Discrete Applied Mathematics

Special Issue

Networks in Computational Biology

Dedicated to our dear teacher and friend
Prof. Dr. Peter Ladislaw Hammer (1936-2006)

Editorial

This special issue on *Networks in Computational Biology* is based on a workshop at Middle East Technical University in Ankara, Turkey, September 10-12, 2006 (http://www.iam.metu.edu.tr/Networks_in_Computational_Biology/). Computational biology is one of the many currently emerging areas of applied mathematics and science. During the last century, cooperation between biology and chemistry, physics, mathematics and other sciences increased dramatically, thus providing a solid foundation for, and initiating an enormous momentum in, many areas of the life sciences. This special issue focuses on *networks*, a topic that is equally important in both, biology and mathematics, and presents snapshots of current theoretical and methodological work in network analysis. Both, discrete and continuous optimization, dynamical systems, graph theory, pertinent inverse problems, and data mining procedures are addressed. The principal goal of this special issue is to contribute to the mathematical foundation of computational biology by stressing its particular aspects relating to network theory.

This special issue consists of 25 articles, written by 65 authors and rigorously reviewed by 70 referees. The guest editors express their cordial thanks to all of them, as well as to the Editors-in-Chief of Discrete Applied Mathematics, Prof. Dr. Endre Boros and his predecessor, Prof. Dr. Peter L. Hammer who was one of the initiators of this special issue but left us in 2006, and to Mrs. Katie D'Agosta who was at our side in each phase of preparation of this DAM special issue.

The articles are ordered according to their contents. Let them shortly be summarized:

In the paper of Jacek Błażewicz, Dorota Formanowicz, Piotr Formanowicz, Andrea Sackmann and Michał Sajkowski, entitled *Modeling the Process of Human Body Iron Homeostasis Using a Variant of Timed Petri Nets*, the standard model of body iron homeostasis is enriched by including the durations of the pertinent biochemical reactions. A Petri-net variant in which, at each node, a time interval is specified, is proposed in order to describe the time lag of the commencement of conditions that must be fulfilled before a biochemical reaction can start.

Due to critical changes in the environment, switches can occur in metabolic networks that lead to systems exhibiting simultaneously discrete and continuous dynamics. Hybrid systems represent

this accurately. The paper *Modeling and Simulation of Metabolic Networks for Estimation of Biomass-Accumulation Parameters* by Uğur Kaplan, Metin Türkay, Bülent Karasözen, and Lorenz Biegler develops a hybrid system to simulate cell-metabolism dynamics that includes the effects of extra cellular stresses on metabolic responses.

Path-finding approaches to metabolic-pathway analysis adopt a graph-theoretical approach to determine the reactions that an organism might use to transform a source compound into a target compound. In the contribution *Path-Finding Approaches and Metabolic Pathways*, Francisco J. Planes and John E. Beasley examine the effectiveness of using compound-node connectivities in a path-finding approach. An approach to path finding based on integer programming is also presented. Existing literature is reviewed. This paper is well illustrated and provides many examples as well as, as an extra service, some supplementary information.

In *A New Constraint-Based Description of the Steady-State Flux Cone of Metabolic Networks*, Abdelhalim Larhlimi and Alexander Bockmayr present a new constraint-based approach to metabolic-pathway analysis. Based on sets of non-negativity constraints, it uses a description of the set of all possible flux distributions over a metabolic network at a steady state in terms of the steady-state flux cone. The constraints can be identified with irreversible reactions and, thus, allow a direct interpretation. The resulting description of the flux cone is minimal and unique. Furthermore, it satisfies a simplicity condition similar to the one for elementary flux modes.

Most biological networks share some properties like being, e.g., “scale free”. Etienne Birmelé proposes a new random-graph model in his contribution *A Scale-Free Graph Model Based on Bipartite Graphs* that can be interpreted in terms of metabolic networks and exhibits this specific feature.

Differential equations have been established to quantitatively model the dynamic behaviour of regulatory networks representing interactions between cell components. In the paper *Inference of an Oscillating Model for the Yeast Cell Cycle*, Nicole Radde and Lars Kaderali study differential equations within a Bayesian setting. First, an oscillating core network is learned that is to be extended, in a second step, using “Bayesian” methodology. A specifically designed hierarchical prior distribution over interaction strengths prevents overfitting and drives the solutions to sparse networks. An application to a real-world data set is provided, and its dynamical behaviour is reconstructed.

The contribution *An Introduction to the Perplex Number System* by Jerry L.R. Chandler derives from his approach to theoretical chemistry, and provides a universal source of diagrams. The perplex number system, a new tool for describing relationships between concrete objects and processes, provides in particular an exact notation for chemistry without invoking “alchemical symbols”. Practical applications to concrete compounds (e.g., isomers of ethanol and dimethyl ether) are given. In conjunction with the real number system, the relations between perplex numbers and scientific theories of concrete systems (e.g., intermolecular dynamics, molecular biology, and individual medicine) are described.

Since exact determination of haplotype blocks is usually impossible, a method is desired which can account for recombinations, especially, via phylogenetic networks or a simplified version. In

their work *Haplotype Inferring via Galled-Tree Networks Using a Hypergraph-Covering Problem for Special Genotype Matrices*, Arvind Gupta, Ján Maňuch, Ladislav Stacho, and Xiaohong Zhao reduce the problem via galled-tree networks to a hypergraph-covering problem for genotype matrices satisfying a certain combinatorial condition. Experiments on real data show that this condition is mostly satisfied when the minor alleles (per SNP) reach at least 30%.

Recently the Quartet-Net or, for short. "QNet" method was introduced by Stefan Grünewald *et al* as a method for computing phylogenetic split-networks from a collection of weighted quartet trees. Here, Stefan Grünewald, Vincent Moulton, and Andreas Spillner show that QNet is a "consistent" method. This key property of QNet not only guarantees to produce a tree if the input corresponds to a tree --- and an outer-labeled planar split network if the input corresponds to such a network. The proof given in their contribution *Consistency of the QNet Algorithm for Generating Planar Split Networks from Weighted Quartets* also provides the main guiding principle for the design of the method.

Kangal and Akbash dogs are the two well-known shepherd dog breeds in Turkey. In the article *Genetic Relationship between Kangal, Akbash and other Dog Populations*, Evren Koban, Çigdem Gökçek Saraç, Sinan Can Açan, Peter Savolainen, and İnci Togan present a comparative examination by mitochondrial DNA control region, using a consensus neighbour-joining tree with bootstrapping which is constructed from pairwise F_{ST} values between populations. This study indicates that Kangal and Akbash dogs belong to different branches of the tree, i.e., they might have descended maternally from rather different origins created by an early branching event in the history of the domestic dogs of Eurasia.

In their contribution *The Asian Contribution to the Turkish Population with respect to the Balkans: Y-Chromosome Perspective***,** Ceren Caner Berkman and İnci Togan investigate historical migrations from Asia using computational approaches. Chikhi *et al*'s admixture method was used to estimate the male genetic contribution of Central Asia to hybrids. The authors observed that the male contribution from Central Asia to Turkish population with reference to the Balkans was 13%. Comparison of the admixture estimate for Turkey with those of neighboring populations pointed out that the Central Asian contribution was lowest in Turkey.

Split-decomposition theory deals with relations between real-valued split systems and metrics. In his work *Split Decomposition over an Abelian Group Part 2: Group-Valued Split Systems with Weakly Compatible Support*, Andreas Dress uses a general conceptual framework to study these relations from an essentially algebraic point of view. He establishes the principal results of split-decomposition theory regarding split systems with weakly compatible support within this new algebraic framework. This study contributes to computational biology by analyzing the conceptual mathematical foundations of a tool widely used in phylogenetic analysis and studies of bio-diversity.

The contribution *Phylogenetic Graph Models beyond Trees* of Ulrik Brandes and Sabine Cornelsen is a phylogenetic analysis is to describe the relationship between species. The authors demonstrate that the phylogenetic tree model can be generalized to a cactus (i.e., a tree of edges and cycles) without losing computational efficiency. A cactus can represent a quadratic rather than linear number of splits in linear space. They show how to decide in some linear time

whether a set of splits has a cactus model and, in that case, how to construct it within the same time bounds. Finally, the authors briefly discuss further generalizations of tree models.

In their paper *Whole-Genome Prokaryotic Clustering Based on Gene Lengths*, Alexander Bolshoy and Zeev Volkovich present a novel method of taxonomic analysis constructed on the basis of gene content and lengths of orthologous genes of 66 completely sequenced genomes of unicellular organisms. They cluster given input data using an application of the information bottleneck method for unsupervised clustering. This approach is not a regular distance-based method distinguishing it from other recently published whole-genome based clustering techniques. This study correlates well with the standard "tree of life".

The approach of Limor Kozobay-Avraham, Sergey Hosida, Zeev Volkovich and Alexander Bolshoy in their article *Prokaryote Clustering Based on DNA Curvature Distributions* compares genomes according to typical distributions of a function respresenting some biological function. Due to an extensive amount of data they were able to define the factors influencing the curvature distribution in promoter and terminator regions such as growth temperature, genome size, and A+T composition. The clustering methods K-means and PAM produced a very similar clustering, reflecting genomic attributes and environmental conditions of the species' habitat.

The paper *Pattern Analysis for the Prediction of Fungal Pro-peptide Cleavage Sites* of Süreyya Özögür Ayzüz, John Shawe-Taylor, Gerhard-Wilhelm Weber and Zümrüt B. Ögel applies support vector machines to predict the pro-peptide cleavage site of fungal extracellular proteins displaying mostly a monobasic or dibasic processing site. A specific kernel is expressed as an application of the Gaussian kernel via feature spaces defined. The novel approach simultaneously performs model selection together with the testing of accuracy and confidence levels. The results are found to be accurate and compared with the ones of the server ProP1.0.

Preetam Ghosh, Samik Ghosh, Kalyan Basu and Sajal Das adopt an "in silico" stochastic event based simulation methodology to determine the temporal dynamics of different molecules. In their paper *Parametric Modeling of Protein-DNA Binding Kinetics: A Discrete Event Based Simulation Approach*, they present a parametric model to predict the execution time for the biological function protein-DNA binding. It considers the actual binding mechanism along with some approximated protein and DNA structural information. A collision theory based approach is used. This model brings the important biological parameters and functions into consideration.

Murat Ali Bayır, Tacettin Doğacan Güney and Tolga Can in their paper *Integration of Topological Measures for Eliminating Non-Specific Interactions in Protein Interaction Networks* propose a novel technique based on integration of topological measures for removing non-specific interactions in a large-scale protein-protein interaction network. After transforming into a line graph, they compute clustering coefficient and betweenness centrality measures for all the edges in the network. The authors use confidence estimates and validate their proposed method by comparing the results of a molecular complex detection to some ground truth set.

In their article *Graph Spectra as a Systematic Tool in Computational Biology* by Anirban Banarjee and Jürgen Jost is about the natural question how biological content can be extracted from the graphs to which the biological data are reduced. It advocates a set of graph invariants that can be easily graphically represented and therefore visually analysed and compared, but

which also yields an essentially complete qualitative characterization of a graph. This is the spectrum of the graph Laplacian. The authors apply this method to the study of protein-protein interaction networks. Contributions from literature are discussed and examples offered.

A Note on Fundamental, Non-fundamental, and Robust Cycle Bases is a study to achieve robustness in biological systems by redundant wiring reflected by the presence of cycles in the graphs connecting the systems' components. The authors Konstantin Klemm and Peter F. Stadler introduce cyclically robust cycle bases to generate all cycles of a given 2-connected graph by iteratively adding basis cycles. Strictly fundamental bases are not necessarily cyclically robust; conversely, cyclically robust are not necessarily fundamental. Furthermore, they present a class of cubic graphs for which cyclically robust bases can be explicitly constructed.

Understanding interplay and function of a system's components also requires the study of the system's functional response to controlled experimental perturbations. For biological systems, it is problematic with an experimental design to aim at a complete identification of the system's mechanisms. In his contribution *Stochastic Dependence and Causal Information Flows*, Nihat Ay employs graph theory and studies the interplay between stochastic dependence and causal relations within Bayesian networks and by information theory. Applying a causal information flow measure provides a quantitative refinement of Reichenbach's common cause principle.

Axel Mosig, Türker Bıyıkoğlu, Sonja J. Prohaska and Peter F. Stadler by their article *Discovering Cis-Regulatory Modules by Optimizing Barbecues* ask for simultaneously stabbing a maximum number of differently coloured intervals from *K* arrangements of coloured intervals. A decision version of this *Best Barbecue Problem* is shown to be NP-complete. Because of the relevance for complex regulatory networks on gene expression in eukaryotic cells, they propose algorithmic variations that are suitable for the analysis of real data sets comprising either many sequences or many binding sites. The studied optimization problem generalizes frequent itemset mining.

The contribution *A Mathematical Program to Refine Gene Regulatory Networks* of Guglielmo Lulli and Martin Romauch proposes a methodology for making sense of large, multiple time-series data sets arising in expression analysis. It introduces a mathematical model to release a reduced and coherent regulatory system, provided a putative regulatory network. Two equivalent formulations of the problem are given and the NP-completeness is proved. For solving large-scale instances the authors implemented an Ant Colony Optimization procedure. The proposed algorithm is validated by a computational analysis on randomly generated test instances.

Jutta Gebert, Nicole Radde, Ulrich Faigle, Julia Ströosser and Andreas Burkovski in their paper *Modelling and Simulation of Nitrogen Regulation in Corynebacterium Glutamicum* aims at understanding and predicting the interactions of macromolecules inside the cell. It derives a theoretical model for biochemical networks and introduce a general method for the parameter estimation, applicable in the case of very short time series. This approach is applied to a special system concerning the nitrogen uptake. The equations are set up for the main components of this system, the optimization problem is formulated and solved, and simulations are carried out.

Gerhard-Wilhelm Weber, Ömür Uğur, Pakize Taylan and Aysun Tezel mathematically model and prediction of gene-expression patterns with a rigorous introducing of the environment and aspects of errors and uncertainty. For this purpose, in their paper *On Optimization, Dynamics and*

Uncertainty: a Tutorial for Gene-Environment Networks they employ Chebyshev approximation and generalized semi-infinite optimization. Then, time-discretized dynamical systems are studied the region of parametric stability is detected by a combinatorial algorithm, before the topological landscape of gene-environment networks becomes analyzed in terms of structural stability.

We are convinced that all papers presented in this special issue not only satisfy the high standards of DAM, but also constitute important contributions to many different areas in Computational Biology, to Discrete Applied Mathematics and its related fields. For their valuable help we thank again all the colleagues who have participated in this exciting endeavour with their care, foresight and vision!

Andreas Dress, Bülent Karasözen, Peter F. Stadler and Gerhard-Wilhelm Weber
(Guest Editors)

Prof. Dr. Andreas Dress
Max-Planck Institute for Mathematics in the Sciences,
Leipzig, Germany, and
CAS-MPG Partner Institute on Computational Biology,
Shanghai, China
E-mail: dress@mis.mpg.de, dress@cfc.nankai.edu.cn

Prof. Dr. Bülent Karasözen
Department of Mathematics and
Institute of Applied Mathematics
Middle East Technical University (METU, ODTÜ)
Ankara, Turkey
E-mail: bulent@metu.edu.tr

Prof. Dr. Peter F. Stadler
Institut für Informatik and IZBI
Universitaet Leipzig
Leipzig, Germany
E-mail: Peter.Stadler@bioinf.uni-leipzig.de

Prof. Dr. Gerhard-Wilhelm Weber
Institute of Applied Mathematics
Middle East Technical University (METU, ODTÜ)
Ankara, Turkey
E-mail: gweber@metu.edu.tr

Assistant to the guest editors: Mrs. Cand. MSc. Bengisen Pekmen (Institute of Applied Mathematics, METU, Ankara)