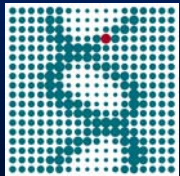


Analysis of Gene Expression Trees in Blood Cell Development

Ivan G. Costa Filho
Stefan Roepcke
Alexander Schliep



Computational Biology Department
Max Planck Institute for Molecular Genetics, Berlin

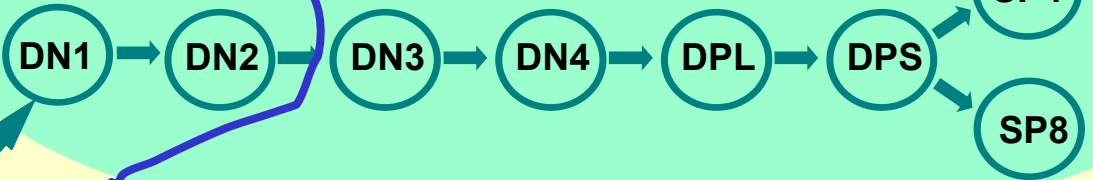
Blood Cell Development

- Motivation
 - understand development system
 - clinical interest: immune cells, leukemia
- Technical aspects
 - pure cell samples easily accessible
 - broadly studied developmental system
 - no computational framework available

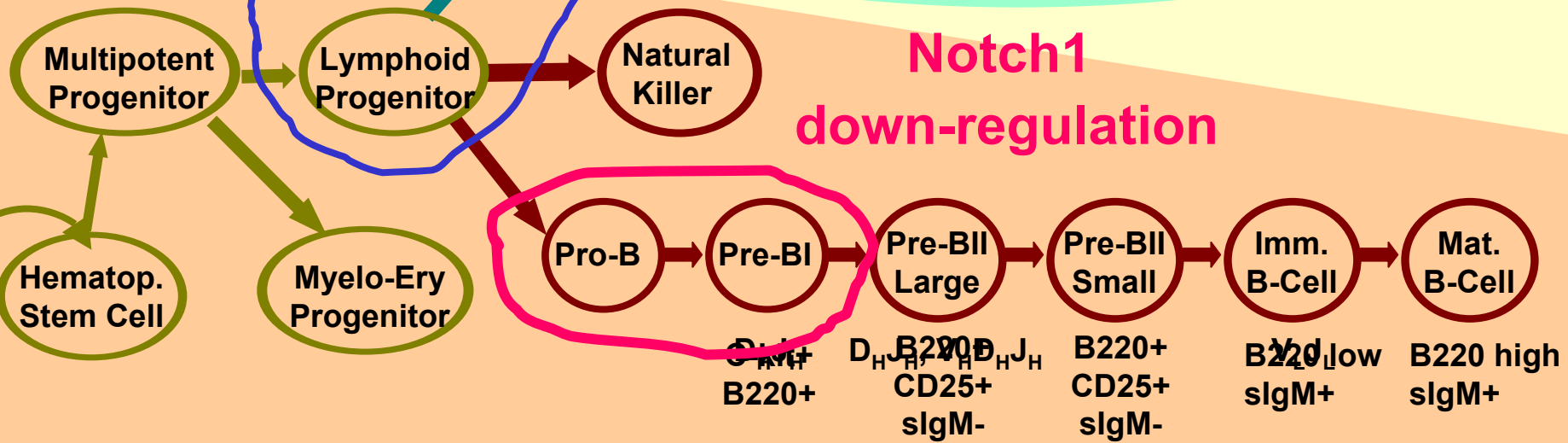
Blood Cell Development

Notch1
up-regulation

Thymus – T cell



Notch1
down-regulation



Bone Marrow

DN – double negatives, DP – double positives, SP – single positives

Goal

Understand **gene regulation**
during **development** of blood cells

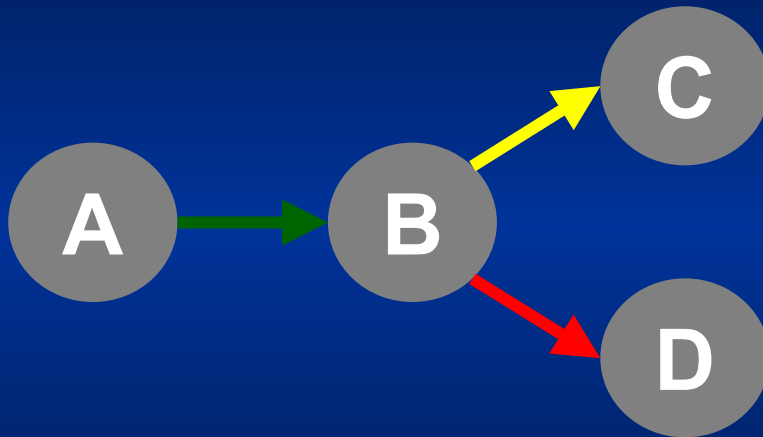
Method Outline

Use a statistical model for

1. finding similar 'development profiles'
 - tree model
2. clustering 'development profiles'
 - combine tree models in a mixture
3. interesting regulatory patterns
 - enrichment of microRNA targets

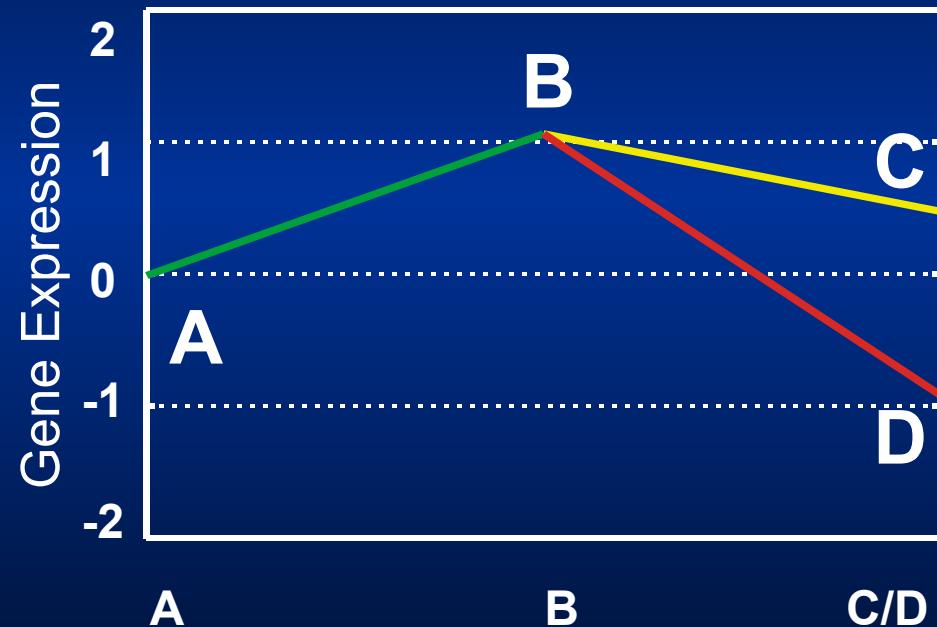
(1) Method Tree Model

Tree Model Development Profile

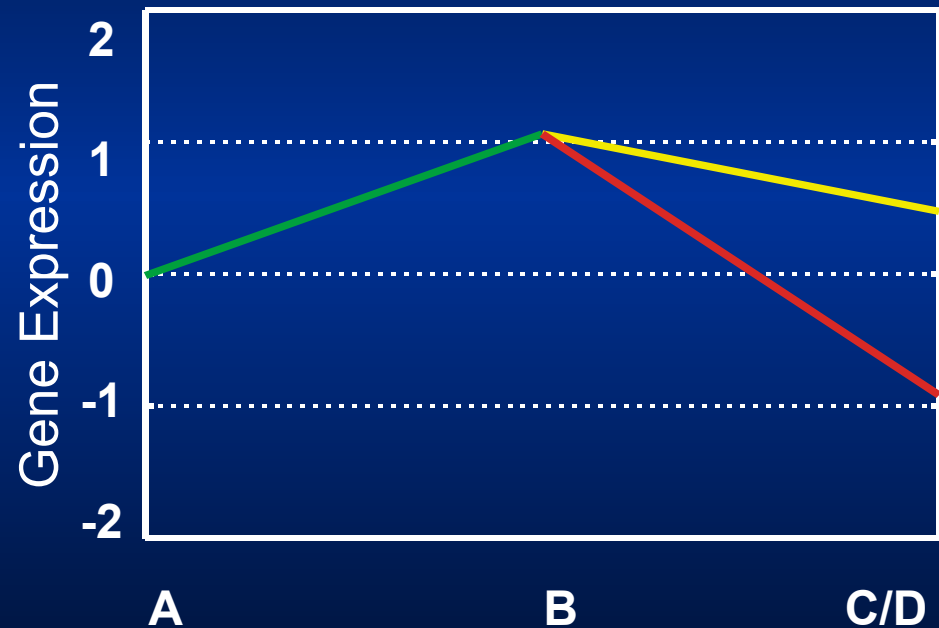
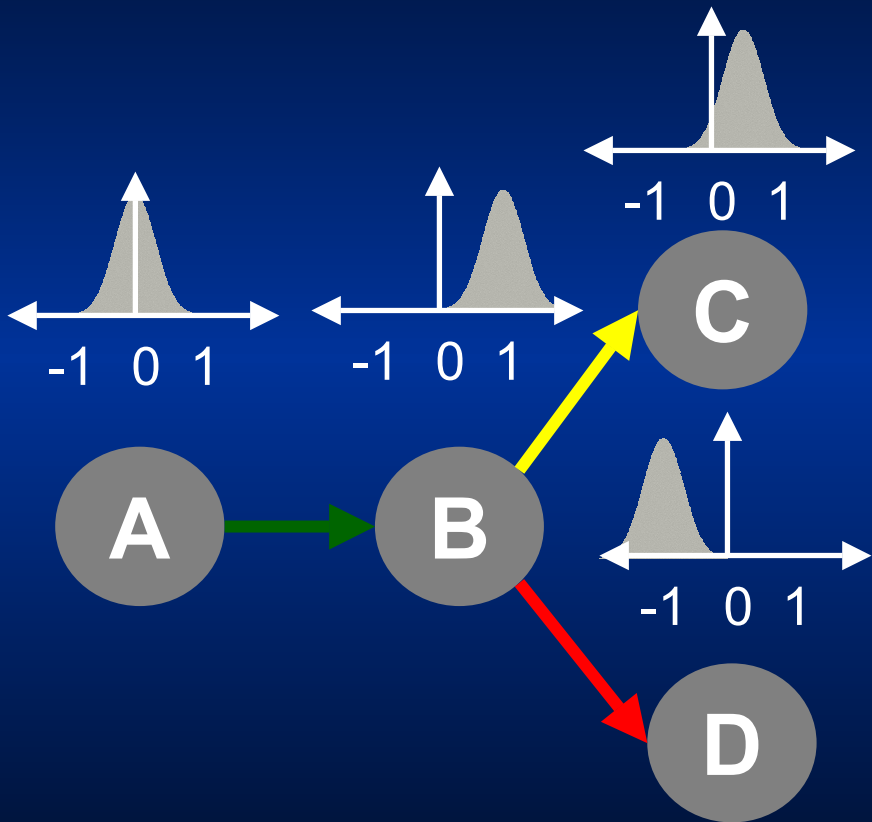


development profile:

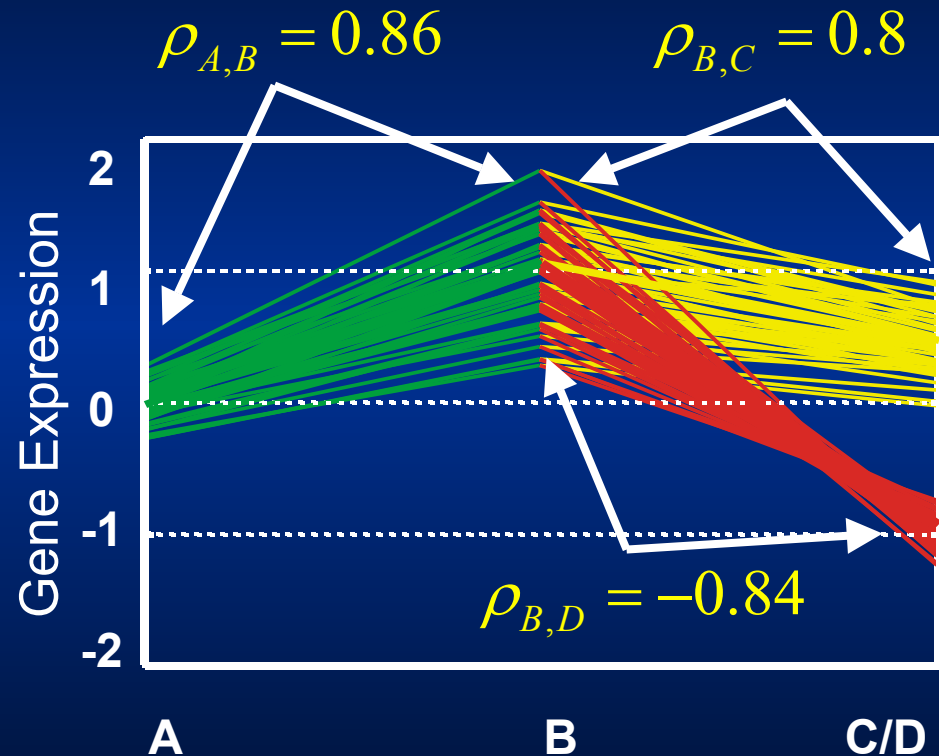
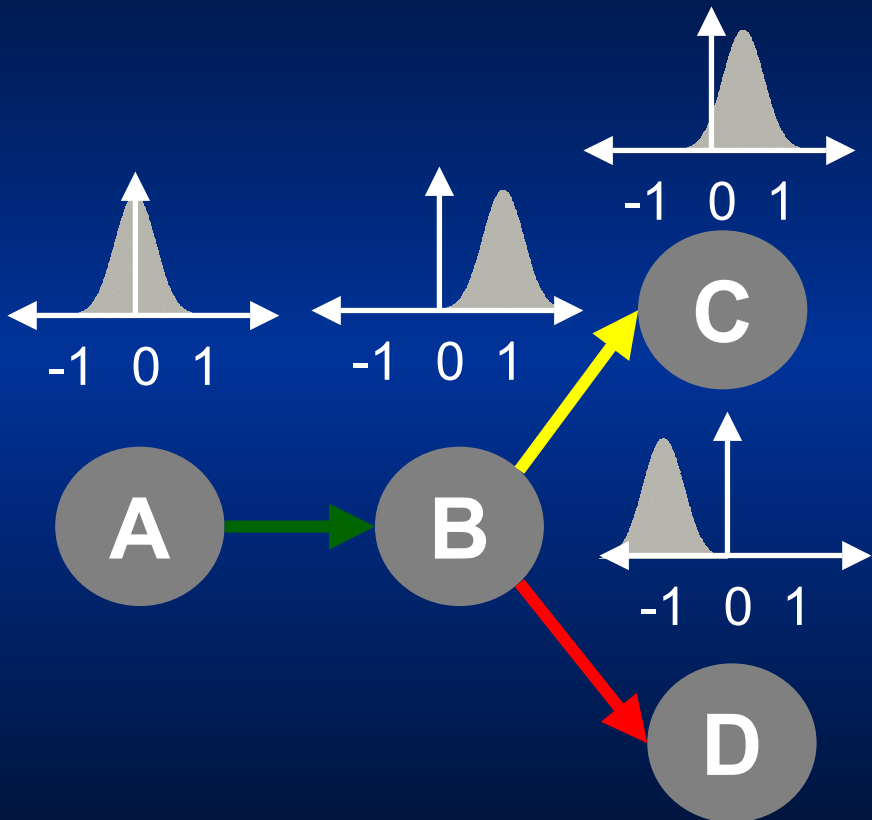
	A	B	C	D
Notch1	0	1	0.5	-1



Tree Model

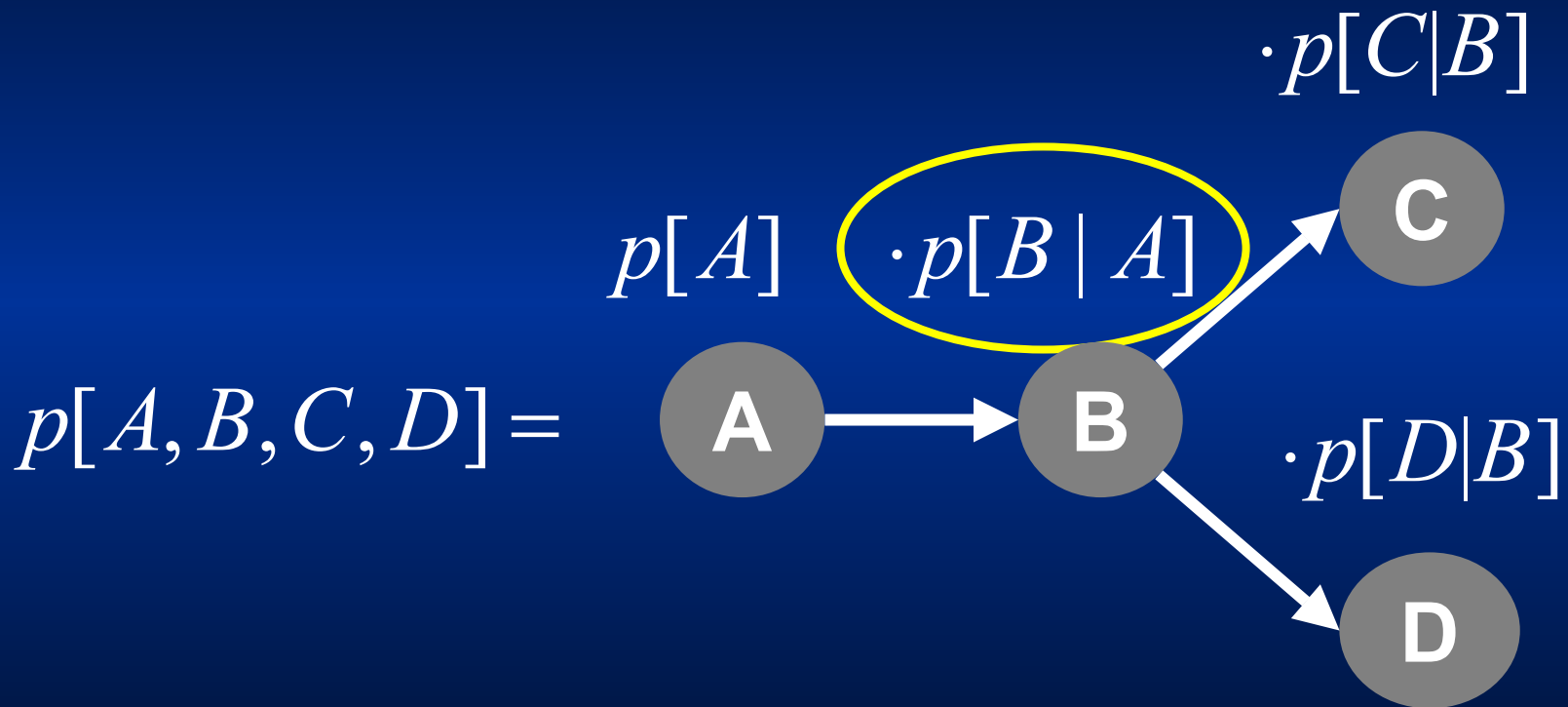


Tree Model



$\rho_{A,B}$ correlation of A and B

Tree Model Assumption



Tree Model

Conditional Gaussian

Conditional Gaussian pdf for $p[b|a]$,

$$p[b | a, \theta] = \frac{1}{\sqrt{2\pi\sigma_{B|A}^2}} \exp\left(\frac{-(b - \mu_B - w_{B|A}(a - \mu_A))^2}{2\sigma_{B|A}^2}\right),$$

Maximum likelihood estimates:

$$\mu_A$$

$$w_{B|A} = \sigma_{AB} / \sigma_A^2$$

$$\sigma_{B|A}^2 = \sigma_B^2 - w_{B|A}^2 \sigma_A^2$$

(2) Method

Mixtures of Trees

Perspective - Clustering

- ‘Classical’ methods assume independence between variables.
- ‘Complex’ models (ie. multivariate gaussians) over fit parameters.
- Methods for time-courses consider temporal dependencies, but not trees dependencies
- Mixtures of trees uses prior knowledge with the requirement of few additional parameters

Clustering Method

1. Combine K tree models in a mixture model

$$p[A, B, C, D | \alpha_1, \dots, \alpha_K, \text{tree}_1, \dots, \text{tree}_K]$$
$$= \sum_{j=1}^K \alpha_k \cdot p[A, B, C, D | \text{tree}_k]$$

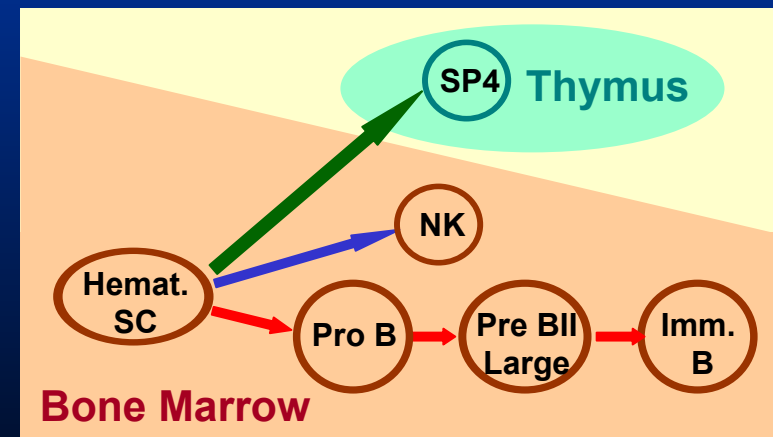
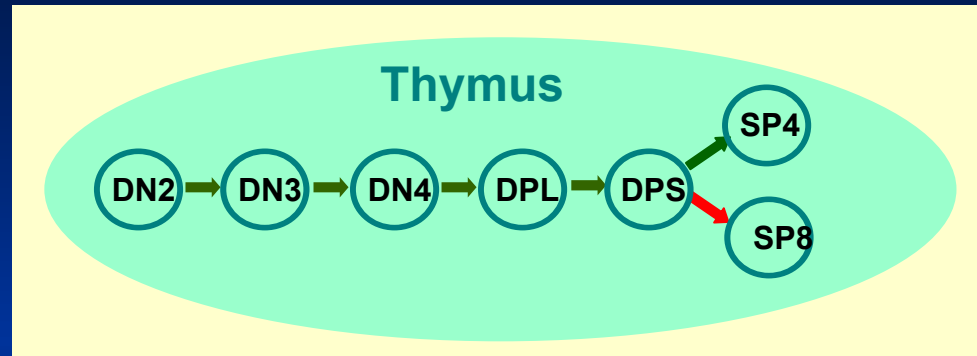
where **all trees share the same topology**

2. Estimate the mixture using EM algorithm

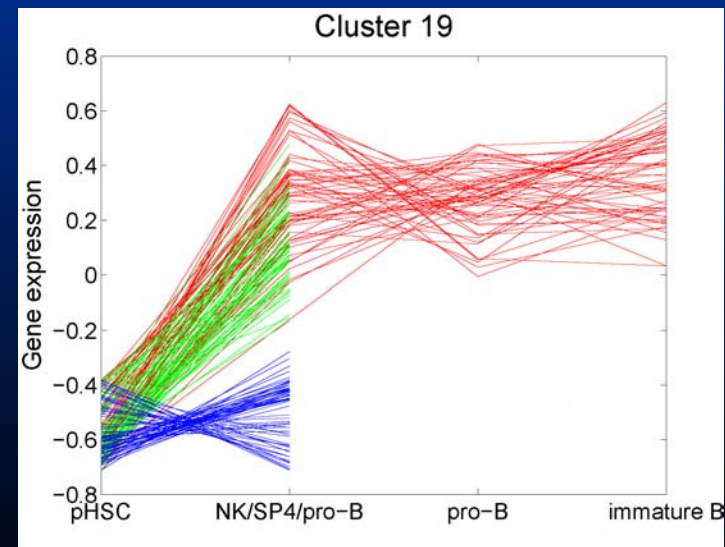
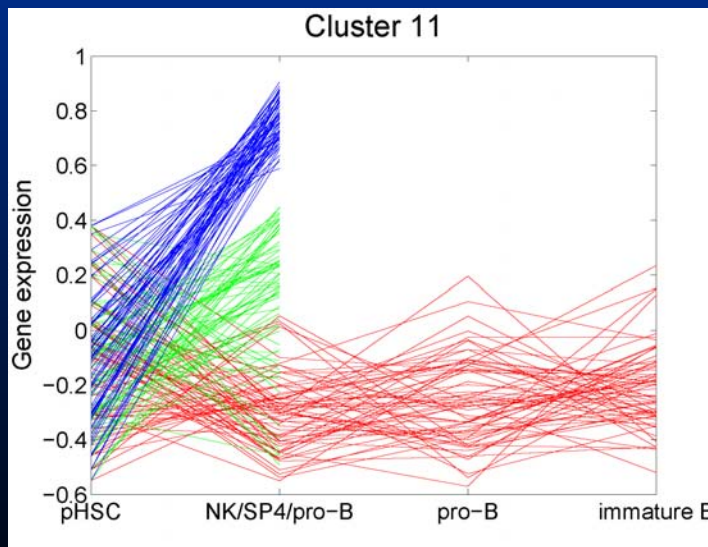
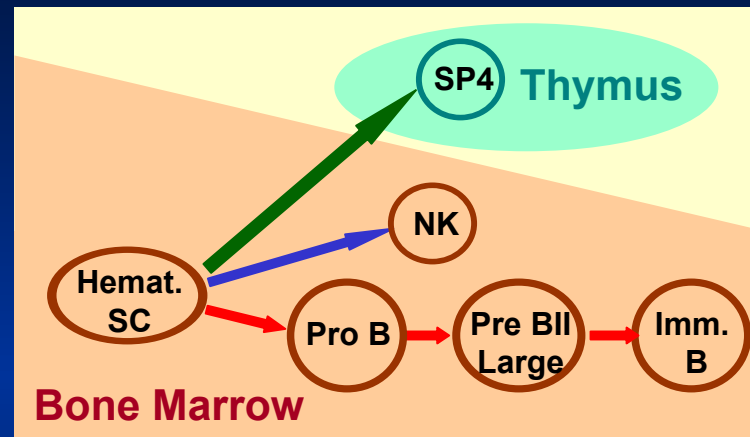
Results

Data

- T Cell [Hoffman at *et al.*]
 - 7 detailed stages of mouse t-cell development
 - 1318 genes after filtering
- Lymphoid Tree
 - 6 stages of mouse lymphoid lineage
 - from three studies [Bystrykh at *et al.*, Poirot at *et al.*, Tze at *et al.*]
 - 1321 genes after filtering



Results – Lymphoid Tree

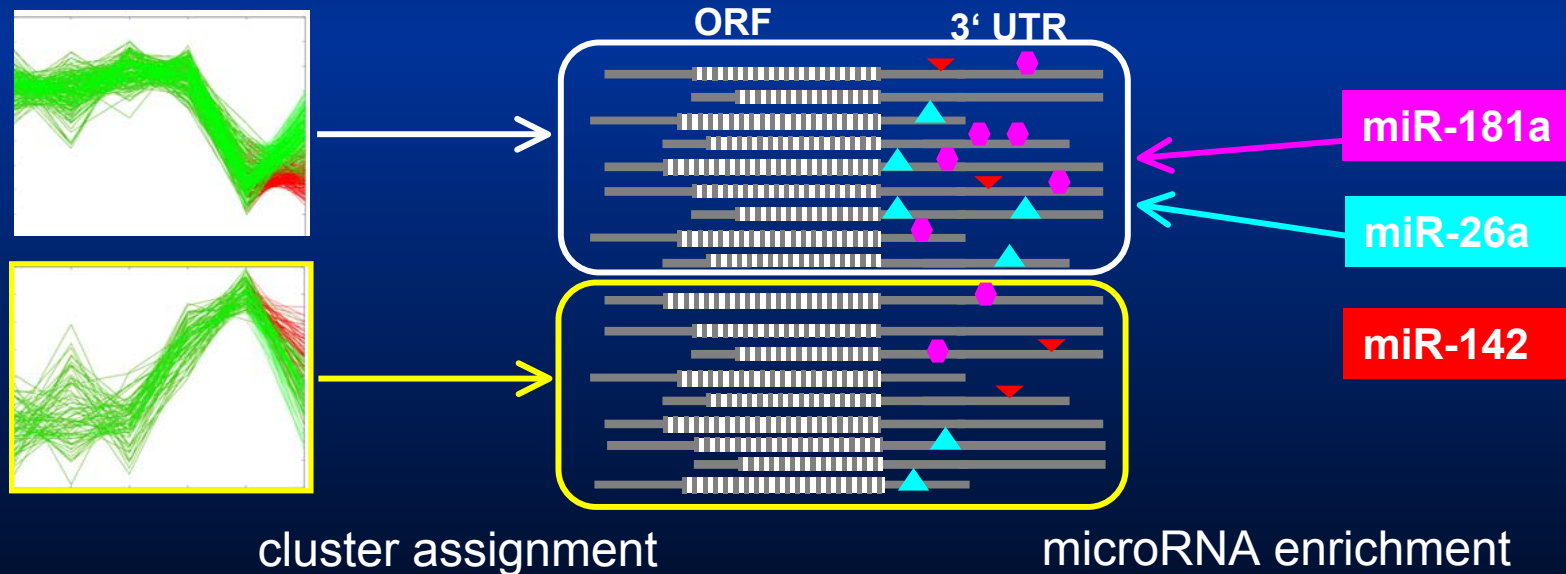


(3) Method microRNA Target Detection

MicroRNA Target Detection

MicroRNA have post-transcriptional roles by repression of target genes

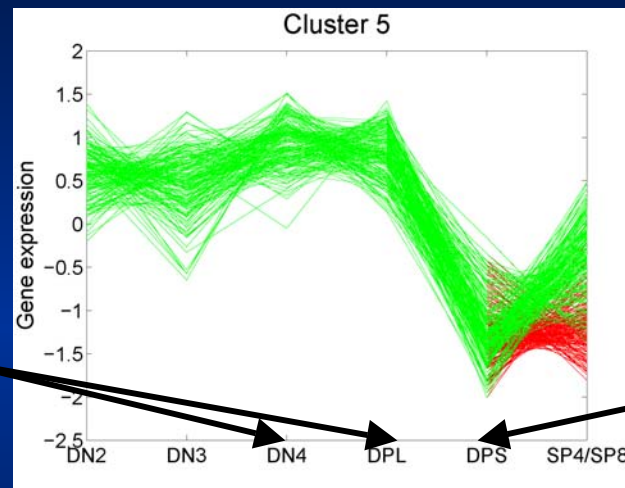
- important in hematopoiesis [Chen *et al.* 2004]
- degradation of targets transcripts [Lin, 2005, Huang, 2006]



Sequence based target detection with mirRanda [Enright, 2003]

Results - microRNA Targets

Receptor rearrangements
proliferating cells



resting cells

microRNA	microRNA Targets
miR-15a, miR-181a, miR-221, miR-24, miR-26a	2410015N17Rik, Alad, Atpif1, Aurkb, Cdc25a, Chek1, Cks1b, Cks2, Eed, H2afx, Kpnb1, Mcm5, Nasp, Pex7, Psmc12, Ranbp5, Rars, Tk1, Trip13, Uchl5

Results - microRNA Targets

Cluster 5
2

Some Numbers

Enrichment was found in:

- 5 out of 20 clusters
- 11 out of 17 microRNAs
- 39 out of 229 predicted targets

Receptor r
prolife

g cells

microRNA	microRNA Targets
miR-15a, miR-181a, miR-221, miR-24, miR-26a	2410015N17Rik, Alad, Atpif1, Aurkb, Cdc25a, Chek1, Cks1b, Cks2, Eed, H2afx, Kpnb1, Mcm5, Nasp, Pex7, Psmd12, Ranbp5, Rars, Tk1, Trip13, Uchl5

Summary

- Novel framework for analysis of development
 - querying, visualization, clustering, ...
- Recovery of well known biological facts
- Discovery of putative regulatory elements
 - concise lists of microRNA targets and insights on microRNA function

Outlook

- Improve microRNA target detection
 - integrate microRNA expression in the target prediction framework [Huang *et al*, 2006]
- Structure learning of tree topologies
- Analysis of a more detailed development tree (with Fritz Melchers at MPI-Infection Biology)
 - microRNA targets validation

Acknowledgments

- *Christopher Hafemeister* - implementations
- *Fritz Melchers* (MPI for Infectious biology) for helpful discussions and encouragement.
- CNPq (Brasil) and DAAD (Germany) for funding.

Poster L32 - Today

Thanks.

Gene Expression Trees in Blood Cell Development



Ivan G. Costa*, Stefan Rospcke*, Alexander Schliep
 Algorithmics Group, Computational Molecular Biology,
 Max Planck Institute for Molecular Genetics, Berlin
<http://algorithmics.molgen.mpg.de>
 E-mail: filho@molgen.mpg.de
 * Authors contributed equally to this work.

Introduction

The regulatory processes that govern cell proliferation and differentiation are central to developmental biology. Particularly well studied in this respect is the hematopoietic system. Gene expression data of cells of various distinguishable developmental stages fosters the elucidation of the underlying molecular processes, which change gradually over time and lock cells in certain lineages. Large-scale analysis of this data requires a computational framework for tasks ranging from visualization, querying, and finding clusters of similar genes, to answering detailed questions about the functional roles of individual genes and their similarities and differences.

Gene Expression from Blood Cell Development



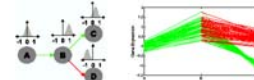
Schematic (left panel) view of blood cell development. Self-renewing hematopoietic stem cells give rise to T cell in the cortex (green), B cell in the bone marrow (blue) and natural killer cells (NK) in intermediate stages. The expression data sets are plotted in this and the next panel as follows: green lines for T cell (T); blue lines for B cell (B); and pink lines for NK cell (NK).

Tree Model

We present a statistical framework designed to analyze gene expression and further heterogeneous data such as miRNA binding as it is collected during the course of development. We extend conditional trees to continuous variables [5]. The main idea behind conditional probability trees (CPT) is to approximate a d continuous variables distribution by a product of $d - 1$ second order distributions. For example, we can approximate the joint probability distribution function (PDF) of four random variables (A, B, C, D) given the CPT in the figure below, by

$$P(A, B, C, D) = P(A)P(B|A)P(C|A, B)P(D|A, B, C) \quad (1)$$

These trees model differentiation with their inherent dependencies naturally, and enable data visualization and querying.



In the left, we depict a simple development tree, where arrows represent dependencies between variables. In the right, we display the gene expression values (y-axis) in the distinct development stages (x-axis). Above each tree variable, we have a distribution related to it and the profiles depicted in the right are in accordance to this model. In the use of conditional Gaussian density functions [6] to represent the conditional densities, from Eq. 1, hence, for a given gene expression profile (a, b, c, d) , the pdf for CPT (4) takes the form

$$p(a, b, c, d) = \frac{1}{\sqrt{(2\pi)^4 |C|}} \exp\left(-\frac{1}{2} (x - \mu)^T C^{-1} (x - \mu)\right) \quad (2)$$

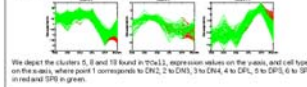
where $\mu = (\mu_1, \mu_2, \mu_3, \mu_4)$ and C are the model parameters.

The maximum likelihood estimates (MLE) for the parameters of the conditional Gaussian are

$$\hat{\mu}_i = \frac{\sum_{j=1}^n x_j^{(i)}}{n}, \hat{\sigma}_{ij} = \frac{1}{n} \sum_{j=1}^n (x_j^{(i)} - \hat{\mu}_i)(x_j^{(j)} - \hat{\mu}_j)$$

Mixture of Trees

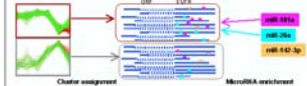
To cluster gene expression in the course of development, we combine several trees with the same, fixed topology taken from the biological literature in a classical mixture model. More formally, we combine a set of k trees in a mixture model $p(x) = \sum_{i=1}^k \alpha_i p_i(x)$, where p_i is a gene expression profile, α_i are the parameters of the LP CPT and $\alpha_i = \frac{1}{k}$ the mixing coefficient \propto the number of genes profiles assigned to that CPT.



We depict the clusters 5, 8 and 18 found in T cell, B cell, and NK cell, respectively. In the x-axis, where point 1 corresponds to DN1, 2 to DN2, 3 to DN3, 4 to DN4, 5 to DN5, 6 to DN6, 7 to DN7, 8 to DN8, 9 to DN9, 10 to DN10, 11 to DN11, 12 to DN12, 13 to DN13, 14 to DN14, 15 to DN15, 16 to DN16, 17 to DN17, 18 to DN18, 19 to DN19, 20 to DN20, 21 to DN21, 22 to DN22, 23 to DN23, 24 to DN24, 25 to DN25, 26 to DN26, 27 to DN27, 28 to DN28, 29 to DN29, 30 to DN30, 31 to DN31, 32 to DN32, 33 to DN33, 34 to DN34, 35 to DN35, 36 to DN36, 37 to DN37, 38 to DN38, 39 to DN39, 40 to DN40, 41 to DN41, 42 to DN42, 43 to DN43, 44 to DN44, 45 to DN45, 46 to DN46, 47 to DN47, 48 to DN48, 49 to DN49, 50 to DN50, 51 to DN51, 52 to DN52, 53 to DN53, 54 to DN54, 55 to DN55, 56 to DN56, 57 to DN57, 58 to DN58, 59 to DN59, 60 to DN60, 61 to DN61, 62 to DN62, 63 to DN63, 64 to DN64, 65 to DN65, 66 to DN66, 67 to DN67, 68 to DN68, 69 to DN69, 70 to DN70, 71 to DN71, 72 to DN72, 73 to DN73, 74 to DN74, 75 to DN75, 76 to DN76, 77 to DN77, 78 to DN78, 79 to DN79, 80 to DN80, 81 to DN81, 82 to DN82, 83 to DN83, 84 to DN84, 85 to DN85, 86 to DN86, 87 to DN87, 88 to DN88, 89 to DN89, 90 to DN90, 91 to DN91, 92 to DN92, 93 to DN93, 94 to DN94, 95 to DN95, 96 to DN96, 97 to DN97, 98 to DN98, 99 to DN99, 100 to DN100.

MicroRNA Target Prediction

Strategy to identify miRNAs and their target genes overrepresented in genes of co-expressed genes (indicated left) as part of a post-transcriptional regulatory mechanism. In the middle, we compare miRNAs clustered according to our mixture results are depicted. We search for potential miRNA binding sites in their 3'UTR using standard sequence analysis tools [7], and look for clusters with statistically significant number of targets for given miRNA.



Results

Computational results for a wide range of data from the hematopoietic system demonstrate the large biological relevance of our framework. We recover well known biological facts and also identify pathways but connecting regulatory elements, genes and functional assignments. In special, our results support that some miRNAs, which have been previously related to hematopoiesis [8], have a regulatory role in reducing the transcript levels of genes that are important for cell proliferation.

References

[1] M. G. Costa et al. A gene expression tree model. *PLoS ONE* 10(12): e0152000, 2015.
 [2] S. Rospcke et al. A gene expression tree model. *PLoS ONE* 10(12): e0152000, 2015.
 [3] S. Rospcke et al. A gene expression tree model. *PLoS ONE* 10(12): e0152000, 2015.
 [4] S. Rospcke et al. A gene expression tree model. *PLoS ONE* 10(12): e0152000, 2015.
 [5] S. Rospcke et al. A gene expression tree model. *PLoS ONE* 10(12): e0152000, 2015.
 [6] S. Rospcke et al. A gene expression tree model. *PLoS ONE* 10(12): e0152000, 2015.
 [7] S. Rospcke et al. A gene expression tree model. *PLoS ONE* 10(12): e0152000, 2015.
 [8] S. Rospcke et al. A gene expression tree model. *PLoS ONE* 10(12): e0152000, 2015.

Conditional Prob. Trees

Definition (1)

The prob. density function of a CPT is

$$p[X | \theta] = \prod_u^D p[X_u | X_{\text{pa}(u)}, \theta]$$

where $X = \{X_1, \dots, X_j, \dots, X_D\}$

X_j takes expression values of stage j

$\text{pa}(u) = v$ for $1 < v < u \leq D$

θ are the tree parameters

Prob. Conditional Trees

Definitions (2)

$$P[x_u | x_v, \theta] = \frac{1}{Z} \exp\left(\frac{-(x_u - \mu_u - w_{u|v}(x_v - \mu_v))^2}{2\sigma_{u|v}^2}\right)$$

- MLE estimates:

$$\mu_u = \bar{x}_u$$

$$w_{u|v} = \text{cov}(x_u, x_v) / \text{var}(x_v)$$

$$\sigma_{u|v}^2 = \text{var}(x_v) - w_{u|v}^2 \text{var}(x_u)$$

- MAP estimates:

$$w_{u|v} : \text{Normal}(0, N\beta_{u|v} \text{var}(x_v)^{-1})$$

$$w_{u|v}^* = \text{cov}(x_u, x_v) / (\text{var}(x_v) + \beta_{u|v}^{-1})$$

- With empirical Bayes

$$\beta_{u|v} = N / \left(\frac{\text{var}(x_u) \text{var}(x_v)}{\text{cov}(x_u, x_v)^2} - 1 \right)$$

Mixture Model Estimation

We want to maximize:

$$P[X | \Theta] = \prod_{i=1}^N \sum_{j=1}^K \alpha_j \cdot P[x_i | \theta_j]$$

By adding a hidden variable Y , we obtain:

$$\begin{aligned} P[X, Y | \Theta] &= P[X | Y, \Theta] P[Y | \Theta] \\ &= \prod_{i=1}^N \prod_{j=1}^K (\alpha_j \cdot P[x_i | \theta_j])^{r_{ij}} \end{aligned}$$

where $Y = \{y_i\}_{i=1}^N$ and $y_i \in \{1, \dots, K\}$

$$r_{ij} = P[y_i = j | x_i, \theta_j]$$

EM for Mixture Estimation

- Input:
 - genes profiles X and the number of clusters k
- Initialization:
 - Randomly assign the posterior probabilities $P[y_i = j | \mathbf{x}_i, \theta_j^0]$
- Iterate (until convergence):
 - Re-estimate Θ^t given $P[y_i = j | \mathbf{x}_i, \theta_j^{t-1}]$ (**M Estep**)
 - Calculate $P[y_i = j | \mathbf{x}_i, \theta_j^t]$ given Θ^t (**E Estep**)

From Mixtures to Groups

1. Maximum posterior assignment:

$$y_i = \arg \max_{1 \leq j \leq K} (\mathbb{P}[y_i = j | x_i, \theta_j])$$

2. Entropy cut-off assignment:

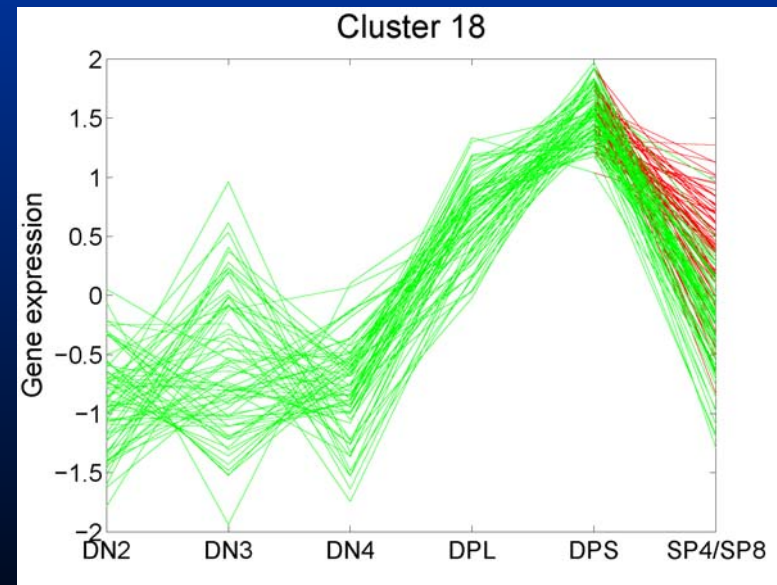
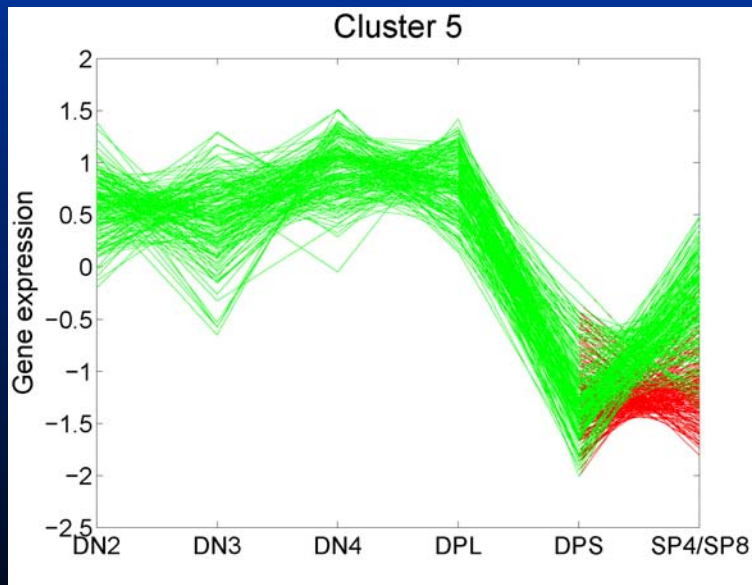
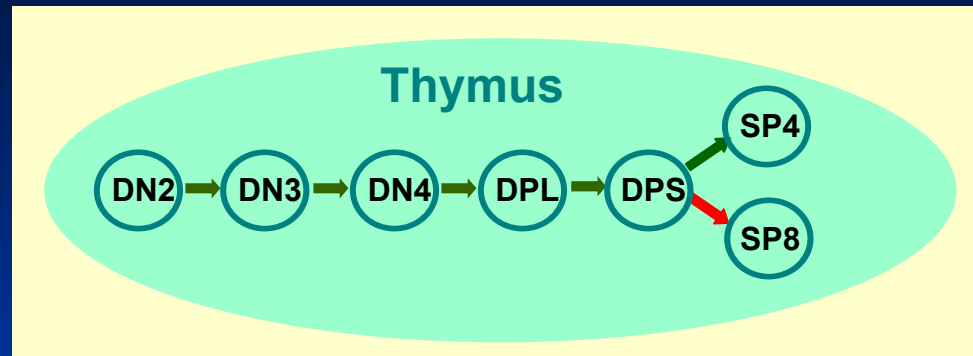
- if $H(\{\mathbb{P}[y_i = j | x_i, \theta_j]\}_{j=1}^K) < \varepsilon$

$$y_i = \arg \max_{1 \leq j \leq K} (\mathbb{P}[y_i = j | x_i, \theta_j])$$

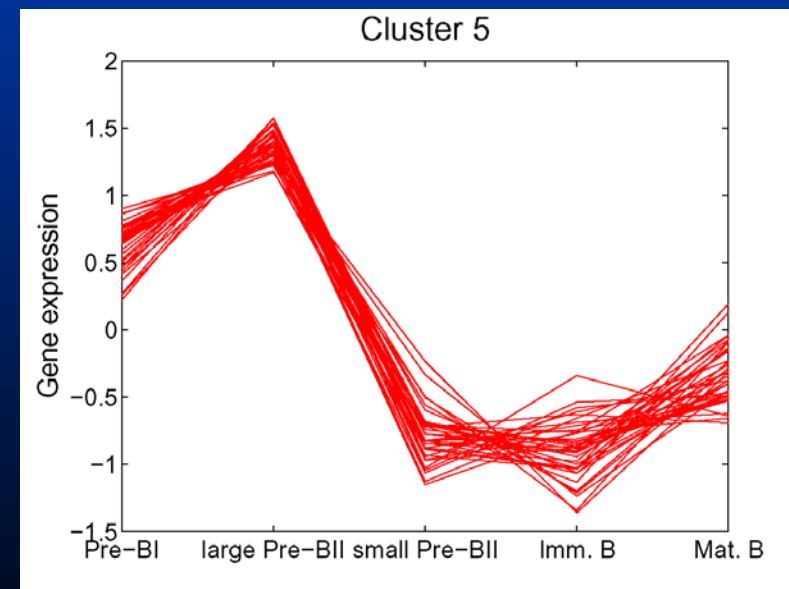
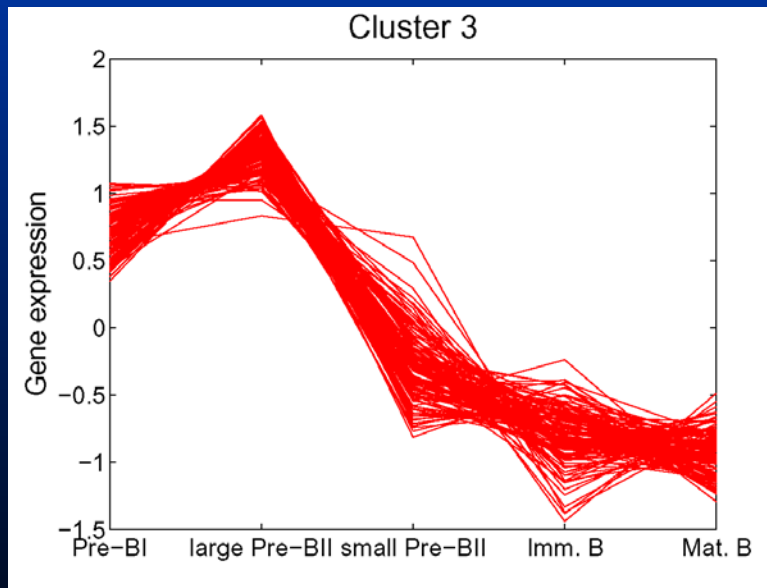
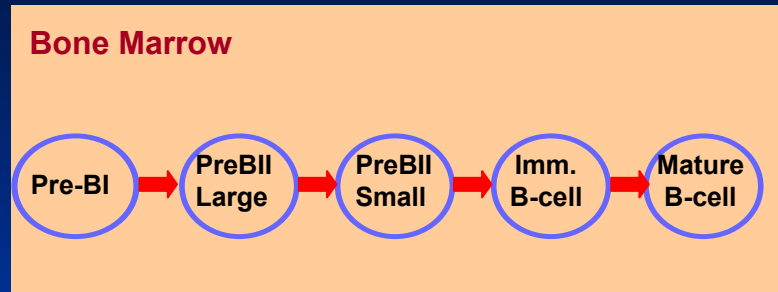
- else

$$y_i = K + 1$$

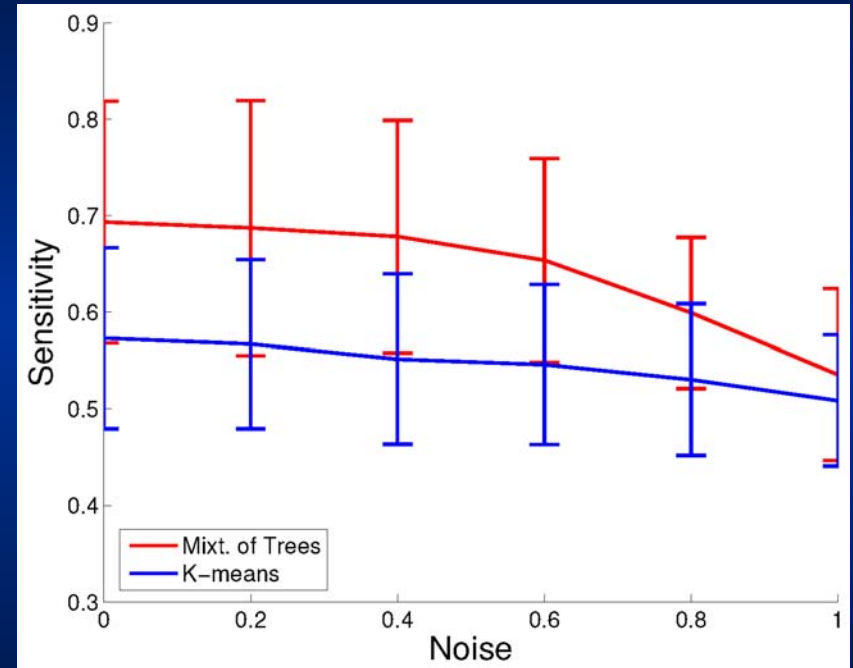
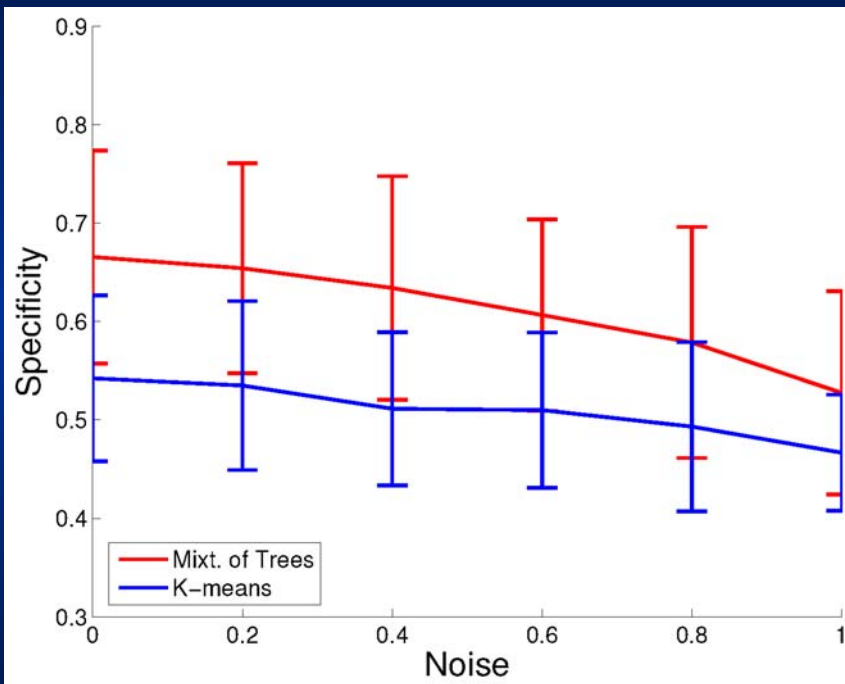
Results – T Cell



Results – Bcell



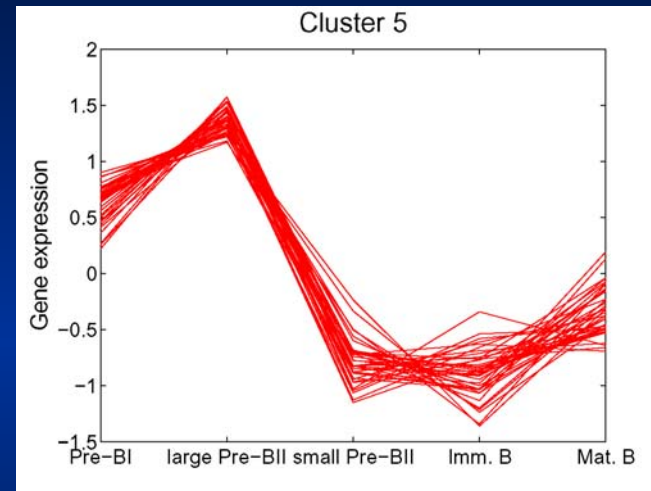
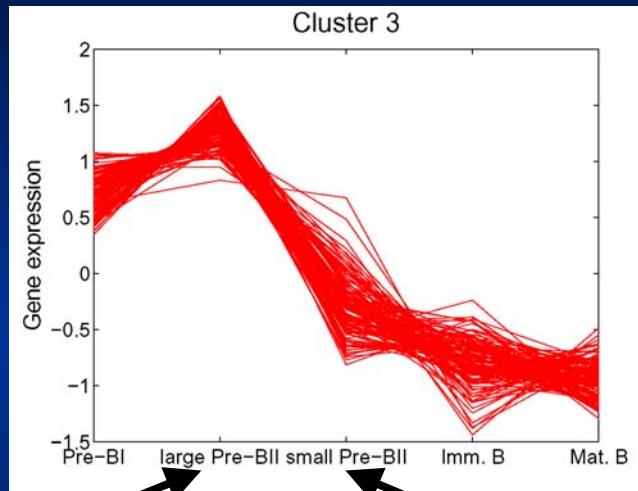
Results - Simulated Data



• Simulated Data

- random parameters for a given mixture of trees
- addition of independent noise

Results - microRNA Targets

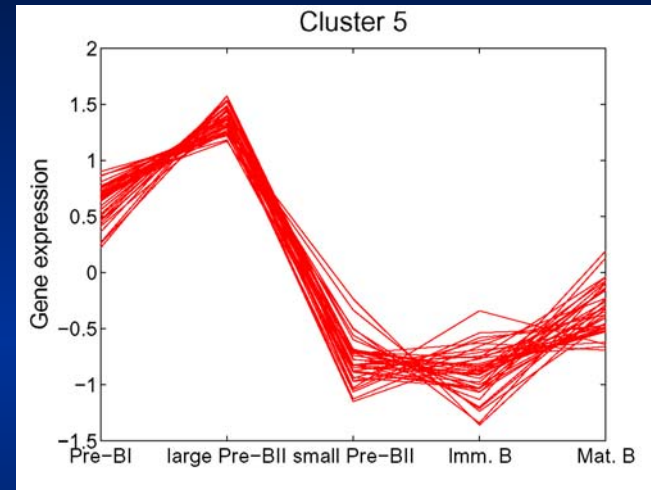
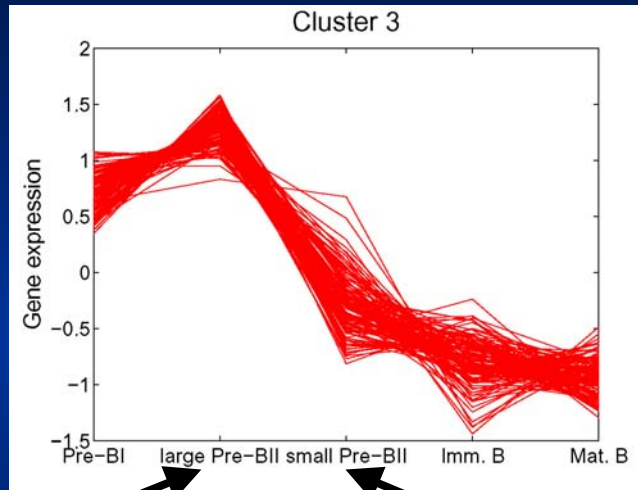


proliferating cells

resting cells

microRNA	microRNA Targets
miR-181b, miR-181c miR-26a	Atpif1, Aurkb, Cbx1, Cdc45l, Cks1b, Cks2, Cox5a, Hmgb2, Melk, Ttk, Uchl5
miR-15a, miR15b, miR-221, miR-223	Cdca4, Chek1, Mcm4, Nasp, Nfyb, Smc4l1, Tuba2 ⁴

Results - microRNA Targets



proliferating cells

resting cells

microRNA	microRNA Targets
miR-181b, miR-181c miR-26a	<i>Atpif1</i> , <i>Aurkb</i> , <i>Cbx1</i> , <i>Cdc45l</i> , <i>Cks1b</i> , <i>Cks2</i> , <i>Cox5a</i> , <i>Hmgb2</i> , <i>Melk</i> , <i>Ttk</i> , <i>Uchl5</i>
miR-15a, miR15b, miR-221, miR-223	<i>Cdca4</i> , <i>Chek1</i> , <i>Mcm4</i> , <i>Nasp</i> , <i>Nfyb</i> , <i>Smc4l1</i> , <i>Tuba2</i>