Bibliography


Bibliography


Bibliography


Appendix A

Gene Ontology enrichment

In order to find GO terms with annotations related to a given group (or cluster) of genes, one should look for annotation terms that are over-represented in this group. The probability that this over-representation is not found by chance can be measured with the use of a hyper-geometric Fisher exact test [199]. This test returns for each cluster and gene ontology term a \( p \)-value describing how statistically significant a GO term is for describing genes in a particular cluster.

Let \( n \) be the total number of annotated genes in GO (reference group), and \( m \) be the number of genes annotated with a specific GO term. This will give us \( m \) positive genes and \( n - m \) negative genes. If we draw \( k \) genes from the reference group (or analogously obtain a cluster with \( k \) genes), we obtain \( q \) positive genes and \( k - q \) negative genes, see Table A.1 for a 2X2 contingency table representation of these terms. We are interested in observing how unusually large this value \( q \) is, given \( n \), \( m \) and \( k \). This can be achieved by calculating a \( p \)-value defined by \( p(X \geq q) \), where \( X \) is defined by \( \{P(x = i)\}_{1 \leq i \leq k} \), and \( P(x = i) \) is defined as below:

\[
P(x = i) = \frac{\binom{m}{i} \binom{n-m}{k-i}}{\binom{n}{k}}
\]

In the thesis, when a particular GO term is over-represented for a given cluster, we state GO Term X is enriched in cluster Y, or we found enrichment for GO Term X in cluster Y.

A later correction of the \( p \)-values is necessary, because of the effects of multiple testing. For example, if we have 1000 GO terms, and a \( p \)-value of 0.1 is used, at least 100 false

<table>
<thead>
<tr>
<th>Table A.1: 2x2 Contingency Table for genes annotated or not annotated by a given GO term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotated Genes</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>in cluster</td>
</tr>
<tr>
<td>not in cluster</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>
positives are expected. To correct this, we apply a false positive discovery ratio proposed in [175].
Appendix B

Analysis of Gene Expression of Lymphoid Development

Table B.1: Contingency Table comparing results from MixDTrees-Dev (columns) versus SOM (lines) for TCell

|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 1 | 6 | 38 | 14 | 1 | 34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 6 | 38 | 14 | 1 | 34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 4 | 12 | 31 | 32 | 25 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 10 | 0 | 13 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 2 | 0 | 0 |
| 5 | 0 | 0 | 0 | 35 | 8 | 88 | 3 | 34 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 1 | 0 | 0 | 1 | 15 | 6 | 9 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 3 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 10 | 7 | 2 | 23 | 9 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 35 | 0 | 49 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 12 | 21 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 47 | 18 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 12 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 4 | 11 | 5 | 7 | 4 | 8 | 5 | 2 | 1 | 0 |
| 13 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 7 | 4 | 7 | 1 | 8 | 0 | 2 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 24 | 23 | 0 | 0 | 0 |
| 15 | 4 | 0 | 5 | 2 | 0 | 5 | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table B.2: Contingency Table comparing results from MixDTrees-Dev (columns) versus SOM (lines) for BCell

|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| 1 | 52 | 0 | 0 | 0 | 0 | 0 | 4 | 5 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 20 | 12 | 10 | 0 | 4 | 0 | 0 | 1 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 64 | 5 | 25 | 2 | 0 | 0 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 14 | 3 | 8 | 4 | 2 | 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 7 | 43 | 5 | 10 | 42 | 2 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 4 | 1 | 0 | 0 | 4 | 2 | 0 | 2 | 1 | 1 | 18 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 1 | 10 | 0 | 0 | 0 | 7 | 1 | 17 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 6 | 0 | 0 | 0 | 18 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 2 | 6 | 18 | 2 | 6 | 1 | 0 | 1 | 0 | 1 |
| 12 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 2 | 6 | 18 | 2 | 6 | 1 | 0 | 1 | 0 | 1 | 0 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 3 | 28 | 3 | 8 | 4 |
| 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 3 | 25 | 24 | 0 | 0 | 3 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
### Table B.3: MicroRNA enrichment per cluster for TCell for MixDTrees-Dev

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>MicroRNA</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>miR-222</td>
<td>0.0006906</td>
</tr>
<tr>
<td>5</td>
<td>miR-15a</td>
<td>0.0019456</td>
</tr>
<tr>
<td></td>
<td>miR-26a</td>
<td>0.0369906</td>
</tr>
<tr>
<td></td>
<td>miR-24</td>
<td>0.0369906</td>
</tr>
<tr>
<td></td>
<td>miR-221</td>
<td>0.0051746</td>
</tr>
<tr>
<td></td>
<td>miR-181a</td>
<td>0.0244306</td>
</tr>
<tr>
<td>7</td>
<td>miR-342</td>
<td>0.0200686</td>
</tr>
<tr>
<td>8</td>
<td>miR-26a</td>
<td>0.0013526</td>
</tr>
<tr>
<td>10</td>
<td>miR-150</td>
<td>0.0012176</td>
</tr>
<tr>
<td></td>
<td>miR-142-3p</td>
<td>0.0000056</td>
</tr>
<tr>
<td>11</td>
<td>miR-16</td>
<td>0.0049776</td>
</tr>
<tr>
<td></td>
<td>miR-146</td>
<td>0.0011936</td>
</tr>
<tr>
<td></td>
<td>miR-181b</td>
<td>0.0049776</td>
</tr>
</tbody>
</table>

### Table B.4: MicroRNA enrichment per cluster for BCell for MixDTrees-Dev

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>MicroRNA</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>miR-26a</td>
<td>0.0358116</td>
</tr>
<tr>
<td></td>
<td>miR-181c</td>
<td>0.0025866</td>
</tr>
<tr>
<td></td>
<td>miR-181b</td>
<td>0.0358116</td>
</tr>
<tr>
<td>5</td>
<td>miR-15b</td>
<td>0.0029956</td>
</tr>
<tr>
<td></td>
<td>miR-15a</td>
<td>0.0029956</td>
</tr>
<tr>
<td></td>
<td>miR-223</td>
<td>0.0029956</td>
</tr>
<tr>
<td></td>
<td>miR-221</td>
<td>0.0323296</td>
</tr>
<tr>
<td>6</td>
<td>miR-191</td>
<td>0.0486736</td>
</tr>
<tr>
<td></td>
<td>miR-155</td>
<td>0.0271276</td>
</tr>
<tr>
<td>19</td>
<td>miR-342</td>
<td>0.0402686</td>
</tr>
<tr>
<td></td>
<td>miR-142-3p</td>
<td>0.0088346</td>
</tr>
</tbody>
</table>
Appendix C

Notation

All chapters

1(e) indicator function, which takes value 1 iff e is true
α_k mixture coefficient of the kth mixture component
E[X] expectation of a random variable X
L likelihood function
K number of clusters or components in a mixture model
μ_x mean value of random variable X
\[ p(x \mid \theta) \] a probability density function over variable X and parameterized by \( \theta \)
\( r_{ik} \) posterior probability that observation \( x_i \) is assigned to the kth mixture component, i.e., \( p(y_i = k \mid x_i, \Theta) \)
\( \Sigma_x \) covariance matrix of random variable X
\( \Theta \) set of parameters of a mixture model
\( \theta_k \) set of parameters of the kth mixture component
X an L dimensional continuous random variable
x an observation vector \((x_1, \ldots, x_L)\) from X
X a data set represented by a \( N \times L \) matrix, where entry \( x_{ij} \) denotes the values of the jth variable from the ith observation
Y an one dimensional discrete random variable
y an observation of Y, where \( y \in \{1, \ldots, K\} \) indicates the mixture component (or cluster) the observation belongs
Y a set of N observations from Y, where \( y_i = k \) denotes that the ith observation belongs to the kth mixture component (or mixture)
\( Y^c \) space of all possible values of Y

Chapter 4

A transition matrix of a HMM, where \( a_{uv} \) represents the probability of going from state \( u \) to state \( v \)
\( d_u \) duration parameter representing the expected number of visits to state \( u \)
Appendix C  Notation

\( M \)  number of states of the HMM
\( \mu_u \)  mean parameter of the emission function of the \( u \)th state
\( \pi_u \)  probability of visiting state \( u \) at time \( t = 1 \)
\( Q \)  an \( L \)-dimensional discrete variable representing the sequence of visited states
\( q \)  observation from \( Q \), where \( q = (q_1, ..., q_t, ..., q_L) \) and \( q_t \in \{1, ..., M\} \) represents the HMM state visited at time \( t \).
\( \sigma^2_u \)  standard error parameter of the emission function from the \( u \)th state
\( \theta_L \)  parameters of a linear HMM

Chapter 5

\( D(p||p^*) \)  relative entropy between the pdfs \( p \) and \( p^* \)
\( H(X) \)  entropy of variable \( X \)
\( I(X_u, X_v) \)  mutual information of variables \( X_u \) and \( X_v \)
\( p^T(x|\Theta) \)  dependence tree pdf
\( p(x_u|x_v, \tau_u) \)  conditional Gaussian pdf
\( pa \)  parent map defining the dependence tree structure
\( \sigma^2_{u|v} \)  standard error of the conditional Gaussian pdf
\( \tau_u \)  parameters of a conditional Gaussian pdf
\( w_{u|v} \)  regression parameter of the conditional Gaussian pdf

Chapter 6

\( \lambda^+ \)  parameter defining the penalty weights of positive constraint violations
\( \lambda^- \)  parameter defining the penalty weights of negative constraint violations
\( W \)  pair \((W^+, W^-)\) representing the positive and negative constraint matrices
\( W^+ \)  positive constraints matrix, where \( w^+_{ij} \) is the positive constraint value for observations \( i \) and \( j \)
\( W^- \)  negative constraints matrix, where \( w^-_{ij} \) is the negative constraint value for observations \( i \) and \( j \)
\( Z \)  an \( L \)-dimensional continuous random variable
\( z \)  an observation \((z_{i1}, ..., z_{il}, ..., z_{iL})\) of \( Z \) representing the pixel intensities of an image
Appendix D

Abbreviations

BCell B cell development data
Bimm immature B cells
BMC Bayesian model collection
Bpre pre B cells
Bpro pro B cells
BIC Bayesian information criteria
CL co-location index
CLP common lymphoid progenitor
CMP common myeloid progenitor
CR corrected Rand index
DAG directed acyclic graph
DN CD4-/CD8- double negative cells
DPL CD4+/CD8+ double positive large cells
DPS CD4+/CD8+ double positive small cells
DTree dependence tree
ECR extended corrected Rand index
ED equal density
EM expectation-maximization algorithm
E-Step expectation step
FACS fluorescence activated cell sorting
GQL Graphical Query Language
GO Gene Ontology
ImaGO Image Gene Ontology
HemoMIR hematopoiesis related microRNAs data
HMM hidden Markov model
HMRF hidden Markov random fields
KEGG Kyoto encyclopedia of genes and genomes
KMC $k$-means model collection
LHMM linear hidden Markov model
MAP maximum-a-posteriori
Appendix D  Abbreviations

MCMC  Monte Carlo Markov Chain
mir  microRNA
MixDTrees  mixture of dependence trees
MixDTrees-Dev  MixDTrees with the developmental tree as structure
MixDTrees-Str  MixDTrees with estimated structure
  MLE  maximum likelihood estimation
  MM  probe mismatch
  MoG  mixture of multivariate Gaussians
MoG Full  MoG with full covariance matrix
MoG Diag  MoG with diagonal covariance matrix
  M-Step  maximization step
  NK  natural killer cells
  NMF  non-negative matrix factorization
  PC  Pearson correlation
  pdf  probability density function
pHSC  pluri-potent, self-renewing hematopoietic stem cells
  PM  probe match
  PPP  pluripotent progenitor
RMC  random model collection
SCC  strongly connected components
Sens  sensitivity index
SIM  simulated data
SOM  self-organizing maps
SSL  semi-supervised learning
Spec  specificity index
SP4  single positive CD4
SP8  single positive CD8
TCell  T cell development data
TCD4  cd4 T cells
TCD8  cd8 T cells
TDN  double negative T cells
TF  transcription factor
TFBS  transcription factor binding site
TNK  natural killer T cells
  TR  transcription regulation data
YCC  yeast cell cycle
  VD  Viterbi decomposition