Third International Conference on Computational Biomedicine, CBM 2016

Hilton University of Florida Conference Center

February 25th – 27th, 2016
Gainesville, Florida, USA.

Sponsors

Informatics Institute
UNIVERSITY of FLORIDA

UF UNIVERSITY of FLORIDA

Springer

	Thursday	Friday	Saturday
08:15	Michailidis, G.		
08:30	Pardalos, P.M.	Conesa, A.	Principe, J.C.
09:00	Thambisetty, M.		
09:30		Prokopyev, O.	Yezerska, O.
10:00	Coffee break	Coffee break	Coffee break
10:30	Henderson, J.	Kammerdiner, A.	Shylo, O.
11:00	Thai, M.T.	Jörnsten, R.	Berry, R.B.
11:30			
12:00	Lunch	Lunch	Lunch
01:00	Bihorac, A.	Aljuaid, A.	
01:30	Li, X.	Panagopoulos, O.	
02:00		Ünlü, R.	
02:30	Coffee break	Coffee break	
03:00	Nicosia, G.	Sackellares, C.	
03:30			
04:00	Guarracino, M.	Moreno-Centeno, E.	
04:30	Stozhkov, V.D.	Pi, J.	
05:00	Zdanovskaya, V.	Kocheturov, A.	
05:30	Wang, P.		
06:00		Dinner	

Thursday, February 25th

7:30AM-8:20AM Conference registration

8:30AM-9:00AM

Panos M. Pardalos
Conference Opening: *Data Science and Optimization Challenges in Computational Biomedicine Research*

9:00AM-10:00AM

Madhav Thambisetty
Plenary Talk: *Seeking biomarkers and understanding mechanisms: building an integrated approach to Alzheimer's disease*

10:00AM-10:30AM Coffee Break

10:30AM-11:00AM

James Henderson
Signed Gene Set Analysis

11:00AM-12:00AM

My T. Thai
Plenary Talk: *Group Testing and Its Application in Biological Screens*

12:00PM-1:00PM Lunch Break

1:00PM-1:30PM

Azra Bihorac
Integrating data, algorithms and clinical reasoning

1:30PM-2:30PM

Xiaolin Li
Plenary Talk: *DeepHealth: Deep Learning and its Applications in Precision Medicine*

2:30PM-3:00PM Coffee Break

3:00PM-4:00PM

Giuseppe Nicosia
Plenary Talk: *Multi-Objective Multiplex Multi-Omic Genome-Scale Models for Cancer Metabolism*

4:00PM-6:00PM

Mario Guarracino
A novel semi-supervised classification technique and its application to transcriptional data
Vladimir D. Stozhkov
Random graph models for biological networks
Victoria Zdanovskaya
Value-at-Risk Support Vector Machines (Var-SVMs): Mixed Integer Programming (MIP) Representations
Pei-Li Wang
Reconstruction of Directed Acyclic Graphs networks based on prior causal ordering information with applications to gene regulatory networks

Friday, February 26$^{\text{th}}$

8:30AM-9:30AM

Ana Conesa
Plenary Talk: *Functional transcriptomics in the post-NGS era: multiomics integration and new technologies*

9:30AM-10:00AM

Oleg Prokopyev
Dynamic Abandon/Extract Decisions for Failed Cardiac Leads

10:00AM-10:30AM Coffee Break

10:30AM-11:00AM

Alla Kammerdiner
Computational studies of operations research models for sensor-based ranking risks of falls

11:00AM-12:00AM

Rebecka Jörnsten
Plenary Talk: *Network modeling of TCGA data: integrative and disease comparative approaches*

12:00PM-1:00PM Lunch Break

1:00PM-2:30PM

Awad Aljuaid
Neuroergonomics Study: Analysis of Brain's EEG Activity During Manual Lifting Tasks
Orestis P. Panagopoulos
Relaxed Support Vector Regression
Ramazan Ünlü
A weighted framework for unsupervised ensemble learning based on internal quality measures

2:30PM-3:00PM Coffee Break

3:00PM-4:00PM

Chris Sackellares
Plenary Talk: *Quantitative Conceptualization of the Epileptic Phenomenon*

4:00PM-5:30PM

Erick Moreno-Centeno
Guided Undersampling for Imbalanced Data Classification of Biomedical Data
Jiaxing Pi
Jointly sparse generalized eigenvalue classifier for feature selection
Anton Kocheturov
Nonlinear Dimensionality Reduction for Analysis of Electroencephalography Records

Saturday, February 27th

8:30AM-9:30AM

Jose C. Principe
Plenary Talk: *A Transient Model for Neuro-Modulation*

9:30AM-10:00AM

Oleksandra Yezerska
2-club clustering

10:00AM-10:30AM Coffee Break

10:30AM-11:00AM

Oleg Shylo
Optimizing Dynamic Interventions in Sleep Studies

11:00AM-12:00AM

Richard B. Berry
Plenary Talk: *Biomedical Computation - Sleep Medicine Opportunities*

12:00PM-1:00PM Lunch Break

Seeking biomarkers and understanding mechanisms: building an integrated approach to Alzheimer's disease

Madhav Thambisetty
NIH, USA.
thambisetty@mail.nih.gov

We have pursued numerous 'omics'-based approaches to identify predictive biomarkers of Alzheimer's disease (AD). In combination with neuroimaging, these methods have been successfully used to discover blood biomarkers reflecting core pathological features of AD. We have recently identified Alpha2 Macroglobulin (A2M) as a candidate serum biomarker predictive of incident AD in cognitively normal individuals. A2M is also associated with cerebrospinal fluid changes related to AD pathology. Using a network analysis, we then identified regulators of A2M gene expression and examined the relationship between peripheral and central A2M gene expression. Regulator of Calcineurin (RCAN1) drives A2M gene expression and inhibits calcineurin, a major brain tau phosphatase. Finally, using label free proteomics, we demonstrate that protein levels of A2M and calcineurin in the brain are positively correlated. These studies identify a novel molecular pathway linking the acute phase protein, A2M and neuronal injury in AD.

Signed Gene Set Analysis

James Henderson
University of Michigan, USA.
jbhender@umich.edu

With more than 200 types and multiple subtypes, cancer is a highly heterogeneous disease characterized by uncontrolled cell growth often thought to be caused by errors in DNA. Fortunately, high-throughput techniques for characterizing nucleic acids have enabled the collection of comprehensive genomic profiles from a large number of cancers and initiatives such as The Cancer Genome Atlas (TCGA) have made a large collection of these profiles available for analysis. A common use of such data is for genome-wide association studies in which one looks for differentially expressed genes between two or more groups, defined by a phenotype, genotype, or clinical outcome of interest. To reduce false discoveries and aid comprehension, it has become standard practice to carry out the analysis in terms of predefined and functionally-related gene sets rather than single genes using gene-set enrichment analysis (GSEA) or related techniques to compare the genomic profiles between groups. Here, we extend GSEA to work with signed gene sets in which each gene is augmented with a direction, i.e. plus or minus. This broadens the scope of GSEA by allowing genes with either cooperative or antagonistic function to be included in a single set. In a variety of settings, it also leads to increased statistical power as we demonstrate using Monte Carlo simulations. In addition, we discuss a number of possible sources for building a collection of signed gene sets. We conclude by presenting a signed gene set enrichment analysis of selected TCGA data.

Group Testing and Its Application in Biological Screens

My T. Thai
University of Florida, USA.
mytrathai@gmail.com

Group testing has various applications in blood testing, chemical leakage testing, coding, multi-access channel communication, and many others. In the context of biology, group testing is usually referred as pooling designs. As the technology for obtaining sequenced genome data is getting mature, more and more sequenced genome data are available to scientific research community, so that the study of gene functions has become a popular research direction. Such a study is supported by a high quality DNA library which usually is obtained through a large amount of testing and screening. Therefore, the efficiency of testing and screening becomes very important. Pooling design is a tool to reduce the number of tests in DNA library screening as well as in DNA microarrays. The construction of pooling design and its ability for decoding are very challenging, especially when inhibitors are present in the biological sample. Towards this end, we are tackling this problem in several layers, from distinguishing positive clones from negative clones to a more complicated model, considering errors and inhibitors.

Integrating data, algorithms and clinical reasoning

Azra Bihorac
University of Florida, USA.
ABihorac@anest.ufl.edu

In the United States, where the average American can expect to undergo seven surgical operations during a lifetime, each year at least 150,000 patients die and 1.5 million develop a medical complication within 30 days after surgery. Reducing postoperative complications (PC) by 20% could potentially save thousands of lives and significantly reduce healthcare costs. The risk for PC arises from the interactions between a patient's preoperative health and physiologic capacity to withstand surgery-related stress, modulated by the type and quality of surgery and anesthesia that the patient undergoes. Cost-effective strategies implemented in a timely fashion can ameliorate risk for PC or prevent their progression to more severe stages. The ability to implement such preventive strategies depends on the timely and accurate identification of patients at the greatest risk of PC. Assessment of surgical risk requires timely, accurate and dynamic synthesis of the large amount of clinical information throughout the continuum of perioperative care. Today we do not have the ability to accurately predict and quantify, for a given patient, a personal and real-time risk for PC that dynamically integrates preoperative risk with the risk incurred by events during surgery. Current surgical risk scores are limited to either a preoperative physician's subjective risk assessment or calculated scores with modest accuracy and limited usability. The interventions that could help prevent PC are applied without consideration of a patient's personal risk profile or not applied at all because the risk is underestimated. Paradoxically the abundant real-time physiologic, laboratory, and other clinical data streams are available in the perioperative electronic health records (EHR) but their magnitude and complexity often overwhelms physicians' ability to comprehend, retain, and organize the information in an optimal and timely way.

Our research is based on premises that clinical data contains unexplored richness sufficient for the development of automated and personalized diagnostic tools that currently do not exist. Our multidisciplinary team of experts in medicine and engineering has formed research partnership to address methodological challenges including real-time data integration and processing, data analytics and knowledge exchange between computers and physicians.Using technological advances in data science and engineering we are developing an intelligent perioperative system (IPS), composed of computers and physicians interacting in real time, which can generate usable medical knowledge with both increased speed and accuracy using complex clinical data obtained in the perioperative period.

DeepHealth: Deep Learning and its Applications in Precision Medicine

Xiaolin Li
University of Florida.
andyli@ece.ufl.edu

In recent years, Deep Learning (DL) has attracted tremendous enthusiasm in both academia and industry, winning numerous competitions in computer vision, natural language processing, and speech recognition. The key breakthrough in DL was due to a series of improvements in artificial neural network, machine learning, big data, and big systems. Researchers have developed large-scale deep learning models with billion to trillion parameters on tens thousands of CPU cores and GPUs. In this tutorial, we will give a broad overview of the state-of-the-art in deep learning: deep neural networks, deep convolutionary neural networks, deep recurrent neural networks, and enhanced models. We will also present representative applications in personalized precision medicine: sensing, monitoring, diagnosis, and prediction for bipolar, cancers, sepsis, and others.

Multi-Objective Multiplex Multi-Omic Genome-Scale Models for Cancer Metabolism

Giuseppe Nicosia
University of Catania, Italy.
nicosia@dmi.unict.it

Cancer cells were found to have a specific metabolism that is remarkably different from the tissues from which they originated, due to their high demand for proteins, lipids, nucleotides and energy levels, all necessary for speed-up and enhanced growth and proliferation.

The metabolic network is highly interconnected and complex, and probably the best characterized biological network in terms of models and multi-scale omic data. This help us to study the global goals and functional implications of the dysregulated metabolism in cancer.

In this talk he will show the effective integration of different sets of omic datasets, multi-objective optimization, multiplex and multi-scale genome-scale models to try to decipher the tumor metabolism.

A novel semi-supervised classification technique and its application to transcriptional data

Mario Guarracino
National Research Council of Italy, Italy.
mario.guarracino@cnr.it

Supervised classification is one of the most powerful data analysis technique, that uses *a-priori* information on the membership of data samples to classes. It is interesting to investigate semi-supervised algorithms that can produce classification models taking advantage of unlabeled samples. In this paper we analyze a novel technique that uses a Laplacian regularized eigenvalue classifier to produce models that are both accurate and parsimonious in terms of labeled data samples. We will provide examples from lRNA-seq experiments.

Random graph models for biological networks

Vladimir D. Stozhkov
University of Florida, USA.
vstozhkov@ufl.edu

In order to test effective algorithms on various biological networks, there is often a necessity to construct an appropriate random graph model.

Many researchers still use random graph models which are too simplified and do not approximate real-world biological networks.

In this talk we bridge this gap by showing what important features different types of biological networks exhibit and how to choose a correct random graph model that serves our goals.

Value-at-Risk Support Vector Machines (Var-SVMs): Mixed Integer Programming (MIP) Representations

Victoria Zdanovskaya
University of Florida, USA.
ladyvi@ufl.edi

SVMs is a widely used data classification technique. A class of Var-SVMs is known to be robust to the outliers in the training dataset. Unfortunately Var-SVM is a non-convex optimization problem. We consider MIP representations of Var-SVM, that can be solved by standard Branch&Bound algorithm. We also consider different techniques that help to dramatically improve computational performance of such formulations. Finally, we analyze the computational results of Var-SVM on the Liver Disorders data set.

Reconstruction of Directed Acyclic Graphs networks based on prior causal ordering information with applications to gene regulatory networks

Pei-Li Wang
University of Florida, USA.
peilicat@ufl.edu

Directed Acyclic Graphs (DAG) capture causal relations between random variables in various fields. However, their reconstruction (estimation) from observational data is a challenging task due to the computational hardness of the learning task. We study this estimation problem when prior information exists on partial orderings of sets of nodes. Specifically, it is assumed that the nodes are partitioned in different sets for which ordering information exists. We develop a framework for a high-dimensional sparse regime, that combines penalized regression to delineate relationships amongst nodes in different sets, with the popular PC-algorithm that identifies the skeleton of the graph within each set. In the final step, we combine the results from the regression and PC-algorithm steps and develop an adjustment step to eliminate redundant edges. The framework is evaluated on simulated data sets from the DREAM3 competition.

This is joint work with George Michailidis.

Functional transcriptomics in the post-NGS era: multiomics integration and new technologies

Ana Conesa
University of Florida, USA.
aconesa@ufl.edu

Next generation sequencing has speed up genome analysis and brought omics research closer to many organisms and biological scenarios. Today an increasing number of research projects propose the combined use of different omics platforms to investigate diverse aspects of genome and transcriptome function. However, combination of high-throughput data from different technologies in not straight forwards and it is not clear what the best way is to combine omics information to achieve interpretable muti-layer systems biology models. On the other hand, short read sequencing has shown to have intrinsic limitations to accurately describe the molecular and functional complexity of the transcriptomes of higher eukaryotes. I will present novel computational approaches in my lab to integrate omics technologies and use third generation sequencing platforms to deepen in the architecture, regulation and functionality of gene expression.

Dynamic Abandon/Extract Decisions for Failed Cardiac Leads

Oleg Prokopyev
University of Pittsburgh, USA.
droleg@pitt.edu

When a cardiac lead fails, physicians implant a new lead and may opt to extract the failed lead and/or any previously abandoned leads. Because the risk of extraction increases in lead age, physicians may extract leads to reduce the future risk of mandatory extraction, due to either infection or limited space in the vein. We develop discrete-time semi-Markov decision process models for various types of cardiac devices to determine patient-specific, lifetime-maximizing extraction policies as a function of patient age and the age of every implanted lead. We use clinical data to calibrate these models and present insightful numerical results, including comparisons to policies commonly used in practice.

This is a joint work with Anahita Khojandi (University of Tennessee), Lisa Maillart, Mark Roberts and Samir Saba (University of Pittsburgh)

Computational studies of operations research models for sensor-based ranking risks of falls

Alla Kammerdiner
New Mexico State University, USA.
alla@nmsu.edu

Falls represent a considerable public health problem. We study risk of falls by simulating an initial perturbation that creates a loss of balance in a laboratory environment. A system of wearable accelerometer sensors is used to record the information about motion of human body. We use these empirical data to evaluate the operations research models for sensor-based surveillance and risk assessment of falls. Rather than aiding in the traditional falls detection and overall risk assessment, the evaluated models aim at continuous monitoring of exposures to falls.

Network modeling of TCGA data: integrative and disease comparative approaches

Rebecka Jörnsten
Chalmers University, Sweden.
jornsten@chalmers.se

Statistical network modeling techniques have the potential to increase our understanding of cancer genomics data. Here, we analyze multiple TCGA data sets via a generalized sparse inverse covariance model, carefully addressing such challenges as unbalanced sample sizes, local network topology, model selection and robust estimation. The method integrates genetic, epigenetic and transcriptional data from multiple cancers, to define links that are present in multiple cancers, a subset of cancers, or a single cancer. The modeling results are available at cancerlandscapes.org, where the derived networks can be explored as interactive web content and be compared with several pathway and pharmacological databases. Network components are shown to fall in mainly two categories: common to all cancers or unique to one type of cancer. We also discuss how network models can be used to construct diagnostic markers (predictors of survival). This is joint work with the Nelander lab, SciLife, Uppsala.

Neuroergonomics Study: Analysis of Brain's EEG Activity During Manual Lifting Tasks

Awad Aljuaid
University of Central Florida, USA.
amjuaid@knights.ucf.edu

Electroencephalogram (EEG) has been confirmed as a reliable tool in health research due to the low cost and less body invasion. The application of EEG has a wide range of interest in cognitive and physical ergonomics. Psychophysical weight test has been used for decades to improve safe work practices by understanding human limitations in manual lifting. The aim of this study is to test the change on different EEG measures during various psychophysical lifting frequencies. High-density wireless dry cell EEG device has been used to record brain signals. Twenty healthy males participated in this experiment performing psychophysical weight lifting. After that subjects repeated the same experiment after two weeks. Analysis of variance ANOVA shows significant differences in different power bands between lifting frequencies at several brain areas.

Relaxed Support Vector Regression

Orestis P. Panagopoulos
University of Central Florida, USA.
opanagopoulos@knights.ucf.edu

Datasets with outliers pose a serious challenge in regression analysis. In this paper, a new regression method called relaxed support vector regression (RSVR) is proposed for such datasets. RSVR is based on the concept of constraint relaxation which leads to increased robustness in datasets with outliers. RSVR is formulated using both linear and quadratic loss functions. Numerical experiments on benchmark datasets and computational comparisons with other popular regression methods depict the behavior of our proposed method. RSVR achieves better overall performance than support vector regression (SVR) in measures such as RMSE and R^2_{adj} while performing on par with other state-of-the-art regression methods such as robust regression. Additionally, RSVR provides robustness for higher dimensional datasets, which is a limitation of robust regression. Furthermore, RSVR can be used on datasets that contain varying levels of noise.

A weighted framework for unsupervised ensemble learning based on internal quality measures

Ramazan Ünlü
University of Central Florida, USA.
ramazanunlu@knights.ucf.edu

Unsupervised ensemble, or consensus clustering, consists in nding the optimal combination strategy of individual clusterings that is robust with respect to the selection of an algorithmic clustering pool. Recently an approach was proposed based on the concept of consensus graph that has profound advantages over its predecesors. Despite its robust properties this approach assigns the same weight to the contribution of each clustering to the nal solution. In this paper, we propose a weighting policy for this problem that is based on internal clustering quality measures and compare against other popular approaches. Results on publicly available datasets show that weights can significantly improve the accuracy performance while retaining the robust properties.

Quantitative Conceptualization of the Epileptic Phenomenon

J. Chris Sackellares
EncephaloDynamics Corporation, USA.
jc.sackellares@gmail.com

Clinical neurophysiology has developed as a science, based on visual inspection of complex patterns of electrical potentials generated by the cerebral cortex. Major advances were the detection of the alpha wave, characteristic of the alert relaxed state, spatiotemporal patterns associated with specific sleep stages, discovery and classification of EEG patterns associated with various types of epileptic seizures, and recognition of patterns associated with diffuse and focal disturbances of brain function, as a consequence of a wide range of pathological conditions. In contrast to EEG, advances in the understanding of neuronal physiology have been approached by quantitative methods. Examples include the resting membrane potential, action potential, and synaptic transmission. The generation and resolution of seizure patterns in the EEG have been described based on visual inspection by expert clinicians. However, this approach does not provide an explanation as to why the patterns develop and resolve repetitively in the epileptic brain. Quantitative investigations into this phenomenon have provided preliminary insights, suggesting that these state transitions are similar to those arising in chaotic systems. However, controversy remains as to whether the brain is a nonlinear chaotic system. More recent evidence indicates that EEG over the surface of the brain can be modeled as a network, and that onset and propagation of seizure patterns are determined by connectivity and other network properties. In this presentation, we will demonstrate results from analysis of spatiotemporal dynamics of EEG patterns in human epileptogenic brain and the rodent chronic limbic epilepsy mode.

Guided Undersampling for Imbalanced Data Classification of Biomedical Data

Erick Moreno-Centeno
Texas A&M, USA.
emc@tamu.edu

Several biomedical datasets are highly imbalanced; that is, one class (the majority class, usually the healthy class) has significantly more instances than the minority class (usually the class of interest). Conventional classification methods have poor minority-class detection performance on imbalanced datasets since they tend to classify most (if not all) test data as majority instances. This talk presents an imbalanced-data classification method which combines a novel guided undersampling method and support vector machines. This method, which we call GU-SVM, outperforms several state-of-the-art imbalanced-data classification methods on a set of widely-used publicly-available imbalanced-datasets; most of them arising in the biomedical context.

Jointly sparse generalized eigenvalue classifier for feature selection

Jiaxing Pi
University of Florida, USA.
jiaxing@ufl.edu

Classification of high dimension and low sample size data has been studied extensively in biomedical applications. Feature selection methods has emerged as an important tool to reduce the dimensionality and improve the classification performance for high dimensional data sets. In this paper, we propose a new embedded feature selection method by combining a reformulated Generalized Eigenvalue Proximal Support Vector Machines (PSVMs) with the $L_{2,1}$-Norm regularization to achieve feature selection and classification simultaneously. An optimization method is provided to solve the formulation efficiently. Computational results on standard data sets are included to compare the proposed framework with other classification methods.

Nonlinear Dimensionality Reduction for Analysis of Electroencephalography Records

Anton Kocheturov
University of Florida, USA.
antrubler@gmail.com

We suggest using nonlinear dimensionality reduction technique called the Local Linear Embedding for analysis of EEG records. This approach enabled us to distinguish between different states of the brain in a more efficient way comparing to the existing machine learning techniques since it is faster and doesn't require training of the algorithm.

A Transient Model for Neuro-Modulation

Jose C. Principe
University of Florida, USA.
principe@cnel.ufl.edu

This talk proposes a novel approach to quantify brain activity taking into consideration the transient nature of the electroencephalogram (EEG) and local field potentials (LFP). In fact, the quantitative structure in the EEG/LFP extracted visually by clinicians relates to transients called bursts or spindles at selected frequencies called the EEG rhythms. Our framework proposes a noisy combination of filtered shot noise models that extract these events from a single channel of EEG/LFP using Matching Pursuit. The filters are defined a priori in a dictionary that spans the event shapes. Conceptually, the filters can also be learned from the data. We present results from rat hippocampal and human EEG data to illustrate the methodology.

2-club clustering

Oleksandra Yezerska
Texas A&M, USA.
yaleksa@tamu.edu

Graph clustering is a helpful tool in understanding complex systems and analyzing their structures and internal properties. A 2-club is a subset of vertices inducing a subgraph with a diameter at most 2. In this work, we study the minimum 2-club clustering problem, which is in partitioning the graph into the smallest number of 2-clubs, develop solution techniques and illustrate the approach on biological networks.

Optimizing Dynamic Interventions in Sleep Studies

Oleg Shylo
University of Tennessee, USA.
oshylo@utk.edu

We discuss a mathematical framework that takes advantage of the technological advances in wearable neuro-headsets to provide an objective, reliable, inexpensive and scalable approach to sleep interventions. This framework is based on semi-Markov decision models that rely on general signal processing methods for continuous sleep assessment.

Biomedical Computation - Sleep Medicine Opportunities

Richard B. Berry
University of Florida, USA.
Richard.Berry@medicine.ufl.edu

Sleep studies acquire a tremendous amount of digital data but currently the methods of analysis are crude and labor intensive. There are numerous opportunities for the development of computational tools to improve our understanding of human sleep and disorders such as sleep apnea. Several areas where biomedical computational techniques would be of great interest are discussed including sleep staging, automated scoring of sleep studies, detection of arousals, improved analysis of the oximetry signal, and estimates of sleep depth will be reviewed.