

## Table of Contents

### Session I. Introduction

Welcome

*Raj Reddy*

Overview of Biological Language Conference

*Judith Klein-Seetharaman and Roni Rosenfeld*

Challenges and Opportunities for Biological Language Modeling in Biomedical High-throughput Genomic and Proteomic Informatics ..... 3

*Jim Lyons-Weiler*

### Session II. Sequence structure mapping

Five hierarchical levels of sequence-structure correlation in proteins..... 7

*Chris Bystroff, Yu Shao and Xin Yuan*

Varieties of Meaning in Protein Sequences: The Nature of the Folding Code..... 22

*Igor B. Kuznetsov and S. Rackovsky*

Prediction of parallel and anti-parallel beta-sheets using Conditional Random Fields..... 35

*Yan Liu, Jaime Carbonell, Judith Klein-Seetharaman and Vanathi Gopalakrishnan*

Yule value tables from protein datasets of different categories: emphasis on membrane proteins..... 44

*Madhavi Ganapathiraju, Deborah Weisser, Raj Reddy and Judith Klein-Seetharaman*

### Session III. Microarray

Computational Challenges in the Analysis of Time Series Expression Data ..... 61

*Ziv Bar-Joseph*

Constructing Genetic Networks from Microarray Data Using Recursive Ridge Regression..... 74

*Fan Li, Yiming Yang, Gulsum Anderson, Judith Klein-Seetharaman, Venkatesh Jitender, Vanathi Gopalakrishnan*

An Analysis of Recursive Feature Elimination Methods for Statistical Classification..... 84

*Fan Li, Yiming Yang and Jian Zhang*

Word Similarity Based Phylogenetic Reconstruction for Validating Micro-arrays Reuse in Closely Related Species 91

*Deepika Jagan and Gautam B. Singh*

### Session IV. Function prediction

Document Classification of Protein Sequences ..... 102

*Betty Cheng, Jaime Carbonell and Judith Klein-Seetharaman*

A Sequence Alignment-Independent Method For Protein Classification..... 123

*John Vries, Rajan Munshi, Dror Tobi, Judith Klein-Seetharaman, Panayiotis V. Benos and Ivet Bahar*

A study of statistical methods for function prediction of protein motifs..... 147

*Tao Tao, ChengXiang Zhai, Xinghua Lu and Hui Fang*

Functional Inference Properties of Protein N-grams ..... 162

*Vijay Ilavarasan, Jay Wren and Gautam B. Singh*

Annotating the proteome..... 173

*M. Couch, M. Paterson and M.B. Swindells*

Improved Gene Prediction through Human/Mouse Ortholog Similarity Constraints..... 181

*Stephanie Seneff, Chao Wang and Chris Burge*

### Session V. Protein dynamics and folding

Extracting hydrogen-bond signature patterns from protein structure data..... 196

*Tejasvini Prasad, Tamilselvi Subramanian, Sridhar Hariharaputran, Chaitra H.S. and Nagasuma Chandra*

Structural features in the model of a thermostable and stress-resistant protein, SP1 from aspen..... 212

*Ravindranath S. Rathore and T. Narasimhamurthy*

Folding dynamics of small proteins with different types of structures..... 218

*Seung-Yeon Kim, Julian Lee and Jooyoung Lee*

The folding core in rhodopsin ..... 229

*AJ Rader, Gulsum Anderson, Basak Isin, Judith Klein-Seetharaman and Ivet Bahar*

Natively Disordered Protein: Functions and Predictions..... 245

*Pedro Romero, Zoran Obradovic, and A. Keith Dunker*

## Session VI. Best idea competition

*All students and Postdocs*

## Session VII. Generation and analysis of multiple sequence alignments

Constructing Sequence Alignments from a Markov Decision model with estimated parameter values..... 259

*Fern Y. Hunt, Anthony J. Kearsley and Agnes O'Gallagher*

Inferring property selection pressure from positional residue conservation..... 271

*Rose Hoberman, Roni Rosenfeld and Judith Klein-Seetharaman*

MSAT – a Multiple Sequence Alignment tool based on TOPS..... 286

*Te Ren, Mallika Veeramalai, Aik Choon Tan and David Gilbert*

Position Dependent and Independent Evolutionary Models Based on Empirical Amino Acid Substitution Matrices 302

*Bernardo Barbiellini, Alexandra Portnova, Anna Chetoukhina and Chia-Hsin Lu, Robert Grothe and Matteo*

*Pellegrini*

## Session VIII. DNA sequence analysis

DNA Sequence Analysis Linguistic Tools: Words and Patterns..... 317

*Alexander Bolshoy*

Haplotype Parsing: Methods for Extracting Information from Human Genetic Variations..... 325

*Russell Schwartz*

Profile Hidden Markov Dirichlet-Multinomial Models for Motif Families..... 340

*Eric P. Xing*

## Session IX. Toolkits

Extensions to the Biological Language Modeling Toolkit (BLMT)..... 353

*Sivaraman Balakrishnan, Madhavi Ganapathiraju, Judith Klein-Seetharaman, N. Balakrishnan and Raj Reddy*

Biological Language Modeling Toolkit Web-Interface..... 366

*Viji Manoharan, Danny Lam and Judith Klein-Seetharaman*

IPSVAC – Integrated Platform for Sequence Visualization, Analysis and Comparison..... 374

*Naveena V.K. Yanamala, Harikrishna V. Rekapalli, Sri Jyothsna Yeleswarapu, Ram Sateesh Talari and Abhijit Mitra*

The Matlab Bioinformatics Toolkit..... 387

*Lucio Cetto and Rob Henson*

## Session X. Interdisciplinary education

Cognitive Assessment Driven Strategy for Integrating Bioinformatics Concepts in Computer Science Curricula..... 390

*Gautam B. Singh and Christine Hansen*

Biomedical Problem Solving Environment..... 398

*Yang Cai, Ingo Snel, Clementine Klein, Suman Bharathi Balasubramaniam, Betty Cheng and Judith Klein-*

*Seetharaman*

## Session XI. Best idea competition award ceremony