BMC Bioinformatics

Introduction

Open Access

Probabilistic modeling and machine learning in structural and systems biology

Samuel Kaski*[1], Juho Rousu[2] and Esko Ukkonen[2]

Address: [1]Adaptive Informatics Research Centre and Helsinki Institute for Information Technology, Laboratory of Computer and Information Science, Helsinki University of Technology, P.O. Box 5400, FI-02015 TKK, Finland and [2]Helsinki Institute for Information Technology, Department of Computer Science, University of Helsinki, P.O. Box 68, FI-00014 University of Helsinki, Finland

Email: Samuel Kaski* - samuel.kaski@tkk.fi; Juho Rousu - juho.rousu@cs.helsinki.fi; Esko Ukkonen - esko.ukkonen@cs.helsinki.fi

* Corresponding author

from Probabilistic Modeling and Machine Learning in Structural and Systems Biology
Tuusula, Finland. 17–18 June 2006

Published: 3 May 2007

BMC Bioinformatics 2007, **8**(Suppl 2):S1 doi:10.1186/1471-2105-8-S2-S1

This article is available from: http://www.biomedcentral.com/1471-2105/8/S2/S1

Abstract

This supplement contains extended versions of a selected subset of papers presented at the workshop PMSB 2007, Probabilistic Modeling and Machine Learning in Structural and Systems Biology, Tuusula, Finland, from June 17 to 18, 2006.

Introduction

The workshop was designed to gather together researchers working on the extremely timely task of integrating advanced machine learning and computational modeling with current biological and medical research problems. The field is particularly active because the new high-throughput measurement techniques and biological databases require advanced modeling methods but the field progresses so rapidly that the models need to be flexible and relatively general-purpose. Modeling approaches on both the systems level and in structural biology have already become a necessary part of normal research practice. While the combination of machine learning and biological research is a particularly good match with lots of opportunities, the work requires expertise in several areas and hence is very challenging and needs frequent interaction between researchers.

The group of researchers working in this field is normally spread thinly in the currently abundant, partly but not fully relevant conferences: both biological and bioinformatics conferences on the one hand, and pure machine learning conferences on the other. The aim of this workshop was to function as a specifically targeted forum.

The workshop started a series which will be continued as MLSB'07, Machine Learning for Structural and Systems Biology, June 28–29, in Evry, France.

Summary of the supplement

Selected submissions were invited based on the papers presented in the workshop. We targeted a subset of around ten best papers, and almost succeeded. This supplement contains a reviewed selection of eleven full papers.

Two of the papers are about modeling of the causal or physical behavior of cellular systems. Rogers et al. [1] introduce a full-Bayesian model of kinetics of the activity of transcription factors in gene regulation, and Opgen-

Rhein and Strimmer [2] use shrinkage methods to estimate autoregressive processes from small samples, to infer causal gene regulatory networks.

Biological networks are modeled in three further papers as well. Geurts et al. [3] predict links in protein-protein interaction networks and enzyme networks with a new kind of kernel-based method. Michoel et al. [4] use a synthetic data generator to evaluate the performance of methods for learning module networks, including a new one they introduce.

Analysis of high-throughput data is a common subtheme in most works. Three of the papers are particularly focused in this task. Yoon et al. [5] introduce a robust preprocessing method for treating missing values in gene expression data, and Bertoni and Valentini [6] decide the number of clusters based on stability against fluctuations caused by random projections. In the only paper on metabonomics, Vehtari et al. [7] introduce a full-Bayesian way of modeling the mapping between NMR spectra and clinical variables. Three of the papers are related to genomics. Landwehr et al. [8] introduce a hidden Markov model-based method for haplotype reconstruction which is a subproblem of gene association studies for uncovering genetic bases of diseases. Dix et al. [9] use compression methods to analyze information content of DNA in a genome-wide scale. Oja et al. [10] use hidden Markov model-based methods to estimate activities of retroviruses residing in human genome, using EST databases.

Finally, Roth and Fischer [11] introduce a kernel-based fusion from multiple data sources for predicting multi-label protein function.

Acknowledgements

We wish to thank particularly warmly the programme committee of the PMSB 2007 conference; many of them participated in refereeing the papers of this supplement as well as the original workshop papers. The programme committee consisted of Florence d'Alché-Buc, Université d'Evry-Val d'Essonne; Jaakko Astola, Tampere University of Technology; Nello Cristianini, UC Davis/University of Bristol; Liisa Holm, University of Helsinki; Mark Girolami, University of Glasgow; Samuel Kaski, Helsinki University of Technology; Matej Orešič, Technical Research Centre of Finland; Juho Rousu, University of Helsinki; Esko Ukkonen, Helsinki Institute for Information Technology; and Jean-Philippe Vert, Ecole des Mines de Paris. We had additionally several other referees; they will be acknowledged according to the normal practices of BMC Bioinformatics to ensure their anonymity.

We thank for financial support the EU FP6 Network of Excellence PASCAL (IST-2002-506778), and Helsinki Institute for Information Technology, and both University of Helsinki and Helsinki University of Technology for providing the infrastructure for the workshop, and finally, all contributors and participants for the successful workshop.

This article has been published as part of *BMC Bioinformatics* Volume 8, Supplement 2, 2007: Probabilistic Modeling and Machine Learning in Structural and Systems Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/8?issue=S2.

References

1. Rogers S, Khanin R, Girolami M: **Bayesian model-based inference of transcription factor activity.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S2.
2. Opgen-Rhein R, Strimmer K: **Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S3.
3. Geurts P, Touleimat N, Dutreix M, d'Alché-Buc F: **Inferring biological networks with output kernel trees.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S4.
4. Michoel T, Maere S, Bonnet E, Joshi A, Saeys Y, Van den Bulcke T, Van Leemput K, van Remortel P, Kuiper M, Marchal K, Van de Peer Y: **Validating module network learning algorithms using simulated data.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S5.
5. Yoon D, Lee EK, Park T: **Robust imputation method for missing values in microarray data.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S6.
6. Bertoni A, Valentini G: **Model order selection for bio-molecular data clustering.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S7.
7. Vehtari A, Mäkinen VP, Soininen P, Ingman P, Mäkelä SM, Savolainen MJ, Hannuksela ML, Kaski K, Ala-Korpela M: **A novel Bayesian approach to quantify clinical variables and to determine their spectroscopic counterparts in 1H NMR metabonomic data.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S8.
8. Landwehr N, Mielikäinen T, Eronen L, Toivonen H, Mannila H: **Constrained Hidden Markov Models for population-based haplotyping.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S9.
9. Dix TI, Powell DR, Allison L, Bernal J, Jaeger S, Stern L: **Comparative analysis of long DNA sequences by per element information content using different contexts.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S10.
10. Oja M, Peltonen J, Blomberg J, Kaski S: **Methods for estimating human endogenous retrovirus activities from EST databases.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S11.
11. Roth V, Fisher B: **Improved functional prediction of proteins by learning kernel combinations in multilabel settings.** *BMC Bioinformatics* 2007, **8(Suppl 3):**S12.