

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

Preface	ix
1 Introduction	1
1.1 Gene Expression: Transcription, Translation and Control	1
1.2 Measuring Gene Expression with Microarrays	3
1.3 Thesis Overview	6
2 Finite Mixture Models	9
2.1 Basics	9
2.2 Mixture Models	10
2.3 Mixture Model Estimation	11
2.3.1 Expectation-maximization Algorithm	11
2.3.2 Method Initialization	15
2.3.3 Mixture of Multivariate Gaussians	15
2.3.4 EM and Local Maxima	16
2.3.5 Determining the Number of Components	17
3 Mixture Models and Clustering	21
3.1 Clustering with Mixture Models	22
3.2 Validation of Mixture Models	23
3.2.1 External Indices	24
3.3 Experiments	26
3.3.1 Simulated Data 1	27
3.3.2 Simulated Data 2	28
4 Analysis of Gene Expression Time Courses	33
4.1 Related Work	34
4.2 Hidden Markov Models	38
4.2.1 Linear HMM and Time Courses	40
4.3 Querying of Time Courses	47
4.4 Mixture of linear HMMs	48
4.4.1 Model Initialization	49
4.4.2 Viterbi Decomposition	50
4.4.3 GQLCluster	50
4.5 Experiments	51

4.5.1	Data	52
4.5.2	Results	53
5	Analysis of Gene Expression in Lymphoid Development	57
5.1	Related Work	58
5.2	Dependence Trees	61
5.2.1	Equivalence of Dependence Trees	62
5.2.2	Parameterization of Dependence Trees	64
5.2.3	Estimation of the Structure of Dependence Trees	65
5.2.4	Dependence Trees and Multivariate Gaussians	67
5.3	Mixture of Dependence Trees	67
5.3.1	MixDTrees with Developmental Tree as Structure	68
5.3.2	Maximum-a-posteriori Estimates	68
5.3.3	MixDTrees with Estimated Structure	72
5.4	Experiments	74
5.4.1	MixDTrees with Developmental Tree Structure	74
5.4.2	MixDTrees with Estimated Structure	82
6	Clustering with Constraints for Integration of Heterogeneous Biological Data	89
6.1	Related Work	90
6.2	Mixture Model Estimation with Constraints	93
6.2.1	Mean Field Approximation	94
6.2.2	Deriving Constraints	96
6.3	Experiments	98
6.3.1	Yeast Cell Cycle with Gene Ontology and Location Analysis	100
6.3.2	Drosophila Syn-Expression	102
7	Discussion	115
	Bibliography	119
A	Gene Ontology enrichment	137
B	Analysis of Gene Expression of Lymphoid Development	139
C	Notation	141
D	Abbreviations	143
E	Zusammenfassung	145
F	Curriculum Vitae	147

List of Figures

1.1	Example of a double stranded DNA molecule	2
1.2	Biology of gene expression	3
1.3	Schema of cDNA (a) and Oligonucleotide (b) microarrays	5
2.1	Examples of densities modeled by mixtures of two Gaussians pdfs	11
2.2	Examples of solutions found by a mixture of Gaussians with distinct co- variance matrices	16
2.3	Examples of mixture models with 1, 6 and 82 components fitting the Galaxy Velocity data	18
3.1	Example of entropy of posterior assignments for a bimodal density	23
3.2	Mean CR and ECR for the the mixture estimation from a bimodal normal density	27
3.3	The data points of data sampled from the base mixture	29
4.1	Example of a linear HMM with two emitting states	40
4.2	Example of linear HMMs modeling up-regulation and up- and down-regulation patterns	42
4.3	Example of up-regulated time courses with 19 time points sampled from a linear HMM	43
4.4	Example of the Viterbi paths of the time courses for a given IHMM	44
4.5	Example of the extension of the linear HMM for modeling periodicity	45
4.6	Heat map plots of the absolute value of the empirical covariance of data sampled from several IHMMs	46
4.7	Heat map plots of the absolute value of the empirical covariance of data sampled from a 19 state IHMM at distinct up-regulation time points	47
4.8	GQLQuery interface	48
4.9	GQLCluster interface	51
4.10	BIC versus number of components for the data set YCC	54
4.11	Results for the HeLa data set	55
4.12	The mean GO specificity level of clusters from HeLa versus the entropy threshold	55
5.1	Schematic view of lymphocyte cell development	58
5.2	Example of a simple developmental tree and a group of developmental profiles	61

5.3	The undirected tree structure of the graph from Figure 5.2 and four possible directed versions	63
5.4	Example of a mixture of four DTrees with the structure defined in Figure 5.2	68
5.5	Example of a developmental tree and its gene expression data	73
5.6	Selected clusters from MixDTrees-Dev for TCell	76
5.7	Selected clusters from MixDTrees-Dev for BCell	77
5.8	Strategy to identify enriched microRNAs	78
5.9	We display the mean sensitivity and mean specificity against five experimental settings	81
5.10	Developmental tree with the stages contained in the Lymphoid data set . .	83
5.11	The DTree and expression profiles of selected groups from the Lymphoid data	85
5.12	Scatter plot comparing the KEGG pathway enrichment of MoG diag and MixDTrees-Str-MAP	87
5.13	Heat-map plot displaying the comparison of KEGG and GO enrichment for 10 distinct clustering methods.	87
5.14	Mean corrected Rand, sensitivity and specificity of true label recovery for distinct clustering methods	88
6.1	Example of semi-supervised clustering.	95
6.2	Examples of positively constrained genes and their corresponding in-situ images	99
6.3	CR, Sens and Spec after clustering YCC with positive, negative and both constraints.	101
6.4	CR obtained by clustering YCC with positive constraints after the inclusion of random labels	102
6.5	CR, Sens and Spec after clustering YCC after filtering of relevant TR . .	102
6.6	Image processing pipeline.	104
6.7	Clustering result of mixture of Gaussians.	106
6.8	Clustering result of mixture of Gaussians with constrains.	107
6.9	Scatter plot comparing the ImaGO term enrichment of MoG and cMoG. . .	109
6.10	Ratio of ImaGO terms with p -value $cMoG < p$ -value MoG against the threshold τ	109
6.11	Scatter plot comparing the ImaGO term enrichment of MoG and cMoG for a $\tau = 0.3$	110
6.12	Averaged in-situ images of clusters C2, C3 and C10 from lateral and dorsal views.	111

List of Tables

3.1	The mean ECR, CR and BIC values for data with distinct distributions against the number of components of the estimated mixtures	30
3.2	The mean ECR, CR and BIC values for data with distinct number of clusters against the number of components of the estimated mixtures	31
4.1	The results of the different methods for YCC	53
5.1	List of <i>LymphMIR</i> enriched in the clusters from <i>MixDTrees-Dev</i> for data sets <i>TCell</i> and <i>BCell</i>	79
6.1	Proportion of <i>ImaGO</i> terms with lower <i>p</i> -values in <i>cMoG</i> compared to <i>MoG</i> for distinct constraints weights	108
A.1	2x2 Contingency Table for genes annotated or not annotated by a given GO term	137
B.1	Contingency Table comparing results from <i>MixDTrees-Dev</i> (columns) versus <i>SOM</i> (lines) for <i>TCell</i>	139
B.2	Contingency Table comparing results from <i>MixDTrees-Dev</i> (columns) versus <i>SOM</i> (lines) for <i>BCell</i>	139
B.3	MicroRNA enrichment per cluster for <i>TCell</i> for <i>MixDTrees-Dev</i> . .	140
B.4	MicroRNA enrichment per cluster for <i>BCell</i> for <i>MixDTrees-Dev</i> . .	140

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 Roi-

der 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].

