EURO-CBBM Workshop

Workshop on OR in Computational Biology, Bioinformatics and Medicine

Prague, Czech Republic, July 8, 2007

Location: The University of Economics, Prague
Room: RB (Rajska Building) 211

WORKSHOP PROGRAM

July 7, 2007, Saturday

19:00-22:00 Workshop Dinner (restaurant to be announced with an e-mail to all workshop participants, fee approximately 30 Euros/person.)

July 7, 2007, Sunday (all events in RB 211 at the University of Economics, Prague)

08:00-09:00 Workshop Registration

09:00-09:10 Welcome and Opening Remarks
Metin Türkay, Koç University

09:10-10:10 Invited Session #1, Session Chair: Jacek Blazewicz

Christodolous A. Floudas, Princeton University
Advances and Challenges in De Novo Protein Design

10:10-10:30 Coffee Break

10:30-11:30 Invited Session #2, Session Chair: Gerhard Wilhelm Weber

Ivet Bahar, University of Pittsburgh
Protein Dynamics and Allostery: Insights from Network Models

Dietrich Rebholz-Schuhmann, EBI, Cambridge, UK
Streaming facts from scientific publications to the scientist and new frontiers in publishing

12:30-13:30 Lunch Break

13:30-15:40 Contributed Session #1, Session Chair: Ceyda Oğuz

F.J. Planes[1,2] and J.E. Beasley[1], [1]Brunel University, [2]University of Navarra
Path finding approaches and metabolic pathways

Gunnar W. Klau, Free University of Berlin
Comparing Structural Information in the Life Sciences: From RNA to Metabolic Networks

Gerhard-Wilhelm Weber, Pakize Taylan, Zeynep Alparslan Gök, Başak Akteke Öztürk, Süreyya Özöğür, Ömür Uğur and Aysun Tezel, Middle East Technical University

A New Mathematical Approach in Environmental Protection: Gene-Environment Networks and Their Dynamics

Fadime Üney Yüksektepe and Metin Türkay, Koç University

Mixed-Integer Programming based Hyper-Box Enclosure Method for Predicting Folding Type of Proteins

Paola Bertolazzi, Giovanni Felici, Istituto di Analisi dei Sistemi e Informatica "Antonio Ruberti", Consiglio Nazionale delle Ricerche

Logic Mining Methods: introduction and applications to bioinformatics

David Gomez-Cabrero[1], Salva Ardid[2], Albert Compte[2], [1]Universitat de Valencia, [2]Institut d'Investigacions Biomèdiques August Pi i Sunyer

Exploring the specificity of the relationship between cortical network function and biological simulation parameters with a Particle Swarm Optimization algorithm.

15:40-16:00 Coffee Break

16:00-18:00 Contributed Session #2 - BIOPTRAIN Track, Session Chair: Jon Garibaldi

Alexandr Kovalev, Poznan University of Technology

Polynomial time approximation algorithms for the Simplified Partial Digest Problem

Andrea Sackmann, Poznan University of Technology

A DNA Computing algorithm for calculating the maxflow in networks

Linda Fiaschi, University of Nottingham

SNPs analysis in common diseases susceptibility

Daniele Soria, University of Nottingham

New ideas for clustering breast cancer data

Pawel Widera, University of Nottingham

Protein comparison in context of protein structure prediction

Invited Talks

Title: Advances and Challenges in *De Novo* Protein Design
Speaker: Professor Christodoulos A. Floudas
Department of Chemical Engineering, Princeton University
Abstract:

The primary objective in *de novo* protein design is to determine the amino acid sequences which are compatible with existing or postulated template backbone structures that may be rigid or flexible. The *de novo* protein design problem is of fundamental importance since it addresses the mapping of the space of amino acid sequences to known protein folds or postulated/putative protein folds. It is also of significant practical importance since it can lead to the improved design of inhibitors, design of novel sequences with better stability, design of catalytic sites of enzymes, and drug discovery.

The first part of this lecture will provide a motivation for the *de novo* protein design problem, a definition of the flexible backbone template structures, and an overview of the advances and limitations. The second part will introduce a novel two-stage approach which takes into account explicitly the flexibility of the templates. The first stage addresses **the *in silico sequence selection* problem** through two key contributions: (a) the development of a distance-dependent Ca-Ca and side chain centroid-centroid distance dependent force fields; and (b) a rigorous quadratic assignment-like formulation for the prediction of a rank-ordered list of sequences with novel mutations. The second stage addresses ***the fold specificity problem*** by performing structure prediction calculations using atomistic level force fields. Two alternative approaches will be presented for the generation of ensembles of protein conformations: (i) the first principles protein structure prediction approach, Astro-Fold, and (ii) an approach motivated by an established NMR structure refinement protocol. Based on the ensembles of protein structures generated, the probabilities of each predicted sequence to fold specifically to the flexible templates are calculated. The theoretical prediction results for several peptides and proteins that include variants of Compstatin, human beta defensin-2, C3a, and gp41 for HIV-1 will be presented. Comparisons with experimental findings will also be discussed.

Title: Protein Dynamics and Allostery: Insights from Network Models
Speaker: Professor Ivet Bahar
Department of Computational Biology and Bioinformatics, School o Medicine,
University of Pittsburgh
Abstract:

Elastic network models have been widely used in recent years for describing protein dynamics. The biomolecular structure is represented therein by a network of beads and springs to examine the collective dynamics of the system of harmonic oscillators. Many studies have shown that the most cooperative modes of motions accessible to biomolecular systems are relevant to their functional motions. Motivated by the success of elastic network models, we have exploited the utility of graph theory and spectral graph methods for exploring the pathways of communication in allosteric proteins. To this aim a Markovian diffusion of information across the structure is assumed and the hitting times between residue pairs are examined. Examination of the information processing characteristics of biomolecular structures reveal the efficient communication tendencies of catalytic and conserved residues, as well as the important role of tertiary contacts between secondary structural elements.

Title: Streaming facts from scientific publications to the scientist and new frontiers in publishing
Speaker: Dr. Dietrich Rebholz-Schuhmann
EBI, Cambridge, UK

Abstract:

Scientific literature is increasingly available in electronic form and early on after its acceptance for publication. Techniques to analyse the literature for contained facts are then applied to deliver individual facts directly to the scientist. This leads to the integration of the scientific literature into the infrastructure of existing IT data resources. This talk will explain how scientists in the biomedical domain profit from an infrastructure consisting of services for information extraction. All services automatically process the documents and interlink them with bioinformatics data resources. In addition they can be integrated into external IT solutions to directly couple experimental results with annotations from the scientific literature and to new solutions to support scientists in their pre-submission publication preparation process.

Contributed Talks

Title: Path finding approaches and metabolic pathways
Authors: *F.J. Planes*[1,2] and J.E. Beasley[1]
[1]Mathematical Sciences, Brunel University, Uxbridge, UB8 3PH, UK; [2]CEIT and TECNUN, University of Navarra, Manuel de Lardizabal 15, 20018 San Sebastian, Spain

Abstract:

The complex world of cellular metabolism has been commonly organised into metabolic pathways. A number of metabolic pathways, which are usually defined as a set of enzyme catalysed biochemical reactions by which a living organism transforms an initial source compound into a final target compound, have been elucidated via experiments on different model organisms. The problem of determining in a computational (algorithmic) fashion these metabolic pathways, given a database of biochemical reactions and compounds associated with a particular organism or cell, has attracted increased interest in recent years.

Path finding approaches to metabolic pathways adopt a graph theory approach to the problem of determining the reactions an organism might use to transform a source compound into a target compound. In this paper the effectiveness of using compound node connectivities in a path finding approach is examined. An approach to path finding based upon integer programming is also presented.

Author: Gunnar W. Klau
Free University Berlin
Title: Comparing Structural Information in the Life Sciences: From RNA to Metabolic Networks

Abstract:

Graphs and networks are powerful means of abstraction in the life sciences, and their comparison is an important task. In structural biology, for instance, we can model the space of possible secondary structures of a given RNA sequence as a graph and formulate the problem of aligning multiple RNA molecules with respect to an additive sequence-structure scoring function as a graph problem. Similar techniques are used in the area of structural proteomics. Here, contact map graphs are used to describe protein structures in order to classify them according to their structural similarities. In systems biology, the comparison of metabolic or

protein-protein interaction networks helps to understand cellular functions, evolutionary relationships, and disease mechanisms.

Unlike for classical sequence alignment, most of these problems are NP-hard---even in the pairwise case---as they involve, for instance, the computation of a maximum common subgraph. Nevertheless, techniques from mathematical programming may be used to compute optimal solutions for large instances quite efficiently.

Starting with the problem of aligning a set of RNA molecules, I will present a novel and unifying approach to the above comparison and alignment problems, which is based on an integer linear programming (ILP) formulation. Inspired by the work of Lancia and Caprara for the contact map overlap problem, we employ Lagrangian relaxation to find provably near-optimal solutions of the ILP in reasonable computation time. I will present results from an extensive computational study on benchmark alignments for the RNA case, which show that our approach outperforms alternative methods in terms of quality. Finally, I will show first results on the non-sequential contact map overlap problem and on the comparison of complex biological networks and stoichiometric matrices.

Author(s): *Gerhard-Wilhelm Weber*, Pakize Taylan, Zeynep Alparslan Gök, Başak Akteke Öztürk, Süreyya Özöğür, Ömür Uğur and Aysun Tezel
Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey

Title: A New Mathematical Approach in Environmental Protection: Gene-Environment Networks and Their Dynamics

Abstract:

A research area of central importance in computational biology, biotechnology - and life sciences at all - is devoted to modeling, prediction and dynamics of gene-expression patterns. However, as clearly understood in these days, this enterprise cannot be investigated in a satisfying way without taking into account the role of the environment in its widest sense. To a representation of past, present and most likely future states, the authors also acknowledge the presence of measurement errors and further uncertainties. This paper surveys and improves recent advances in understanding the mathematical foundations and interdisciplinary implications of the newly introduced gene-environment networks. Moreover, it integrates the important theme carbon dioxide emission reduction into the context of our networks and their dynamics. Given the data from DNA microarray experiments and environmental records, we extract nonlinear ordinary differential equations which contain parameters that have to be determined. This is done by some modern kinds of (so-called generalized Chebychev) approximation and (so-called generalized semi-infinite) optimization. After this is provided, time-discretized dynamical systems are studied. Here, a combinatorial algorithm constructing and following polyhedra sequences, allows to detect the region of parametric (in)stability. Finally, we analyze the topological landscape of gene-environment networks in its structural stability. With the example of CO2-emission control and some further perspectives we conclude.

This pioneering work is practically motivated and theoretically elaborated; it tries to support improvements in health care, medicine, education, more healthy living conditions and environmental protection. The authors discuss structural frontiers and invite the interested readers to future research.

Authors: *Fadime Uney Yuksektepe* and Metin Turkay

Title: Mixed-Integer Programming based Hyper-Box Enclosure Method for Predicting Folding Type of Proteins

Abstract:

A novel integer programming based hyper-box enclosure method is presented in order to predict the folding types of proteins. Proteins are mainly categorized in four different classes depending on their secondary structure content. In addition, it is shown that the folding types of proteins are correlated to their amino acid compositions. Traditional approaches using hyperplanes, such as support vector machines, that are very effective in classifying data sets into two groups perform poorly in multi-class problems. A novel hyper-box enclosure method is developed to overcome difficulties and inefficiencies of these approaches. The method uses Boolean variables to define the boundaries of the hyper-boxes that include all or some of the points in that class.

The efficiency of the proposed approach is illustrated on a benchmark problem that includes a training set of 120 proteins (30 from each group). The self-consistency and cross-validation test results are also performed. Moreover, the same data set is studied by different classifiers of the software WEKA. The results show that the proposed method has a better classification accuracy on this data set compared other methods.

Author(s): *Paola Bertolazzi*, Giovanni Felici
Istituto di Analisi dei Sistemi e Informatica “Antonio Ruberti” Consiglio
Nazionale delle Ricerche

Title: Logic Mining Methods: introduction and applications to bioinformatics

Abstract:

We present a particular class of data mining and machine learning techniques that can be applied to solve Classification, Clustering and Feature Selection problems. Such techniques are based on the use of logic variables and logic models that represent information and explanatory models.

We show how these methods can be fruitfully adopted in bioinformatics applications. We describe effective methods to extract hidden information from large data sets, with particular focus on the integer programming models and algorithms on which they are based.

Finally, several bioinformatics applications of the described data mining and machine learning techniques concerning automatic classification of microarray data, DNA sequence classification, feature selection from protein sequences, and Barcode analysis, will be presented.

Authors: *David Gomez-Cabrero*[1], Salva Ardid[2], Albert Compte[2]
[1]david.gomez@uv.es (speaker)
Departamento de Estadística e Investigación Operativa, Universitat de Valencia, Spain.
[2]jsardid@umh.es, acompte@clinic.ub.es Institut d'Investigacions Biomèdiques
August Pi i Sunyer (IDIBAPS) Carrer Villarroel 170, 08036 Barcelona

Title: Exploring the specificity of the relationship between cortical network function and biological simulation parameters with a Particle Swarm Optimization algorithm.

Abstract:

Mechanistic aspects of brain function can be studied with the use of biologically detailed computational models. Often, a critical question that arises from these computational models is how critically the conclusions depend on a particular choice of the simulation parameters. It is difficult to establish that the presented network model is unique in producing the relevant phenomenology. This issue has been approached before for the case of single neurons or small networks of neurons.

Here, we design a computational strategy to explore this for the case of large-scale biological neural network simulations. We focus on a specific neural function: visuo-spatial working memory, and we construct a biological neural network that mimics the cortical network.

We search for sets of parameters such that the network sustains persistent activity. We design several evaluation functions that quantify this ability and weigh them in a proper way. To guide the search we rely on the Particle Swarm Optimization. The first objective is to find if there exists a unique solution or a set of significant different solutions. In the second case, we explore and typify the different areas of solutions; for this second objective we rely on different neighbor policies among the particles.

Contributed Talks, BIOPTRAIN Track

Author: Alexandr Kovalev
Poznan University of Technology
Title: Polynomial time approximation algorithms for the Simplified Partial Digest Problem

Abstract:

The Simplified Partial Digest Problem (SPDP) is a mathematical model for a recently proposed simplified partial digest method of genome mapping. This method is easy for laboratory implementation and robust with respect to the experimental errors. SPDP is NP-hard in the strong sense for the case of measurement errors and for the error-free case. One way to tackle this problem is to develop efficient (polynomial time) approximation algorithms. I will present two optimization versions of the original problem, which are called SPDP-Min and SPDP-Max, and give results on the worst-case efficiency of approximation algorithms for these problems. Then I will describe a graph-theoretic model for SPDP-Min and SPDP-Max, which can be used to reduce the search space for an optimal solution in either of these problems. I will also outline three heuristic polynomial time algorithms based on the graph-theoretic model. The results of computer experiments with these algorithms on randomly generated data and real data from GenBank will be given.

Author: Andrea Sackmann
Poznan University of Technology
Title: A DNA Computing algorithm for calculating the maxflow in networks

Abstract:

The main idea of DNA Computing is to encode information as DNA sequences and use known bimolecular techniques (based on the chemical properties of DNA) for manipulating the molecules to solve computational problems. The approach's inherent value is its massive computational parallelism and its power of large database searching. The concept of DNA Computing was introduced in 1994 to solve the Hamiltonian path problem in a directed graph. Since then various approaches have been developed dealing with mainly graph theoretical but also other mathematical problems (including NP hard ones). An effort of the current research optimizes the procedure on the one hand for example by reducing the external interference by utilizing biochemical reactions which autonomously process computations. And on the other hand to optimize the designing of the sequences used.

To the best of my knowledge an application of DNA Computing for solving net work problems has not been proposed yet. In the presented work I introduce an algorithm based on DNA Computing to solve the maxflow problem of a given network N = (G; c; s; t). Here, G is a directed graph G = (V;E) with weighted edges and these weights c denote the capacity of each edge. The vertices s and t are the source, and the sink of N, respectively. According to the Maxflow min-cut theorem the maximum value of a flow on a network is equal to the minimal capacity of a network's cut. Thus, the structure of the algorithm is as follows:

1. Building templates for the vertices,
2. Generating random partitions (*S; T*) of *V* (as the combinatorial library),

3. Finding the capacities of the cuts of *N*,
4. Detecting a minimal cut and read out its capacity.

Finding suitable DNA sequences and carrying out the corresponding biological experiments are subject of my future research.

Author: Linda Fiaschi
University of Nottingham
Title: SNPs analysis in common diseases susceptibility
Abstract:

Many common human diseases are caused by genetic variations within a single gene or influenced by complex interactions among multiple genes as well as environmental and lifestyle factors. It is currently difficult to measure and evaluate environmental and lifestyle effects on a disease process, therefore my study is focused on the analysis of individual predisposition to develop a disease based on genes and hereditary factors.

In this talk I will give an overview of my specific task within the BIOPTRAIN project which consists in the SNPs analysis applied to Alzheimer and Pre-eclampsia diseases.

Author: Daniele Soria
University of Nottingham
Title: New ideas for clustering breast cancer data
Abstract:

In the talk I will give an overview of my PhD project, pointing out the overall objectives, the work done so far and the main goals for the future. I have applied different clustering techniques on a large dataset of tissue microarray information of invasive breast cancer data.

Now I am trying to solve the problem of assigning each patient to a specific cluster with a certain probability, aiming on finding an automatic algorithm for it. I think this method could be especially useful for the new data the Nottingham City Hospital will provide me.

Author: Pawel Widera
University of Nottingham
Title: Protein comparison in context of protein structure prediction
Abstract:

Although research in protein structure prediction has been continued for over 30 years, not many optimisation techniques has been applied to that field yet. Moreover, measuring the quality of prediction also remains problematic, despite of over 10 years of constant progress in CASP experiment.

In the talk I would like to introduce the problem of protein structure comparison and it's relevance to the ab initio protein structure prediction, sharing my thoughts about possible improvements of state of the art methods used in CASP.