Mixture Models for the Analysis of Gene Expression: Integration of Multiple Experiments and Cluster Validation

Ivan Gesteira Costa Filho

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Joachim Selbig

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Prof. Dr. Joachim Selbig
Tag der Promotion: 29 Mai 2008

Contents

List of Figures

List of Tables

Preface

Acknowledgments

First of all, I would like to thank Dr. Alexander Schliep for his supervision and support during my Ph.D. studies. He introduced me to the field of mixture models, motivated me to do state-of-the-art bioinformatics research, and was a valuable source of research ideas. Some of such ideas that I could turn into reality are the main parts of this thesis.

I am grateful to Dr. Stefan Roepcke for introducing me to the Lymphoid cells research area and for great collaborative work. And more importantly, for keeping me aware of the other side (biology!). I would also acknowledge Prof. Fritz Melchers for his inspiring “private lectures” on B (and T) cells development, and encouragements on my work on Lymphoid development. I thank Dr. Alexander Schoenhut for his work on topology learning and Viterbi decomposition for HMMs and several discussions on the analysis of gene expression time courses. I am also in debt to Dr. Roland Krause for his valuable opinions, as well as for his work on the analysis of Drosophila development data. I acknowledge the aid of Lennart Optiz in the in-situ image pre-processing. I thank Dr. Marcilio de Souto for his advices and great help with the thesis corrections. Furthermore, we carried out a collaboration work on clustering of cancer gene expression data, together with several Brazilian colleagues: Prof. Teresa Ludemir, Dr. Francisco de Carvalho, Dr. Ricardo Prudncio, Daniel Arajo e Rodrigo Soares. I am also grateful to Prof. Joachim Selbig for participating in my Ph.D. thesis committee and for his valuable remarks.

It was a great experience to be a member of the Algorithms group. The group was a perfect environment for scientific discussions and for the collaborative development of “open source” software libraries (e.g., ghmm, GQL, GATO and pymix). The latter was particularly important for development of this thesis. In special, I would like to thank Christopher Hafemeister for his great work in the software of mixture of dependence trees; Benjamin Georgi for many interesting discussions; Jane Grunau for his prompt support on the GHMM library; and Rubens Schilling for both his work on GQL and his help with image processing analysis.

I owe most of my knowledge on computational biology to the Department of Computational Molecular Biology (CMB) and its members. The interaction with the department members through seminars, meetings and retreats was of great importance to my development as a computational biologist. In particular, I would like to thank Dr. Steffen Grossman for discussions on Gene Ontology, cluster validation and for sharing the office; Helge Roider for discussion on gene regulation and his help with the promoter analysis of Lymphoid development; Hughes Richards for several discussions; Willy for prompt answers on technical problems; and Utz Pape for his aid with this thesis corrections. Finally, I am very thankful to Prof. Martin Vingron for being my supervisor and giving me the opportunity to be a member of the department.

This work would not have been possible without the support, friendship and love from special friends and family members: Carlos Venezuela, Fernandão, Gui, Hugo, Jorgito, Julia, Lauro Mostrinho, Lili, Lucciano, Marcilio, Paulete, Philipão, Reginho, Sandra, Scylla and Stefan. Thanks! I would also like to thank Gabi for all love and companionship on earlier years of this work. Last but not least, I am very grateful to my parents: my father, for being my role model of wisdom, which pursue lead to my curiosity for knowledge, which consequently brought me to science; my mother for inciting and fostering my need for knowledge and for their unconditional love. I dedicate this thesis for them.

Publications

Parts of this thesis have been previously published. Chapter 3 includes results of a paper in the Annual Conference of the German Classification Society 2005 [51]. Also, parts of the results in Chapter 4 were published in the journals IEEE Transactions of Bioinformatics and Computational Biology [185] and Bioinformatics [53]. Chapter 5 includes results presented at the PLoS Track of the International Conference on Intelligent Systems for Molecular Biology 2006 and published in the journal BMC Immunology [50] and results accepted for publication in the International Conference on Intelligent Systems for Molecular Biology 2008 [49]. Chapter 6 contains results presented at the NIPS Workshop on New Problems and Methods in Computational Biology 2005, published in the ECML Workshop of Data and Text Mining for Integrative Biology 2006 [52], and in the journal BMC Bioinformatics [48]. Some collaborative work during my Ph.D studies, which were not described in this thesis, also lead to publications on the topics of cluster validation [47, 59, 60] and semi-supervised learning [189].