classification model construction, all relevant sub-structures are available allowing the classifier to intelligently select the most discriminating ones. The computational scalability is ensured by the use of highly efficient frequent subgraph discovery algorithms coupled with aggressive feature selection. Our experimental evaluation on eight different classification problems shows that our approach is computationally scalable and outperforms existing schemes by 10% to 35%, on the average.

# *Gene Ontology-based Similarity Measures for Gene Clustering and Knowledge Discovery*

James M. Keller<sup>1</sup>, Mihail Popescu<sup>2</sup>, and Joyce Mitchell<sup>2</sup>

<sup>1</sup> Health Management and Informatics Department

<sup>2</sup> Electrical and Computer Engineering Department

University of Missouri

Columbia, MO 65211-2060

kellerj@missouri.edu

## **Abstract**

In clustering and subsequent knowledge discovery on unknown gene products, the primary features to date are the gene sequence and expression values found following a microarray experiment. One major goal is to determine the function of this gene product and its similarity in function or structure to other up-regulated or down-regulated gene products. Many measures have been proposed to calculate closeness of sequences. However, for many gene products, additional information comes from the set of Gene Ontology (GO) annotations and the set of journal abstracts related to the gene product. For these genes, it is reasonable to include similarity measures based on the terms found in the GO and/or the index term sets of the related documents (MeSH annotations). In both cases we deal with comparing two sets of terms arranged in a taxonomy (GO or MeSH.). Some measures have been constructed to assess closeness of terms in a taxonomy, including shortest path length between terms and information theory-related values where node probabilities are estimated using a corpus of relevant documents. Utilizing such factors in addition to sequence and expression should aid in the process of knowledge discovery. It will be easier to annotate clusters, for example, when they share common descriptive terms. When an unknown gene product joins the group via sequence and expression, it is reasonable to conjecture that this gene will also share the cluster annotations (at least partially).

In this talk we propose a fuzzy measure-based similarity (FMS) for computing the similarity of two sets of terms found in a taxonomy (and hence, the two gene products annotated with terms from the taxonomy). The advantage of FMS is that it takes into consideration the context of the whole set when computing the similarity. In dealing with large groups of terms and/or documents describing the objects under consideration, not only do we determine the similarity between the document pairs, but, by introducing the Choquet integral, we fuse this partial agreement function on pairs of documents into a single value relating the gene products. The measures for the final integral fusion can be tailored to produce order weighted average (OWA) operators (e.g., “at least two documents must support the connection”) or can be based on assessments of the “worth” of individual and subsets of documents towards building the strength of connection. We present examples and comparisons to other approaches from the literature. Additionally, we display preliminary clustering results utilizing these measures. Finally, we indicate a potential use of the GO approach in determining function of “unknown” genes.

*Multifractal and Recurrence Time Based Methods for Deciphering the Structures of Genomic DNA Sequences*

**Jianbo Gao**

Department of Electrical and Computer Engineering  
University of Florida

`gao@ece.ufl.edu`

**Abstract**

The completion of the human genome and genomes of many other organisms calls for the development of faster computational tools which are capable of easily identifying the structures and extracting features from DNA sequences. Here we discuss two novel algorithms. One is for finding genes based on the observation that multifractal features of a DNA segment often have abrupt changes near the borders of gene-gene and coding-noncoding regions. The other is based on recurrence time statistics, which has its root in nonlinear dynamical systems theory and can conveniently study all kinds of periodicity and exhaustively find all repeat-related features from a genomic DNA sequence. A convenient coding region indicator based on the recurrence time statistics can be developed. Applications of these two methods to other disciplines of biomedicine will also be discussed.

# *Combinatorial Optimization Problems in the Study of Human Polymorphisms*

**Giuseppe Lancia**

Dipartimento di Matematica e Informatica  
University of Udine, Udine, Italy

`lancia@dimi.uniud.it`

## **Abstract**

A polymorphism is a trait which shows variability in a population (e.g., the blood type): without polymorphisms, we would all look the same! The possible values of the trait are called alleles. At genomic level, a polymorphism is a DNA region (string of A, T, C and Gs) whose content varies in a population. The smallest such polymorphism consists of a single base, and is called Single Nucleotide Polimorphism (SNP, pronounced “snip”). Determining the allele values for a set of SNPs, given the output from lab experiments, for either an individual or an entire population, gives rise to a set of nice and challenging combinatorial problems, which can be attacked by means of Optimization techniques. These problems have been extensively studied in the last few years, by many researchers. In this talk, we will review the main results of the very recent years in this area, while presenting them in a unified setting. We will give polynomial algorithms for some special cases, NP-hardness results for some others, and exponential (branch and bound) algorithms for still others.

# *Discovering Disease-Relevant Genes in DNA Microarray Data*

**Stanislav Busygin and Panos M. Pardalos**

Department of Industrial and Systems Engineering  
University of Florida

busygin@ufl.edu

## **Abstract**

Due to the DNA microarray chips produced by Affymetrix it has become possible to discover and predict genetic patterns relevant for various diseases on the basis of exploration of massive datasets provided by DNA microarray probes. A number of data mining techniques have been used for such exploration to achieve the desirable results. However, high dimensionality and uncertain accuracy of microarray datasets remain the major obstacles in revealing the most crucial genetic factors determining a particular disease. In this presentation we describe a microarray data processing technique based on the correspondence analysis helping to handle this issue.

# *The Pattern-based Approaches to Mining Microarray Data*

**Jian Pei**

Department of Computer Science and Engineering  
University at Buffalo, The State University of New York

`jianpei@cse.buffalo.edu`

## **Abstract**

With the rapid technical advances in biology, more and more microarray data becomes available. Mining microarray data is of great interest and important value in biomedicine and bio-informatics research. In this talk, we will discuss the problem of mining various patterns hidden deeply in microarray data. In particular, we examine three interesting problems: mining phenotypes, pattern-based clustering and interactive exploration of gene expression patterns. We present our novel pattern-based approaches that use machine learning and database techniques. The experimental results on real data sets show that some biologically meaningful patterns are found.

This talk reports the research results from the joint work with my colleague Dr. Aidong Zhang and our students at the State University of New York at Buffalo, and Dr. Haixun Wang and Dr. Philip S. Yu at IBM T.J. Watson Research Center.

# *A Novel approach for Clustering Proteomics Data using Bayesian Fast Fourier Transform*

**Halima Bensmail<sup>1</sup>, O. John Semmes<sup>2</sup> and Abdelali Haoudi<sup>2</sup>**

<sup>1</sup>University of Tennessee, Department of Statistics, 334 Stokely Management Building, Knoxville, TN 37996-0532,

<sup>2</sup>Eastern Virginia Medical School, Department of Microbiology and Molecular Cell Biology, Norfolk, VA 23501

bensmail@utk.edu

## **Abstract**

Proteomics studies can provide a wealth of information and rapidly generate large quantities of data from the analysis of biological specimens from healthy and diseased individuals. The high dimensionality of data generated from these studies requires the development of improved bioinformatics tools for efficient and accurate data analyses.

We present novel algorithms that can organize, cluster and derive meaningful patterns of expression from large scaled proteomics experiments. We processed raw data using graphical-based algorithm by transforming it from real space data-expression to a complex space data-expression using discrete Fourier transformation; then we used a thresholding approach to denoise and reduce the length of each spectra. Bayesian clustering was then used on the reconstructed data. Using this approach, we were able to successfully denoise proteomic spectra and reach up to a 99% total reduction of the number of peaks compared to the original data. In addition, the Bayesian-based approach generated better classification rate of 97% (misclassification error rate is 3%).

This new finding will allow us to apply the Fourier transformation for the selection of the protein profile for each sample, and to develop a novel bioinformatic strategy based on Bayesian clustering for optimal diagnosis and biomarker discovery.

*DNA Library Screening and Pooling Designs*

**Weili Wu<sup>1</sup>, Chih-Hao Huang<sup>1</sup>, Yingshu Li<sup>1</sup> and Ding-Zhu Du<sup>1,2</sup>**

<sup>1</sup> Computer Science Department, University of Minnesota

<sup>2</sup> National Science Foundation

ddu@nsf.gov

**Abstract**

The pooling design is used in DNA screening. In this talk, we explain some new ideas to construct pooling designs with simplicial complexes, graph properties, and finite element geometry.

# *Medical Data Classification by Differential Latent Semantic Indexing Approach*

**Liang Chen**

Computer Science Department,  
University of Northern British Columbia,  
Prince George, BC, Canada V2N 4Z9

chenl@unbc.ca

## **Abstract**

Medical data sets are always obtained under situations in which not all the measurements are available and accurate, but we could believe that the attributes of medical data should always be correlated, by observing that a physician's diagnosis is always based on the patient's integrated symptoms. Latent Semantic Indexing (LSI) is originally proposed for content-based text document classification, it is most effective when some entries of some data are missing. It has been successfully extended to the area of content-based image data analysis. A Differential Latent Semantic Indexing (DLSI) approach has been proposed for information retrieval and demonstrated improved performance over standard LSI approach by adapting to the unique characteristics of individual document. This paper investigates the application of DLSI approach in medical numerical data analysis.

# *Data Mining For Fast Detection of Emerging Pathogens and Bioterrorism Attacks*

**Chris Jermaine and Sanjay Ranka**

Computer and Information Science Department  
University of Florida

cjermain@cise.ufl.edu, ranka@cise.ufl.edu

## **Abstract**

With the slow response to last winter's SARS outbreak and the recent concerns regarding bioterrorism, there is a growing worry that the existing infrastructure for rapid detection of emerging or virulent diseases is inadequate. Even in the United States, the current system for detecting the occurrence of known, dangerous diseases like smallpox and anthrax is based on phone calls and faxes, and does not make use of current technology. The existing system relies on doctors and other health professionals to correctly recognize the symptoms of a set of particularly dangerous diseases, and to then take the initiative to notify the proper officials in the case of a suspected infection. In the case of emerging pathogens whose symptoms are still unknown, it is up to public officials to notice that something out of the ordinary is taking place, without the support of any systematic data collection and analysis tools.

In this talk, we will describe our long-term vision for a system supporting the automatic detection and reporting of emerging pathogens and bioterrorism attacks. The system we envision will perform automatic data collection and integration over a wide range of sources, such as hospitals, pharmacies, schools, and so on. Over a long-term time period, the system will construct a statistical baseline that describes what is "normal." That is, given environmental conditions such as recent weather, season, day of the week, and so on, the system automatically maintains an evolving model that can be used to determine how unusual any recently observed pattern is. Recent observations will be considered in the context of this baseline in order to find anomalies that may correspond to a disease outbreak, and thus require further investigation by human experts.

At a high level, the system will rely on real-time data mining. The new and recent data are always checked against the historical data to find worrisome patterns, without a precise, a priori definition of exactly what a "worrisome pattern" is. The benefit of this approach is that it addresses a fundamental weakness of traditional epidemiology, which is concerned with statistical modeling of known pathogens. Epidemiological techniques are much less useful for fast detection of poorly understood or emerging pathogens.

Clearly, such a system is far from being a reality. Thus, in our talk we will describe some of the research challenges that must be addressed for this vision to be realized. We will consider technical issues related to privacy and security, information integration, and statistical anomaly detection.

# *Sparse Component Analysis: a New Tool for Data Mining*

Pando Georgiev<sup>†</sup>, Fabian Theis<sup>†‡</sup>, Andrzej Cichocki<sup>†</sup> and Hovagim Bakardjian<sup>†</sup>

<sup>†</sup> Brain Science Institute, RIKEN

Lab. for Advanced Brain Signal Processing

2-1, Hirosawa, Wako-shi, Saitama, 351-0198, Japan

<sup>‡</sup> Institute of Biophysics, University of Regensburg D-93040 Regensburg, Germany

georgiev@bsp.brain.riken.go.jp

## Abstract

In many practical problems for data mining the data  $\mathbf{X}$  under consideration (given as  $(m \times N)$ -matrix) is of the form  $\mathbf{X} = \mathbf{A}\mathbf{S}$ , where the matrices  $\mathbf{A}$  and  $\mathbf{S}$  with dimensions  $m \times n$  and  $n \times N$  respectively (often called mixing matrix or *dictionary* and source matrix) are unknown ( $m \leq n < N$ ). We formulate conditions (k-SCA-conditions) under which we can recover  $\mathbf{A}$  and  $\mathbf{S}$  uniquely (up to scaling and permutation), such that  $\mathbf{S}$  is *sparse of level  $k$*  in sense that each column of  $\mathbf{S}$  has at most  $m - k$  nonzero elements ( $k = 1, 2, \dots, m - 1$ ). We call this *k-Sparse Component Analysis* problem (k-SCA).

We present new algorithms for identification of the mixing matrix (under k-SCA-conditions), and for source recovery (under identifiability conditions). The methods are illustrated with examples showing good performance of the algorithms. Typical examples are EEG data sets, in which the k-SCA algorithm allows us to detect some features of the brain signals. Special attention is given to the application of our method to the transposed system  $\mathbf{X}^T = \mathbf{S}^T \mathbf{A}^T$  utilizing the sparseness of the mixing matrix  $\mathbf{A}$  in appropriate situations. We note that the sparseness conditions could be obtained with some preprocessing methods and no independence conditions for the source signals are imposed (in contrast to Independent Component Analysis).

# *The Logical Analysis of Data and Applications to Biomedical Informatics*

**Peter L. Hammer**

RUTCOR, Rutgers University

hammer@rutcor.rutgers.edu

## **Abstract**

The Logical Analysis of Data (LAD), a combinatorics, optimization, and Boolean algebra-based methodology for extracting information from data, was first proposed in 1986, with a first paper on the topic published in 1988 (*Annals of Operations Research* 16, 1988), and a first report on its implementation, along with some applications, published in 2000 (*IEEE Transactions on Knowledge and Data Engineering* 12, 2000). LAD's results have been used for classification (including diagnosis and prognosis), for the analysis of the importance and the role of variables, for discovering new classes, for development of decision support systems, etc. One of the most important areas in which LAD has been recently applied is biomedical informatics.

The survey will present an outline of the basic concepts, techniques and algorithmic issues (e.g. support set selection, discretization, pattern enumeration, model formation, construction of discriminants) of LAD. We shall also present results of recent applications of LAD to some problems in biomedical informatics, including ovarian cancer diagnosis using proteomic datasets (*Proteomics*, March 2004), risk stratification among cardiac patients (*Circulation* 106, 2002), and some yet unpublished ones concerning biomaterial design optimization, the analysis of genomic data for the prediction of metastases development in breast cancer and the distinction between various types of diffuse large B-cell lymphomas.

# *Using Xpress-Mosel for Modeling and Solving Data Mining Problems*

**Alkis Vazacopoulos**

Dash Optimization, Inc.

av@dashoptimization.com

## **Abstract**

In this presentation, we will use the modeling and programming environment Xpress-Mosel to demonstrate how to model and solve complicated Data Mining Problems. We will focus on showing how Linear, Mixed Integer, Quadratic and Nonlinear Programming algorithms can be utilized to efficiently solve difficult classification and regression models.

*Entropy estimation in high dimensional feature spaces with applications to image registration*

**Aditee Kumthekar and Anand Rangarajan**

CISE Department  
University of Florida

anand@cise.ufl.edu

**Abstract**

Estimation of entropy in high dimensional feature spaces is frequently encountered in many machine learning and pattern recognition applications. Most previous work in this area begins with density estimation (using kernel regression and the like) with subsequent estimation of entropy from the estimated density. However, only recently has it been realized that direct nonparametric estimation of entropy in high dimensional feature spaces is possible and indeed practical. Hero et al. 2001, following Steele 1988, have shown that the Renyi alpha-entropy can be approximately computed from the Euclidean minimum spanning tree formed from an appropriately defined feature graph of Euclidean power weighted edges. The significance of this result lies in the fact that the computation of the minimum spanning tree can be accomplished approximately in  $O[(E \log E)d]$  time where  $E$  is the number of edges and  $d$  the dimension of the feature space. We apply this method to image registration, wherein we seek to maximize the mutual information between two images w.r.t. an unknown pose.

# *Applying data mining concepts to multidimensional time series analysis*

**José Principe**

Computational NeuroEngineering Laboratory, Electrical and Computer Engineering  
University of Florida

`principe@cnel.ufl.edu`

## **Abstract**

Time series analysis is normally applied to single channel data, and the problem is to find its structure in time. However, in many modern applications in plant controls, monitoring and medicine, many channels must be processed to fully characterize the phenomenon under analysis. Most often, not all the channels are equally important and the least valuable ones should be discarded to yield smaller models that maximize generalization. In such cases tools for variable selection utilized in data mining may be applicable. We will present current work on choosing important channels in epilepsy and brain machine interfaces.

# *Ontology Search and Text Mining of MEDLINE Database*

**Hyunki Kim and Su-Shing Chen**

CISE Department  
University of Florida

suchen@cise.ufl.edu

## **Abstract**

With the explosion of biomedical data, information overload and users' inability of expressing their information needs may become more serious. To solve those problems, this paper presents a text data mining method that uses both text categorization and text clustering for building concept hierarchies for MEDLINE citations in the Knowledge Grid. The approach we propose is a three-step data mining process for organizing MEDLINE database: (1) categorizations according to MeSH terms, MeSH major topics, and the co-occurrence of MeSH descriptors; (2) clustering using the results of MeSH term categorization; and (3) visualization of categories and hierarchical clusters. The hierarchies automatically generated can be used to generate a medical ontology and to construct multiple viewpoints of a collection. Providing multiple viewpoints of a document collection and allowing users to move among these viewpoints will enable both inexperienced and experienced searchers to more fully exploit the information contained in a document collection. User interfaces with multiple viewpoints for this underlying system are also presented.

# *Data Clustering Via Entropy Minimization and Evidence Gathering*

**Anthony Okafor and Panos M. Pardalos**

Department of Industrial and Systems Engineering  
University of Florida  
aokafor@ufl.edu

## **Abstract**

Data analysis often requires the unsupervised partitioning of the data set into clusters. Clustering data is an important but a difficult problem. In the absence of prior knowledge about the shape of the clusters, similarity measures for a clustering technique is hard to specify.

In this work, we propose a framework for learning the structure of the data by evidence gathering. Evidence gathering is accomplished by randomly applying the K-means algorithm multiple times on the data via entropy minimization in order to obtain similarity measures between pairs of patterns. Final data clustering is obtained by applying some threshold on the similarity measures.

# *A Branch-and-Bound Algorithm for the Closest String Problem*

Cláudio N. Meneses<sup>2</sup>, Carlos A.S. Oliveira, and Panos M. Pardalos

Department of Industrial and Systems Engineering  
University of Florida

oliveira@ufl.edu

## **Abstract**

We are concerned with the Closest String Problem (CSP), which can be defined as follows: given a finite set  $\mathcal{S} = \{s^1, s^2, \dots, s^n\}$  of strings, each string with length  $m$ , find a center string  $t$  of length  $m$  minimizing  $d$ , such that for every string  $s^i \in \mathcal{S}$ ,  $d_H(t, s^i) \leq d$ . By  $d_H(t, s^i)$  we mean the Hamming distance between  $t$  and  $s^i$ . This problem has applications in Molecular Biology and Coding Theory.

The CSP is known to be NP-hard [2]. Even though there are good approximation algorithms for this problem (including a PTAS [3]), and exact algorithms for instances with constant difference [1], there are no studies reporting on the quality of integer programming formulations for the general case, when the maximum distance  $d$  between a solution and the strings in  $\mathcal{S}$  is variable.

We propose three integer programming formulations and prove some containment relations among them. A heuristic algorithm is also provided, which allows us to find good upper bounds on the value of an optimal solution. We report computational results of a branch-and-bound algorithm based on the IP formulations, and of the heuristic, on a set of instances representative of real applications. These results show that it is possible to solve CSP instances of moderate size to near optimality using optimization techniques.

## **References**

- [1] P. Berman, D. Gumucio, R. Hardison, W. Miller, and N. Stojanovic. A linear-time algorithm for the 1-mismatch problem. In WADS'97, 1997.
- [2] M. Frances and A. Litman. On covering problems of codes. *Theor. Comput. Syst.*, 30:113–119, 1997.
- [3] M. Li, B. Ma, and L. Wang. On the closest string and substring problems. *Journal of the ACM*, 49(2):157–171, 2002.

---

<sup>2</sup>This author is supported by the Brazilian Federal Agency for Post-Graduate Education (CAPES) - Grant No. 1797-99-9.

# *Capacitance, Hardness and the Potential Energy Function: Chemical Reactivity Prediction*

**Christian Cardenas-Lailhacar**

Department of Industrial and Systems Engineering  
University of Florida

cardenas@ise.ufl.edu

## **Abstract**

The potential energy function (E) for chemical reactivity has been of great interest in chemistry, particularly for its applications in drug design, prediction of protein and molecular structure, industrial chemical processes, catalysis, environmental studies, nano-technology, biomedicine, etc. The multiple levels of theory that quantum chemistry has developed for its study, have generated a variety of tools that have shown their great ability to extract the desired information from systems of interest. Moreover, they allow theoreticians to reproduce experimental results, to suggest the repetition of some, and to predict properties and structures.

The theoretical basis to study reactivity parameters of molecular systems, this is, to its resistance to change its electron density distribution, namely the molecular hardness ( $\eta$ ) has been provided lately by density functional theory (DFT). In the myriad of properties that help to characterize molecular systems and their interactions, Capacitance (C), a property of molecular systems commonly studied in solid state systems, is interesting in drugs and biological systems as it provides with information regarding the ability of a system to store charge. As in the potential energy function and hardness, its evolution along a given reaction coordinate is examined in order to better understand the chemical reactivity of the system of interest. We have previously developed analytical expressions for E and  $\eta$ . In this work we show expressions for the Capacitance, and moreover, analytical expressions to relate Hardness and Capacitance with the Potential Energy function. The importance of the molecular symmetry of the system is included, and its influence explored at the point to suggest corrections to today accepted expressions and values. Furthermore, we study compounds that are often used as prototypes to study their linkage in proteins.

# *Preictal Transition in Temporal Lobe Epilepsy*<sup>1</sup>

**J.C. Sackellares**<sup>1,2,3,8,9</sup>, **D.-S. Shiau**<sup>1,8,9</sup>, **L.D. Iasemidis**<sup>10,11</sup>, **P.R. Carney**<sup>6,8</sup>  
**P.M. Pardalos**<sup>3,4,5,7</sup>, **W. Chaovaitwongse**<sup>12,13</sup>

Departments of <sup>1</sup>Neuroscience, <sup>2</sup>Neurology, <sup>3</sup>Biomedical Engineering, <sup>4</sup>Industrial and Systems Engineering, <sup>5</sup>Computer and Information Science and Engineering, and <sup>6</sup>Pediatrics  
<sup>7</sup>Center for Applied Optimization, <sup>8</sup>McKnight Brain Institute  
University of Florida, Gainesville

<sup>9</sup>Malcolm Randall V.A. Medical Center, Gainesville, Florida  
Department of <sup>10</sup>Biomedical Engineering and

<sup>11</sup>Center for Systems Science and Engineering Research  
Arizona State University, Tempe, Arizona.

<sup>12</sup>Corporate Strategic Research, ExxonMobil Research and Engineering Company, Annandale, NJ  
08801, USA;

<sup>13</sup>Department of Industrial and Systems Engineering, Rutgers, The State University of New Jersey,  
Piscataway, NJ 08854, USA

sackellares@epilepsy.health.ufl.edu

## **Abstract**

**Rationale/Objective:** Recurrent seizures, a characteristic of epilepsy, appear to occur spontaneously as a result of complex dynamical interactions among many regions of the brain, we have defined this phenomenon as Dynamical Entrainment. The dynamical entrainment among brain areas during the preictal transition period can be reflected by the convergence of short-term maximum Lyapunov exponents (STLmax), a nonlinear dynamics technique based on the chaos theory. Through the STLmax analysis on the EEG recordings from ten patients with a total of 161 epileptic seizures, the objective of this study is to test the following hypotheses of the dynamical characteristics during the preictal, ictal and postictal states: (1) spatio-temporal chaotic state is more ordered during the ictal period than during the preictal period, (2) seizure occurs only when the epileptogenic focus is interacting with other cortical areas during the preictal period, and (3) seizures serve to reset the brain and reverse the abnormal interactions among cortical sites.

**Methods:** Multichannel EEG recordings from 10 patients with a total of 161 seizures were analyzed. STLmax values were calculated sequentially over time for each 10.24 non-overlapping time window. Mean STLmax values (over channels) in the preictal and ictal periods were compared to test the temporal order in hypothesis 1. Based on the T-index of STLmax values, percentage of entrained pairs between brain areas in the preictal and ictal periods were compared to test the spatio order in hypothesis 1. T-index values between epileptogenic focus and normal areas during the preictal period were calculated to test the interactions in hypothesis 2. Finally, T-index values during the postictal period with respect to the abnormally interacted cortical sites were calculated to test the resetting mechanism in hypothesis 3.

**Results:** In all 10 patients, the mean STLmax values are significantly lower during the ictal period than during the preictal periods ( $p < 0.05$ ), and the entrainment percentage is significantly higher during the ictal period than during the preictal period ( $p < 0.05$ ). Between focus and normal areas, 89.2% of the seizures show entrainment during the preictal period, and 82.1% of which show disentrainment after the seizure (postictally).

**Conclusion:** Significantly lower mean STLmax values during the seizure periods suggest that the chaotic state is more temporal-ordered during the seizure periods. Significantly higher percentage of entrainment among brain areas during the seizure period suggests that the chaotic state is more spatio-ordered during the seizure periods. High percentage of seizures observes entrainment between epileptogenic focus and normal cortical sites suggests that seizure may occur only when the epileptogenic focus is interacting with other cortical areas during the preictal period. High percentage of disentrainment after the seizures suggests that seizures may serve to reverse the abnormal interactions between epileptogenic focus and normal cortical areas.

---

<sup>1</sup>This research is supported in part by the NIH, NSF, VA, Whitaker, and DARPA research grants.

# *Data Mining in EEG: Application to Epilepsy*

**Wanpracha Chaovalitwongse**

Corporate Strategic Research, ExxonMobil Research and Engineering Company, Annandale, NJ  
08801, USA;

Department of Industrial and Systems Engineering, Rutgers, The State University of New Jersey,  
Piscataway, NJ 08854, USA

arty@mbi.ufl.edu

## **Abstract**

Epilepsy is one of the most common brain disorders. Estimates of incidence rates range from 24 to 53 per 100,000 and 25-30% of all patients remain unresponsive to anti-epileptic drug treatment, which is the standard therapy for epilepsy. One of the most disturbing aspects of epilepsy is the occurrence of seizures, which appears to be random and unpredictable. Uncontrolled epilepsy poses a significant burden to society due to associated healthcare cost; therefore, there is a growing body of research of interest in detecting state transition of epileptogenic process, the development of seizures, from intracranial electroencephalogram (EEG). Although EEG time series are considered to be complicated multidimensional complex systems, it has been shown that characterization and quantification of dynamics of EEG may enable us to detect the state transition of epileptic seizures, which follow a change in dynamics of EEG.

In this research, we herein apply 3 quantitative analyses (Short-Term Maximum Lyapunov Exponents, Angular Frequency, and Entropy), which measure chaoticity and complexity of EEG signals, to detect and classify the state transition of seizures. Each of these measures can be used to display state transition of seizures, which 3 different states of patients can be classified (interictal, pre-ictal, and post-ictal states). A cross validation was implemented to estimate the accuracy of a classification. These results indicate that it may be possible to use this data mining technique to design and develop efficient seizure warning algorithms for diagnostic and therapeutic purposes.

---

<sup>1</sup>Joint work with P.M. Pardalos, L.D. Iasemidis, W. Suharitdamrong, P.R. Carney, and J.C. Sackellares. Research was partially supported by the Medical Research Service of the Department of Veterans Affairs, grants from the Department of Veterans Affairs Research, the NSF grants DBI-980821, EIA-9872509, and NIH grant R01-NS-39687-01A1.

# *A Degradomic Data Mining Algorithm for Brain Diseases*

**Su-Shing Chen<sup>1</sup>, William Haskins<sup>3</sup>, Nancy Denslow<sup>2</sup>, Andrew K Ottens<sup>2,3</sup>,  
Ronald L. Hayes<sup>3</sup>, and Kevin K. W. Wang<sup>3</sup>**

<sup>1</sup> Computer Information Science and Engineering,

<sup>2</sup> Interdisciplinary Center for Biotechnology Research (ICBR),

<sup>3</sup> Departments of Psychiatry and Neuroscience, McKnight Brain Institute,  
University of Florida.

suchen@cise.ufl.edu

## **Abstract**

While information about post-translational modifications (PTMs), particularly proteolytic processing, is often lost during in vitro proteolysis (e.g., trypsin digestion) prior to tandem mass spectrometry (MS2)-based protein identification, a novel degradomic data-mining algorithm can be developed to elucidate protein degradation pathways and distinguish in vivo proteolysis from in vitro proteolysis.

Proteolytic processing is a complex type of post-translational modification (PTM) that generates biologically active peptides and protein fragments with important cell signaling properties. Proteolytic processing of precursor proteins is performed by proteases (i.e., endo- and exopeptidases) during intracellular transport from the rough endoplasmic reticulum, where the proteins are folded, to the Golgi apparatus, where various other PTMs are performed. Further proteolytic processing occurs during stimulation-induced exocytotic release from secretory vesicles (SVs) into the extracellular fluid (ECF). Prediction of protease cleavage sites and peptide products is precluded by a number of unknowns including the consensus sequences and subcellular locations of the proteases, the kinetics of proteolysis and the tertiary and quaternary structures of the proteases and substrates. For example, it has recently been shown that proteolytic processing produces neuropeptides that are not predicted from known protease cleavage sites. Our data also shows that proteolysis of cellular proteins is a major biochemical editing event following brain injury. Therefore, novel experimental approaches are required to identify peptide products and protein fragments and novel computational approaches are required to reconstruct the various proteolytic processing pathways. This paper reports some novel results that we have recently obtained.

# *Clique-Detection Approaches in Biochemistry and Genomics*

**Sergiy Butenko<sup>1</sup>, Panos M. Pardalos<sup>2</sup>, Gerald Samson<sup>2</sup>**

<sup>1</sup> Industrial Engineering Department, Texas A&M University

<sup>2</sup> Industrial and Systems Engineering Department, University of Florida

butenko@tamu.edu

## **Abstract**

Many important data mining problems arising in biochemistry and genomics can be formulated in terms of certain combinatorial optimization problems. Some of these problems, including comparative modeling of protein structure, integration of genome mapping data and macromolecular docking can be solved using clique-detection approaches on specially constructed graphs. In this talk, we present a new algorithm for finding all maximal cliques in a graph. The results of preliminary numerical experiments are also reported.

# *Quality of Data in Biomedicine*

**Hans van Oostrom**

Biomedical Engineering Department,  
University of Florida

`hans@bme.ufl.edu`

## **Abstract**

The availability of electronic medical data has vastly increased in recent years due to advances in medical instrumentation. There is a risk of trusting the validity electronic data too much. In this talk we will explore issues with electronic medical data and propose solutions.

# *An Ensemble Method of Discovering Sample Classes Using Gene Expression Profiling*

Dechang Chen<sup>1</sup>, Zhe Zhang, Xiuzhen Cheng<sup>2</sup>

<sup>1</sup> Department of Preventive Medicine and Biometrics, Uniformed Services University

<sup>2</sup> Department of Computer Science, George Washington University

cheng@gwu.edu

## **Abstract**

Cluster methods have been successfully applied in gene expression data analysis to address tumor classification. Central to cluster analysis is the notion of dissimilarity between the individual samples. In clustering microarray data, dissimilarity measures are often subjective and predefined prior to the use of clustering techniques. In this paper, we present an ensemble method to define the dissimilarity measure through combining assignments of observations from a sequence of data partitions produced by multiple clusterings. This dissimilarity measure is then subjective and data dependent. We present our algorithm of hierarchical clustering based on this dissimilarity. Experiments on gene expression data show that the ensemble method is efficient in discovering sample classes.

# *CpG Island Identification with Higher Order and Variable Order Markov Models*

**Zhenqiu Liu and Dechang Chen<sup>1</sup>**

<sup>1</sup> Department of Preventive Medicine and Biometrics, Uniformed Services University

dchen@usuhs.mil

## **Abstract**

Identifying the location and function of human genes in a long sequence of genome is difficult due to lack of sufficient information about genes. Experimental evidences have suggested that there exist strong correlation between CpG islands and genes immediately following them. Much research has been done to identify CpG islands in a DNA sequence using various models. In this paper, we introduce two alternative models based on high order and variable order Markov chains. Compared with the popular models such as the first order Markov chain, HMM, and HMT, these two models are much easier to compute and have higher identification accuracies. One unsolved problem with the Markov model is that there is no way to decide the exact boundary point between CpG and non-CpG islands. In this paper, we provide a novel tool to decide the boundary points using the sequential probability test. Sequential data from GeneBank are used for the experiments in this paper.

# *A topology-based characterization of protein structure*

**Bala Krishnamoorthy**

Department of Operations Research,  
University of North Carolina, Chapel Hill, NC

kbala@email.unc.edu

## **Abstract**

We study ways of characterizing and categorizing protein structure by means of topological descriptions of strand adjacencies. Proteins are represented by a partial Delaunay complex, called the alpha complex, whose vertices are the alpha-carbon atoms of residues. This complex grows as a function of a distance threshold defining “neighboring” alpha-carbon atoms. Boundaries of neighborhoods of each residue, and of contiguous strands of varying lengths, are defined as sub-complexes of this complex at different distance thresholds. The Betti numbers (ranks of homology groups) of these sub-complexes are found. Secondary and tertiary structural motifs are identified by patterns in these Betti numbers.

A second characterization of protein structure is developed by defining a signature based on a residue-wise color map. A unique color is assigned for each triangle in the boundary of neighborhood of the back-bone chain, describing the position in the chain of the unique “nearest” non-adjacent alpha carbon. Information about secondary and tertiary structure is efficiently captured by patterns in this color signature.

# *A Quantitative EEG Method for “Real-Time” Detection of Neonatal Seizures in the Neonatal Intensive Care Unit<sup>1</sup>*

**Paul R. Carney<sup>1,2</sup>**

<sup>1</sup> Department of Pediatrics

<sup>2</sup> McKnight Brain Institute  
University of Florida

carnepr@peds.ufl.edu

## **Abstract**

**Significance of project:** Seizures may be the most frequent, and often the only clinical sign of central nervous system dysfunction in the neonate (Fenichel, 1985; Volpe, 1995). Seizures raise immediate concerns about the underlying cause of brain disorder, associated clinical conditions, the effect seizures may have on the developing brain, the need for antiepileptic drugs, and the effect antiepileptic drugs may have on the neonate with seizures. Despite the clinical significance of neonatal seizures, the universal acknowledgement of their importance, and the need for urgent and appropriate action when they are recognized, the inherent difficulty in recognizing seizures may limit the effectiveness of the clinician in the care of these infants (Mizrahi, 1987). In addition, because of the developmental immaturity of the brain, neonatal seizures may have unique clinical manifestations when compared to seizures in older infants, children, and adults. Seizures may be brief and infrequent and may also occur when trained personnel are not observing infants. Thus, the clinician is left to rely on historical information that may be incomplete or inaccurate. As a result, neonatal seizures may be more difficult to recognize than those of older age groups.

Electroencephalography (EEG) has an important role in the diagnosis and management of neurologic disorders in neonates. Its unique place in clinical care setting is based on its providing of “real-time” and continuous information concerning brain function. This is in contrast to other techniques, such as head ultrasonography and neuroimaging studies, which assess brain structure. The clinical significance of the EEG is that it may identify pathological changes in brain function and provide prognostic markers, which may have important roles in monitoring response to therapy and in specifically aiding clinical management, such as determining duration of therapy. It is generally agreed that more sensitive and specific quantitative EEG measures can provide relevant information about brain function prior to clinical manifestation that may represent a window of opportunity for appropriate interventions.

The major new thrust of research work in EEG analysis is to extract information from EEGs that is not available by visual analysis of the raw recording. Our team evolved out of the common interest and belief that research based upon rigorous mathematical principles can lead to scientific breakthroughs in the understanding of neurological systems, both normal and pathological. By quantitatively examining EEG data in patients with epilepsy, we discovered measurable changes in brain dynamics that precede and accompany seizures (Iasemidis et al., 2000). Brain electrical activity evolves to a measurable nonlinear state of maximum order just prior to seizure onset.

**Preliminary results:** Recently we have found that this same approach may help to detect neonatal seizures. Quantitative EEG analyses were performed on two ( $n = 2$ ) day of life 1, 38 weeks conceptional age neonates with seizures following hypoxic ischemic brain injury (Apgar scores  $< 5$ ) and one ( $n = 1$ ) age-matched control (Apgar scores  $> 7$ ). Nonlinear dynamical (i.e., short-term maximum Lyapunov exponent) and linear (i.e., energy) measures were performed on 1-hour continuous digital

14-channel scalp video-EEGs. Results demonstrate significantly lower mean short-term maximum Lyapunov exponent (STLmax) in neonates during seizures. Energy analysis of the same EEG data failed to identify differences between the two groups.

The present study will test the Hypothesis that nonlinear EEG analyses can detect seizures by quantitatively measuring changes in brain function during seizure periods; neonates with seizures will demonstrate significantly lower STLmax values relative to age-matched controls. The long-term goal of this research is to develop computer-based system for analyzing dynamical features of EEG signals that can be implemented at the bedside in order to help clinicians diagnose and manage neonatal seizures more effectively.

**Specific aims and expected outcomes:** *1) Develop and implement a computer-based system for analyzing dynamical features of EEG signals. Off-line algorithms will be integrated into a single, computer-based platform for on-line, real-time bedside detection of seizures. 2) Determine the specificity and sensitivity of the mathematical algorithms to be employed in order to detect periods of neonatal seizure susceptibility. We anticipate that the proposed algorithms will be able to detect seizure periods with a high rate of sensitivity and specificity based on our experience with seizure prediction in humans and animal models. 3) Measure changes in brain dynamics following pharmacological anticonvulsant interventions. Algorithms developed in specific aims 1 and 2 will be employed in aim 3.*

**Experimental methods and study design:** 38 to 42 week conceptional age neonates with hypoxic ischemic encephalopathy and (Apgar scores < 5) will undergo continuous video-EEG monitoring during seizure susceptible periods. EEGs will be visually evaluated, and all electrographic seizures will be noted and compared with quantitative nonlinear dynamical measures (STLmax and autocorrelation). The rationale for using these measures is that they have shown to be reliable in detecting seizure onsets. Results will be compared with age matched control neonates. We postulate that significant differences will exist between the two groups. In order to demonstrate 0.05 of significance and 95% power of the test (two-sample t test), at least 12 subjects are required in each group to detect the difference = 0.1 (SD) between the two groups. For 90% power, at least 10 subjects are required in each group. We will also determine global (long-term) as well as local exponents of the EEG signal in order to determine whether more detailed analysis of brain function provides more reliable and sensitive methods for identifying neonates at risk for seizures.

---

<sup>1</sup>Co-Investigators: Leonidas Iasemidis, Ph.D., Associate Professor of Bioengineering, Arizona State University, Tempe, AZ, J.C. Sackellares, M.D., Professor of Neurology and Bioengineering, Affiliate Professor of Neuroscience, University of Florida, P.M. Pardalos, Ph.D., Professor of Industrial and Systems Engineering, University of Florida, D.S. Shiau, Ph.D., Research Scientist, University of Florida V.A. Yatsenko, Ph.D., Visiting Professor, University of Florida, Oleg A. Prokopyev, Student, Department of Industrial and Systems Engineering, University of Florida

# *Real-Time Prospective Seizure Prediction and Statistical Assessment*<sup>1</sup>

**D.-S. Shiau**<sup>1,8,9</sup>, **L.D. Iasemidis**<sup>10,11</sup>, **P.R. Carney**<sup>6,8</sup>, **P.M. Pardalos**<sup>3,4,5,7</sup>,  
**W. Chaovalitwongse**<sup>12,13</sup>, **J.C. Sackellares**<sup>1,2,3,8,9</sup>

Departments of <sup>1</sup>Neuroscience, <sup>2</sup>Neurology, <sup>3</sup>Biomedical Engineering, <sup>4</sup>Industrial and Systems Engineering, <sup>5</sup>Computer and Information Science and Engineering, and <sup>6</sup>Pediatrics  
<sup>7</sup>Center for Applied Optimization, <sup>8</sup>McKnight Brain Institute  
University of Florida, Gainesville

<sup>9</sup>Malcolm Randall V.A. Medical Center, Gainesville, Florida  
Department of <sup>10</sup>Biomedical Engineering and

<sup>11</sup>Center for Systems Science and Engineering Research  
Arizona State University, Tempe, Arizona.

<sup>12</sup>Corporate Strategic Research, ExxonMobil Research and Engineering Company, Annandale, NJ  
08801, USA;

<sup>13</sup>Department of Industrial and Systems Engineering, Rutgers, The State University of New Jersey,  
Piscataway, NJ 08854, USA

shiau@epilepsy.health.ufl.edu

## **Abstract**

**Rationale/Objective:** Epilepsy is a common neurological disorder characterized by spontaneous recurrent seizures. In spite of advances in pharmacology, neuroimaging, clinical neurophysiology, and neurosurgery, seizures remain uncontrolled in many patients. The ability to predict epileptic seizures well before clinical onset promises new diagnostic applications and novel approaches to seizure control. Our group initially reported the existence of preictal transition and predictability of seizures based on the quantitative analysis of EEG signal characteristics (Iasemidis and Sackellares, 1996; Iasemidis et al., 1998; Sackellares et al., 1999). This finding has been confirmed by other investigators (Lehnertz and Elger, 1998; Quyen et al., 1999; Litt et al., 2001; Iasemidis et al., 2001, 2002). We have previously described an automated seizure prediction algorithm (Iasemidis et al., 2003). The objective of this study was to statistically evaluate the performance of an improved automated algorithm for seizure prediction.

**Methods:** Continuous long-term (total 132.4 days) intracranial and scalp EEG recordings obtained from 17 patients with intractable epilepsy (total 239 recorded seizures with mean seizure interval 10.2 hours) were analyzed to test the proposed algorithm. The algorithm utilizes concepts from nonlinear dynamics, statistics, an optimization method for the selection of critical cortical sites, and a novel method for the detection of the preictal transitions using adaptive thresholds according to the state of the EEG dynamics. Prediction receiver operating characteristic (ROC) curves from each patient were compared to ones produced by a statistically derived optimal naive prediction method for nondecreasing hazard rate (ONND) and two other naive prediction methods (periodic and random). Area above ROC curve is utilized to quantify the overall performance of the prediction method. Standard nonparametric Wilcoxon sign rank test was employed to obtain the overall significance (p-values) of the proposed prediction algorithm with respect to the ONNDH, periodic and random prediction algorithms.

**Results:** For patients with intracranial recordings, the mean ROC area for the proposed algorithm is 0.079, whereas the mean areas are 0.173, 0.167 and 0.195 for periodic, random and ONNDH prediction schemes, respectively. For scalp recording patients, the mean ROC areas are 0.1, 0.160,

0.159, and 0.158 for the proposed, periodic, random and ONNDH prediction schemes, respectively. The p-value for claiming that the proposed algorithm is better than any of the other once is less than 0.01, and the differences among the three naive prediction schemes are not significant (p-value > 0.05). Further, in the patients with intracranial recordings, with at least 80% prediction sensitivity, the proposed algorithm predicted 86.4% of seizures with an overall false warning rate 0.124 per hour (one per 8 hours). For scalp recording patients, the algorithm predicted 83.3% of seizures with an overall false warning rate 0.146 per hour ( $\approx$  one per 7 hours). The average prediction time for an impending seizure is 55.2 minutes for intracranial patients and 66.7 minutes for scalp patients.

**Conclusions:** The results of this study suggest that it is possible to predict an impending seizure based on the quantitative analysis of multichannel EEG recordings. By employing a statistical evaluation method based on the comparison of the ROC areas, the prediction performance of the proposed prediction algorithm is superior to the three compared naive prediction schemes. The performance of this algorithm, in this small sample of patients, is sufficient for a wide range of patient monitoring applications and therapeutic interventions, using implantable devices. Further studies in a larger sample of patients are warranted.

---

<sup>1</sup>This research is supported in part by the NIH, NSF, VA, Whitaker, and DARPA research grants.

*Network-Based Techniques in EEG Data Analysis and Epileptic Brain Modeling*

**Oleg A. Prokopyev<sup>4</sup>, Vladimir L. Boginski<sup>4</sup>, Wanpracha Chaovaitwongse<sup>10,11</sup>,  
Panos M. Pardalos<sup>3,4,5,7</sup>, J. Chris Sackellares<sup>1,2,3,8,9</sup>, Paul R. Carney<sup>6,8</sup>**

Departments of <sup>1</sup>Neuroscience, <sup>2</sup>Neurology, <sup>3</sup>Biomedical Engineering, <sup>4</sup>Industrial and Systems Engineering, <sup>5</sup>Computer and Information Science and Engineering, and <sup>6</sup>Pediatrics  
<sup>7</sup>Center for Applied Optimization, <sup>8</sup>McKnight Brain Institute  
University of Florida, Gainesville

<sup>9</sup>Malcolm Randall V.A. Medical Center, Gainesville, Florida

<sup>10</sup>Corporate Strategic Research, ExxonMobil Research and Engineering Company, Annandale, NJ 08801, USA;

<sup>11</sup>Department of Industrial and Systems Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

`oap4ripe@ufl.edu`

**Abstract**

We discuss a novel approach of modeling the behavior of the epileptic human brain, which utilizes network-based techniques in combination with statistical preprocessing of the electroencephalographic (EEG) data obtained from the electrodes located in different parts of the brain. In the constructed graphs, the vertices represent the “functional units” of the brain, where electrodes are located. Studying dynamical changes of the properties of these graphs provides valuable information about the patterns characterizing the behavior of the brain prior to, during, and after an epileptic seizure.