Tampereen teknillinen yliopisto. Julkaisu 548
Tampere University of Technology. Publication 548

Harri Lähdesmäki

Computational Methods for Systems Biology:

Analysis of High-Throughput Measurements and Modeling of Genetic Regulatory Networks

Thesis for the degree of Doctor of Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 27th of October 2005, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2005

ISBN 952-15-1454-X
ISSN 1459-2045

Abstract

High-throughput measurement techniques have revolutionized the field of molecular biology by gearing biological research towards approaches that involve extensive collection of experimental data and integrated analysis of biological systems on a genome-wide scale. Integration of experimental and computational approaches to understand complex biological systems—computational systems biology—has the potential to play a profound role in making life science discoveries in the future. Analysis of massive amounts of measurement data and modeling of high-dimensional biological systems inevitably require advanced computational methods in order to draw valid biological conclusions.

This thesis introduces novel computational methods for the problems encountered in the field of systems biology. The content of the thesis is three-fold.

The first part introduces methods for high-throughput measurement preprocessing. Two general methods for correcting systematic distortions originating from sample heterogeneity and sample asynchrony are developed. The former distortion is typically present in experiments conducted on non-homogeneous cell populations and the latter is encountered in practically all biological time series experiments.

The second topic focuses on robust time series analysis. General methods for both robust spectrum estimation and robust periodicity detection are introduced. Robust computational methods are preferred because the exact statistical characteristics of high-throughput data are generally unknown and the measurements are also prone to contain other non-idealities, such as outliers and distortion from the original wave form.

The third part is devoted to integrated analysis of genetic regulatory networks, or biological networks as they are also called, on a global scale. The effect of certain Post function classes on general properties of genetic regulatory networks, such as robustness and ordered and chaotic behavior, is studied in the Boolean network framework. In order to facilitate the analysis of generic properties of biological networks, efficient spectral methods for testing membership in the studied Post function classes and the class of forcing functions (as well as its variants) are introduced. Fast optimized search algorithms are developed for the inference of regulatory functions from experimental data. Relationships between two commonly used stochastic networks models, probabilistic Boolean networks (PBN) and dynamic Bayesian networks (DBN), are also established. This connection provides a way of applying the standard tools of DBNs to PBNs and the other way around.

Acknowledgements

I am grateful to my advisor Prof. Olli Yli-Harja for guiding me through the doctoral studies. His sense of humor and continuous encouragement and support have inspired me along the way. I would also like to thank my friends and colleagues in our Computational Systems Biology group.

I am highly grateful to Prof. Ilya Shmulevich for his invaluable guidance and deep involvement during this process. Ilya's influence and contribution to this dissertation cannot be overemphasized. I am also indebted to Prof. Wei Zhang for the excited and insightful guidance he gave me on systems biology.

This work has been carried out at the Institute of Signal Processing in Tampere University of Technology and partly at the Cancer Genomics Laboratory in The University of Texas M. D. Anderson Cancer Center during my research visit there. Special thanks go to all the personnel in both institutes. The financial support of the Tampere Graduate School in Information Science and Engineering (TISE), Academy of Finland, Emil Aaltonen Foundation, Kauhajoki Cultural Foundation, Jenny and Antti Wihuri Foundation and Instrumentarium Foundation are gratefully acknowledged.

I would like express my gratitude to my parents, mother Liisa and father Kari, and to my sister Riitta for their constant support. And finally, my warm thanks goes to my wife Marianna.

Contents

List of Publications

This thesis is based on the following publications. In the text, these publications are referred to as Publication-I, Publication-II, etc.

I Lähdesmäki, H., Huttunen, H., Aho, T., Linne, M.-L., Niemi, J., Kesseli, J., Pearson, R. and Yli-Harja, O. (2003) Estimation and inversion of the effects of cell population asynchrony in gene expression time-series. *Signal Processing*, Vol. 83, No. 4, pp. 835–858.

II Lähdesmäki, H., Shmulevich, I. and Yli-Harja, O. (2003) On learning gene regulatory networks under the Boolean network model. *Machine Learning*, Vol. 52, No. 1–2, pp. 147–167.

III Shmulevich, I., Lähdesmäki, H., Dougherty, E.R., Astola, J. and Zhang, W. (2003) The role of certain Post classes in Boolean network models of genetic networks. *Proceedings of the National Academy of Sciences of the USA*, Vol. 100, No. 19, pp. 10734–10739.

IV Pearson, R.K., Lähdesmäki, H., Huttunen, H. and Yli-Harja, O. (2003) Detecting periodicity in nonideal datasets. In *SIAM International Conference on Data Mining 2003*, Cathedral Hill Hotel, San Francisco, CA, May 1-3.

V Shmulevich, I. Lähdesmäki, H. and Egiazarian, K. (2004) Spectral methods for testing membership in certain Post classes and the class of forcing functions. *IEEE Signal Processing Letters*, Vol. 11, No. 2, pp. 289–292.

VI Lähdesmäki, H., Shmulevich, I., Yli-Harja, O. and Astola, J. (to appear) Inference of genetic regulatory networks via Best-Fit extensions. To appear in W. Zhang and I. Shmulevich (Eds.) *Computational And*

Statistical Approaches To Genomics (2nd ed.), Boston: Kluwer Academic Publishers.

VII Lähdesmäki, H., Shmulevich, I., Dunmire, V., Yli-Harja, O. and Zhang, W. (2005) *In silico* microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics*, 6:54.

VIII Lähdesmäki, H., Hautaniemi, S., Shmulevich, I. and Yli-Harja, O. (to appear) Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks. To appear in *Signal Processing*.

IX Ahdesmäki, M.,† Lähdesmäki, H.,† Pearson, R., Huttunen, H. and Yli-Harja, O. (2005) Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, 6:117.

The author's contribution to Publications II, VI, VII and VIII is as follows. As the first author of these publications, H. Lähdesmäki designed and implemented the computational methods, derived the mathematical proofs, and wrote the manuscript for most part, with the exception that Publication VI was co-written with I. Shmulevich. W. Zhang and I. Shmulevich also contributed to Publication VII by providing essential ideas and assisting in drafting the manuscript.

Publication I was a result of collective efforts. As the first author, H. Lähdesmäki had a major role in writing the manuscript. The author was also mainly responsible for the development of those computational methods that are covered in this thesis. Other subtopics to which the author did not make the main contribution, such as the proposed blind deconvolution method developed by Dr. H. Huttunen, are not discussed in this thesis in detail.

In Publications III and V, the author assisted in developing the computational methods and co-performed the simulations. In Publication IV, the author performed the simulations and helped in refining the computational methods.

M. Ahdesmäki and H. Lähdesmäki were equal contributors to Publication IX. H. Lähdesmäki developed the statistical methods, assisted in performing the simulations and mainly drafted the manuscript. M. Ahdesmäki carried out an implementation of the methods, performed the most of the extensive simulations and co-drafted the manuscript.

The author has also published the following related publications. In the text, these publications are referred to as Publication-A, Publication-B and Publication-C.

A Lähdesmäki, H., Hao, X., Sun, B., Hu, L., Yli-Harja, O., Shmulevich, I. and Zhang, W. (2004) Distinguishing key biological pathways between primary breast cancers and their lymph node metastases by gene function-based clustering analysis. *International Journal of Oncology*, Vol. 24, No. 6, pp. 1589–1596.

B Hao, X., Sun, B., Hu, L., Lähdesmäki, H., Dunmire, V., Feng, Y., Zhang, S.-W., Wang, H., Wu, C., Wang, H., Fuller, G.N., Symmans, W.F., Shmulevich, I. and Zhang, W. (2004) Differential gene and protein expression in primary breast malignancies and their lymph node metastases as revealed by combined cDNA microarray and tissue microarray analysis. *Cancer*, Vol. 100, No. 6, pp. 1110–1122.

C Lähdesmäki, H., Yli-Harja, O., Zhang, W. and Shmulevich, I. (2005) Intrinsic dimensionality in gene expression analysis. In *IEEE International Workshop on Genomic Signal Processing and Statistics 2005*, Hyatt Regent Hotel, New Port, Rhode Island, May 22-24.