Clustering Biological Data Using Mixture Models

Alexander Schliep

Department of Computer Science
BioMaPS Institute for Quantitative Biology
Rutgers University

Outline

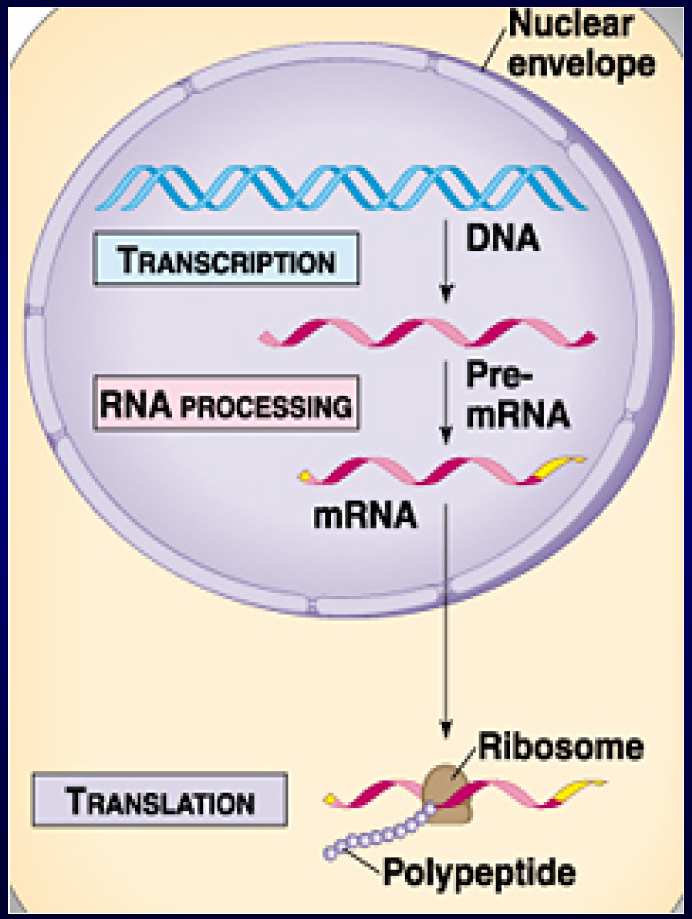
1. **Motivation**
   Gene function and regulation, functional modules
2. **Mixture models**
   Definition, examples and motivation
3. **Case study: Hematopoiesis**
   Blood cell development, sparsely parameterized models
4. **Case study: Syn-expression in Drosophila Development**
   Semi-supervised learning, heterogeneous data, molecular imaging

Motivation

Understanding life at the cellular level

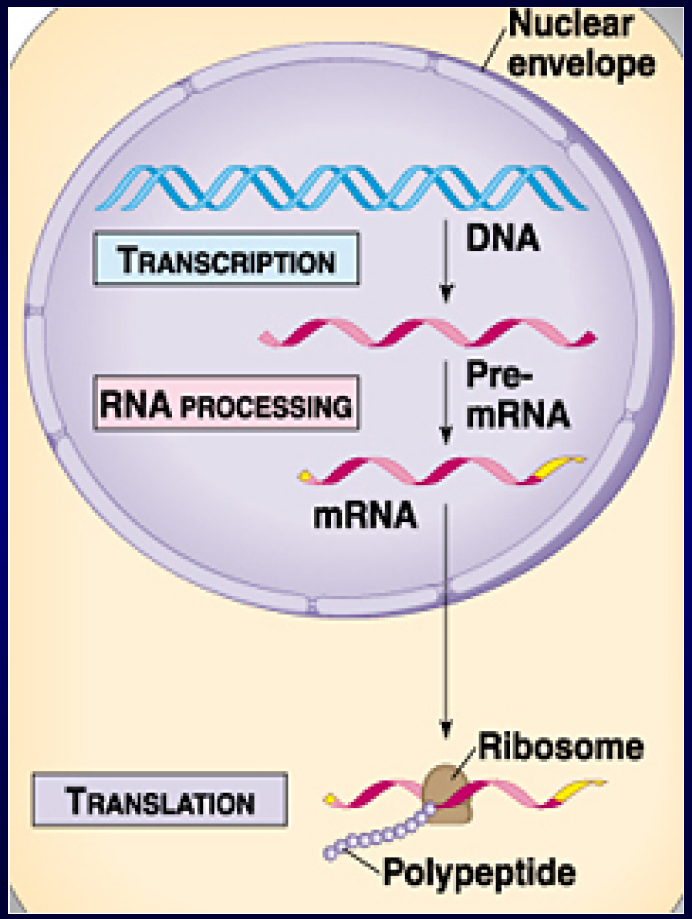
- How does genetic information influence cellular processes?
- How does environment influence cellular processes?
- What are the effects of variations in the genetic information?
- *Which genes do jointly perform which cellular function?*

Classical dogma of molecular biology



- DNA (Deoxyribonucleic acid) double-stranded chain molecule. Four-letter alphabet $\{A, C, G, T\}$
- RNA (Ribonucleic acid) single-stranded chain molecule. Four-letter alphabet $\{A, C, G, U\}$
- Proteins: polypeptides, chain of amino acids. Twenty-letter alphabet

Classical dogma of molecular biology



- Dogma: Information flow DNA $\longrightarrow$ mRNA $\longrightarrow$ Protein
- Gene: segment (substring) of DNA coding for a protein
- Assumption: one-to-one correspondence of gene, protein *and* function

Complexity of gene function

- *C. elegans*: ~19,000 genes
- *D. melanogaster*: ~13,500 genes
- *H. sapiens*:
  - Highest guesses before Human Genome:
    › 140,000 genes
  - Today: ~21.000 protein-coding genes

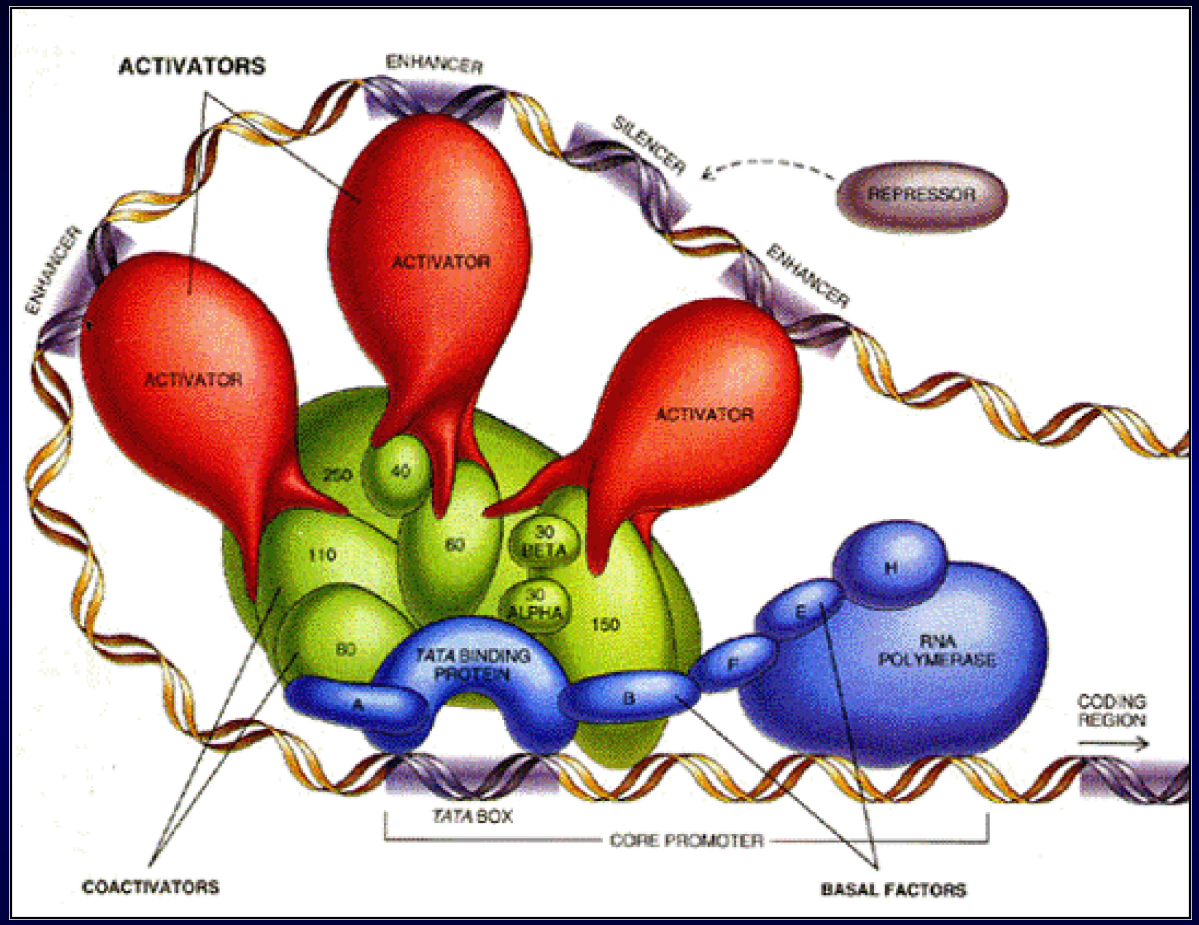
Question

Where does the apparent complexity of Humans come from?

Complexity of gene function

- Alternative splicing (Eukaryotes):
  one gene codes for *multiple* proteins
- Complex formation:
  groups of proteins *jointly* perform a cellular function
- Gene regulation:
  Control of number and type of proteins produced from a gene.

Transcriptional regulation



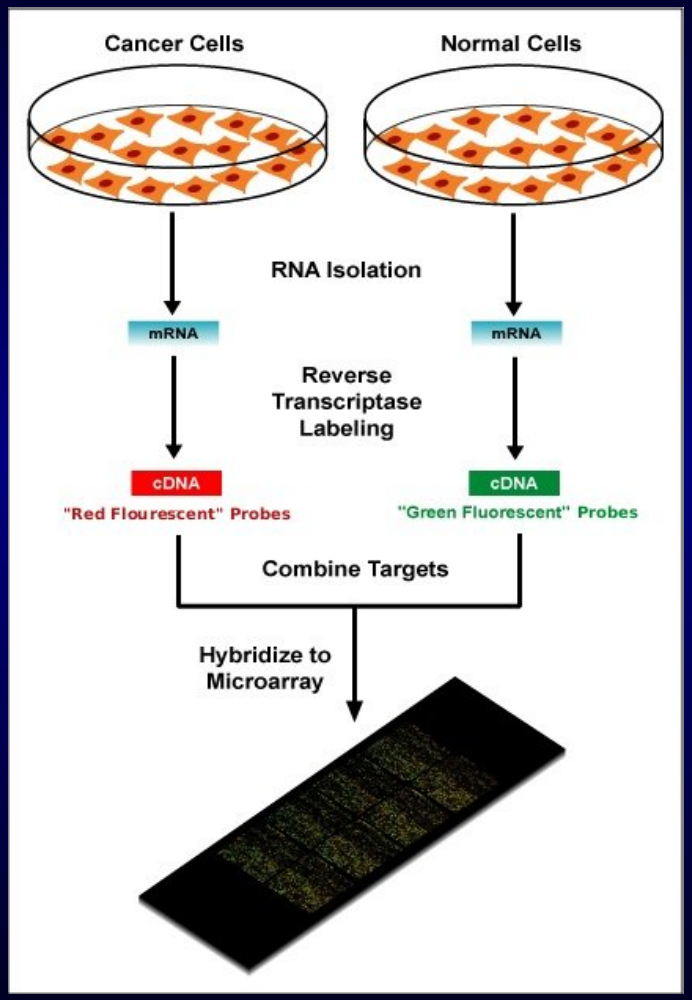
- Control of transcription mechanism
- Molecular mechanism:
  Binding of proteins to DNA

Functional modules

Functional module: sets of genes jointly associated with a cellular function

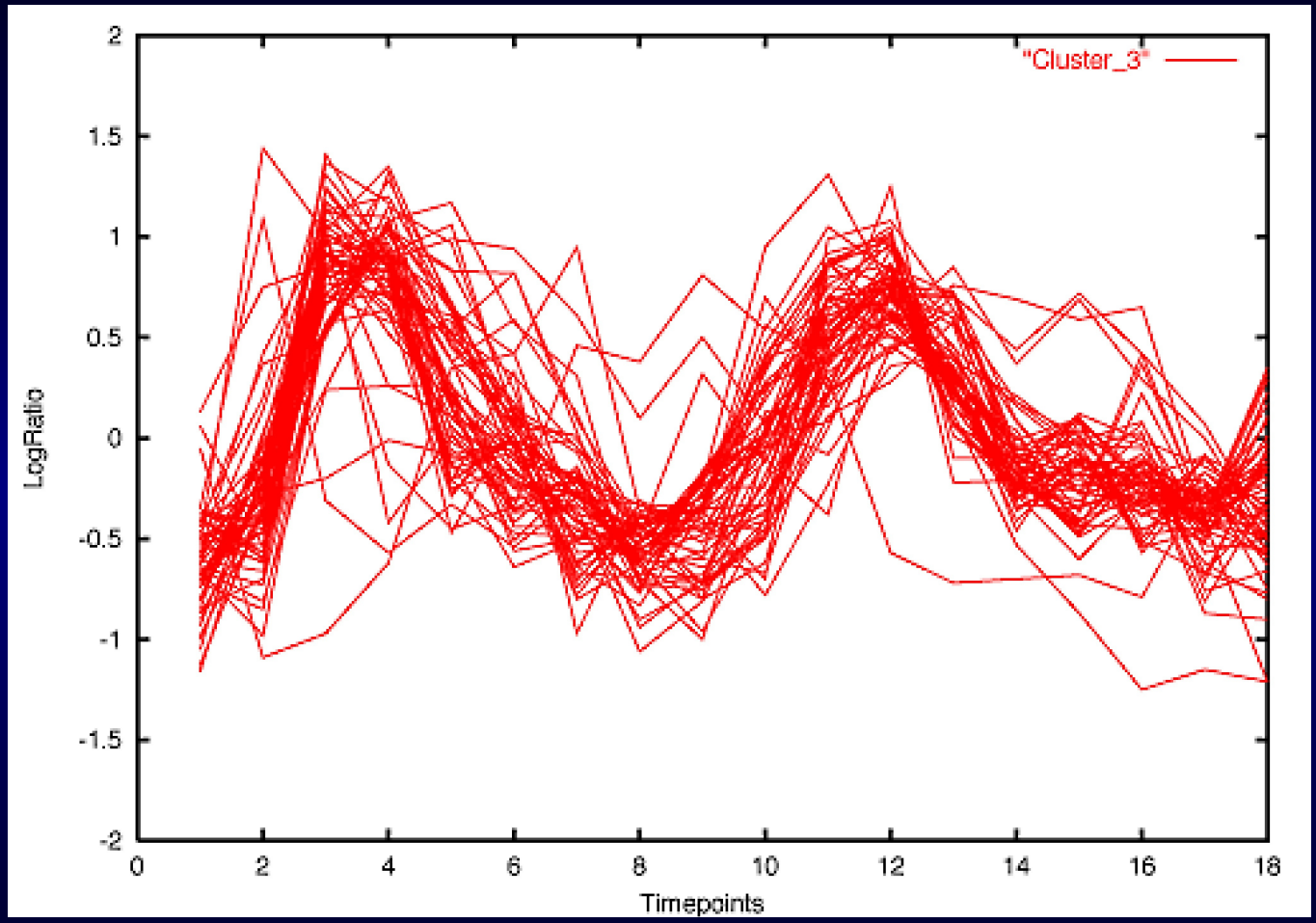
- Assumption: genes share regulators
- Assumption: genes detectable by similarity of transcript levels
- Detection by Clustering

DNA Microarrays



- *One* experiment for mRNA levels or expression of 10,000s of genes
- Comparing populations: Differential gene expression
- Time-courses
  - Cell-cycle
  - Disease progression
  - Reaction to external factors
- Caveat: Data quality

Functional modules: Example



Validity of findings: Internal and external indices

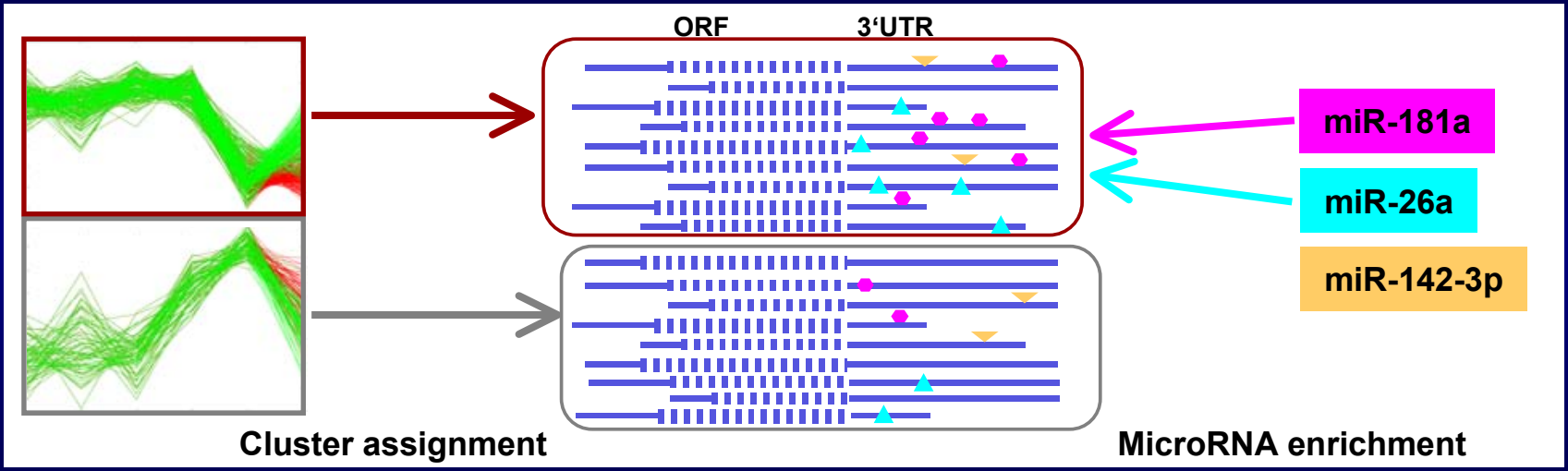
Data: Spellman (1998) Yeast cell cycle data

Functional modules: External validation

Enrichment analysis:

- Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) provide functional annotation
- Check: more genes with same annotation in clusters than expected?
- Hyper-geometric test/Fisher's exact test assess significance

Functional modules: Identifying regulators

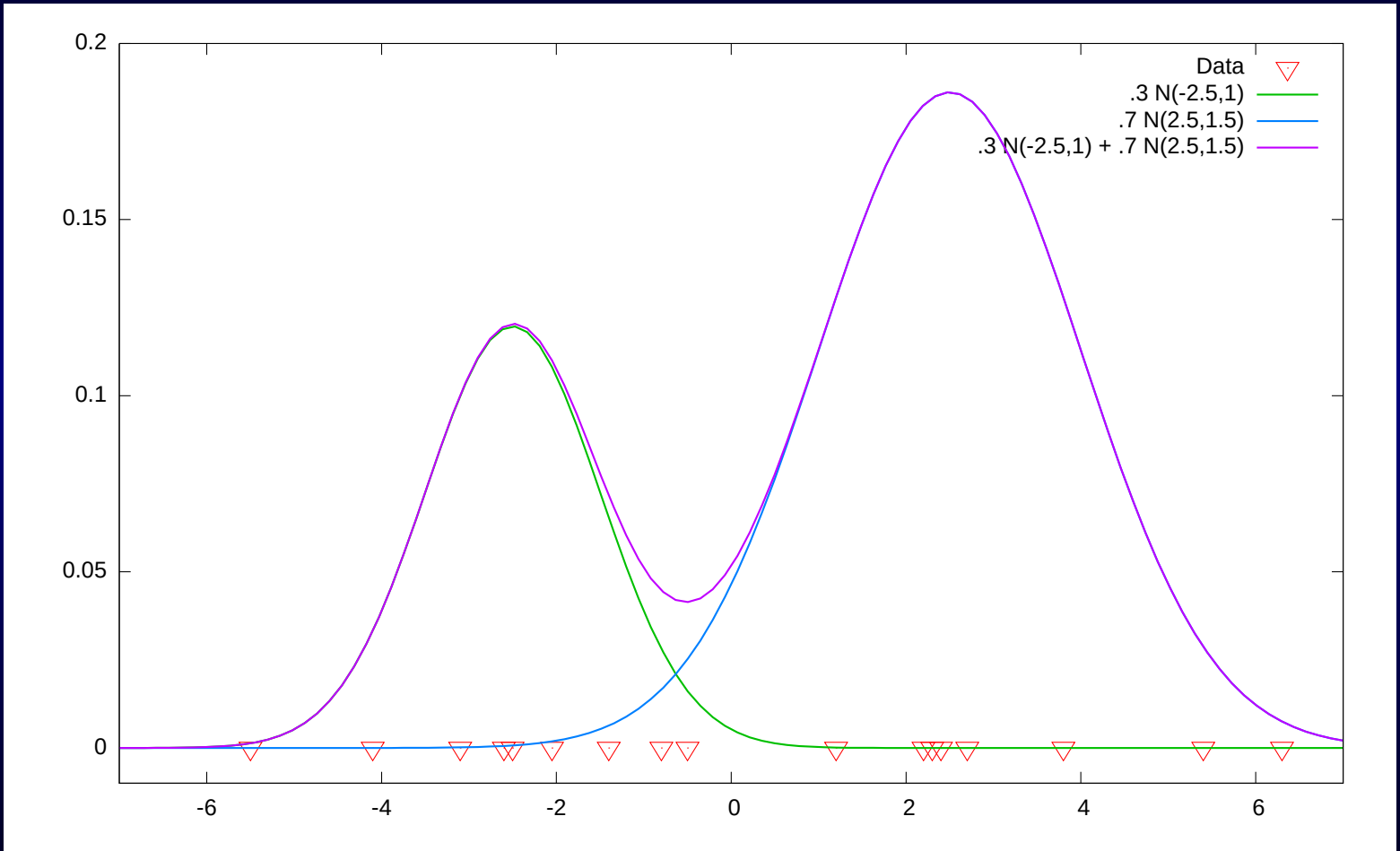


Theme

Discover functional modules of genes from large scale high-throughput experimental data, with a focus on *temporal* and *spatial* information.

Mixture models

Example: Mixture



Mixture model

Definition

A mixture is a convex combination of density functions f_k

$$f(x|\Theta) = \sum_{k=1}^{K} \alpha_k f_k(x|\theta_k)$$

Notation: $x \in \mathbb{R}^d, K \in \mathbb{N}$, θ_k component density parameters,

$\alpha_k \geq 0, \sum_{k=1}^{K} \alpha_k = 1$ mixture weights, $\Theta = (\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K)$.

Remark

Generative model: sample k according to $(\alpha_1, \ldots, \alpha_K)$, then sample $x \sim f_k(x|\theta_k)$. Note: k unobserved.

Component posterior

Definition

Posterior probability of component given data

$$P[k|x] = P[k|x,\Theta] = \frac{\alpha_k f_k(x|\theta_k)}{\sum_{l=1}^{K} \alpha_l f_l(x|\theta_l)}$$

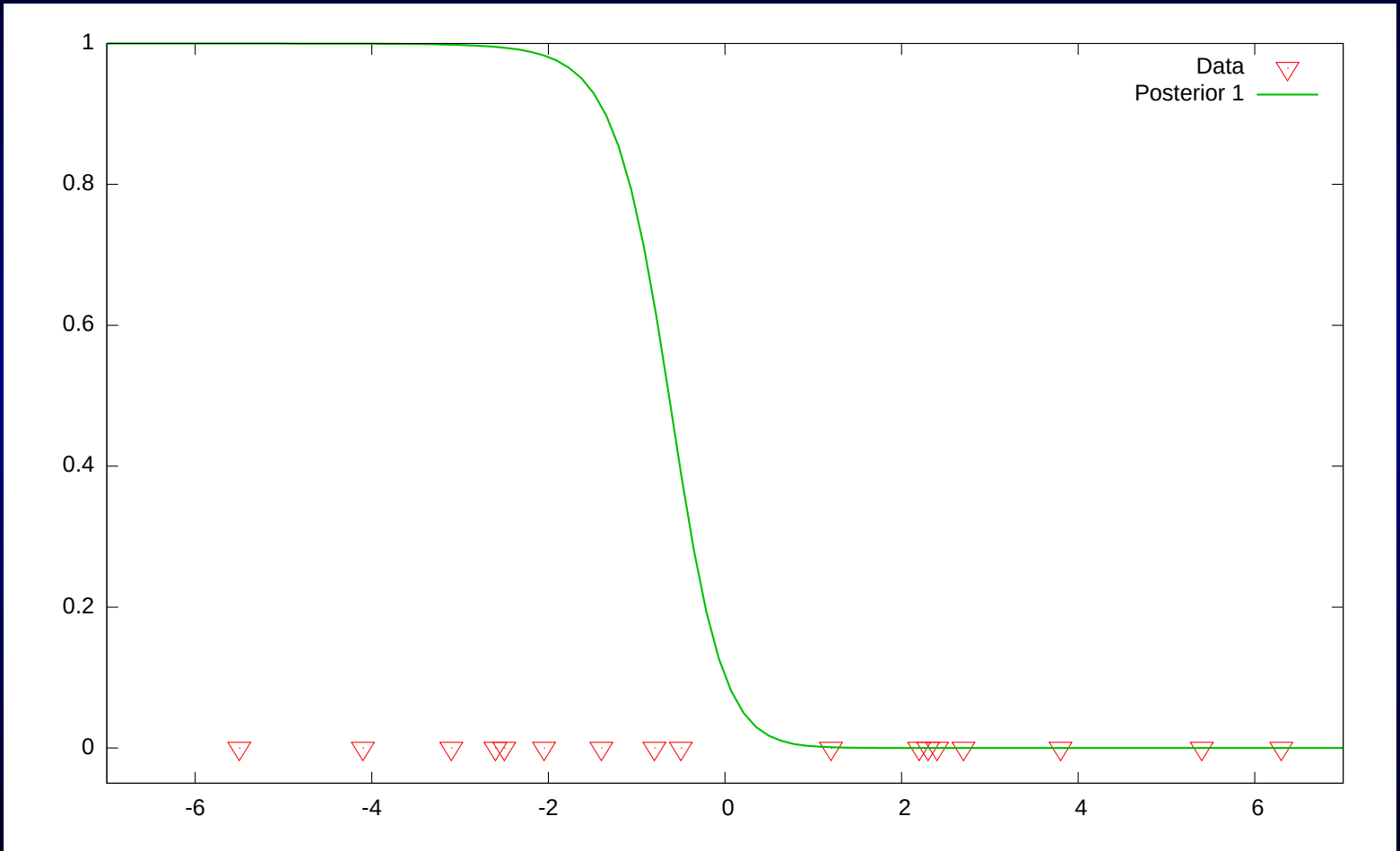
Proposition (Clusters from a mixture)

Compute K-partition $C_1,\ldots,C_K$ of $X = \{x_1,\ldots,x_n\}$, $x_i \in \mathbb{R}^d$ by

$$i \in C_k \Leftrightarrow k = \arg\max_{1\le l\le K} P[l|x_i]$$

Here: $C_k \subset \{1,\ldots,n\}$, $\dot{\bigcup}_{k=1}^{K} C_k = \{1,\ldots,n\}$

Example: Posterior



Log-likelihood

Definition

For $X = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^d$, assuming independence

$$\log L(\Theta|X) = \log \left(\prod_{i=1}^{n} f(x_i|\Theta) \right) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \alpha_k f_k(x|\theta_k) \right)$$

Goal: Find $\Theta^\star$ maximizing $\log L$
Postulate existence of variables $Y = (y_1, \ldots, y_n)$,
$y_i \in \{1, \ldots, K\}$, indicating generating component.

Complete data log-likelihood

Proposition

If Y known then

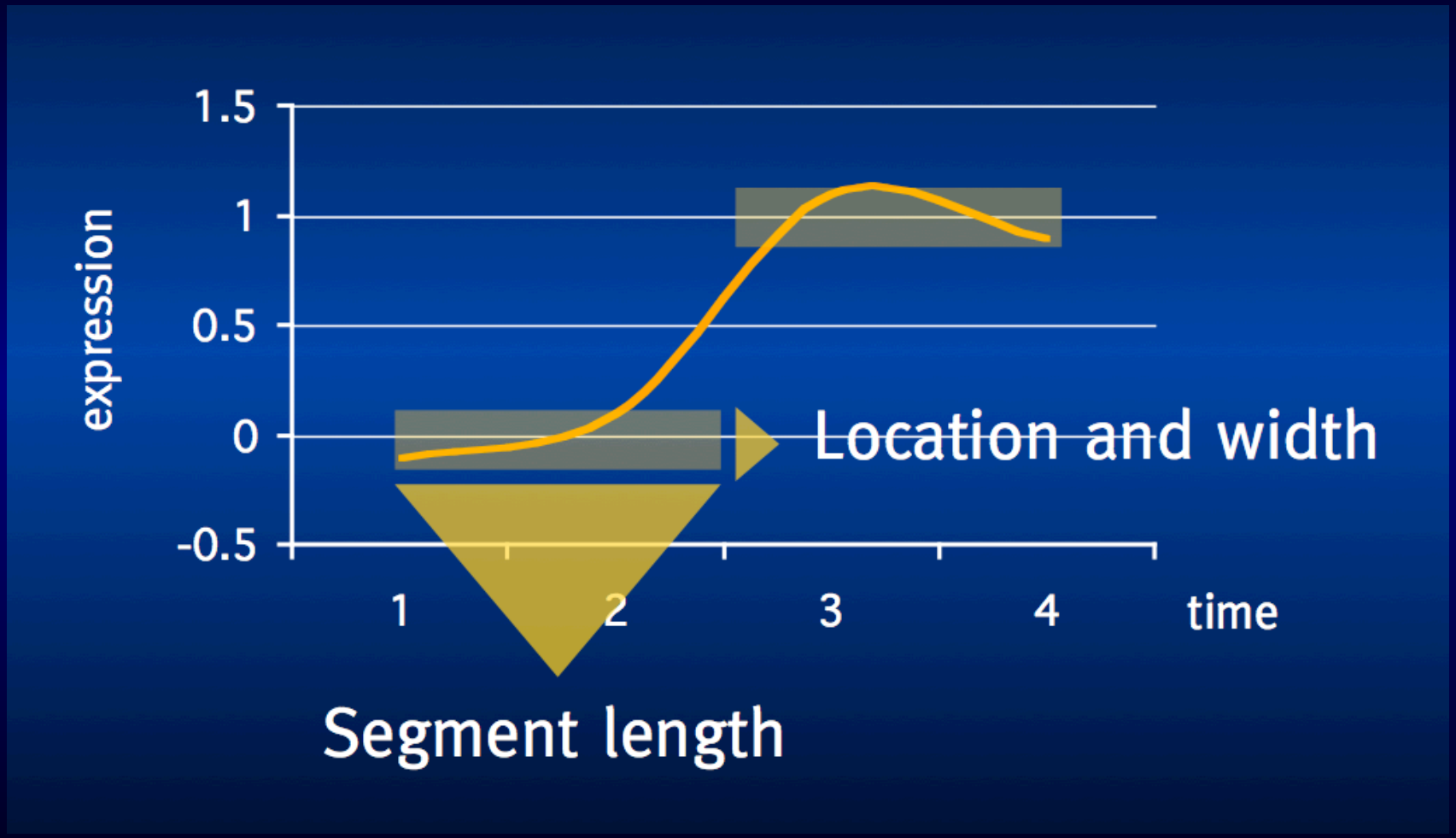
$$\log L(\Theta|X,Y) = \sum_{i=1}^{n} \log\left(\alpha_{y_i} f_{y_i}(x|\theta_{y_i})\right)$$

Algorithm (Expectation-Maximization (EM))

Iterative procedure to find local maximum of $\log L$

- *E-step: Estimate component posteriors given* Θ_t
- *M-step: Maximize the expected log-likelihood given component posteriors, yielding* Θ_{t+1}

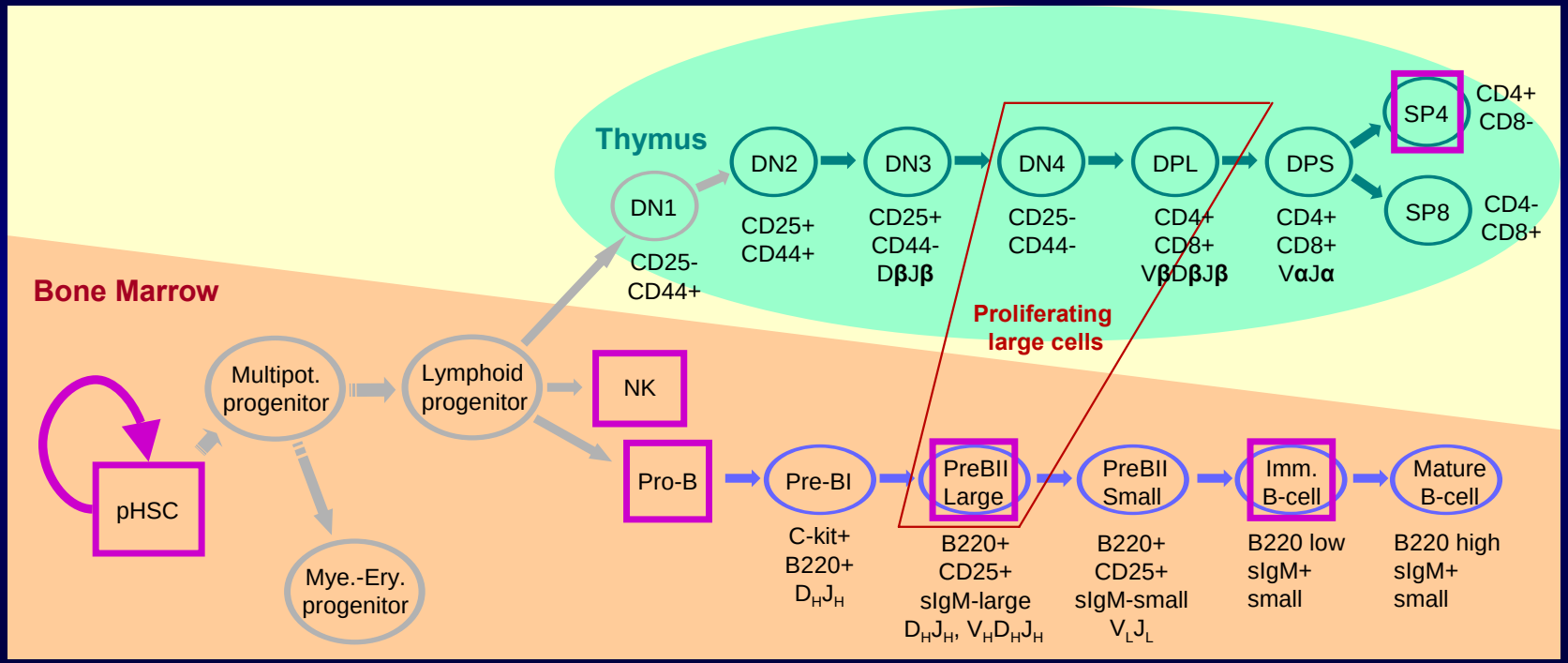
Focus: component models



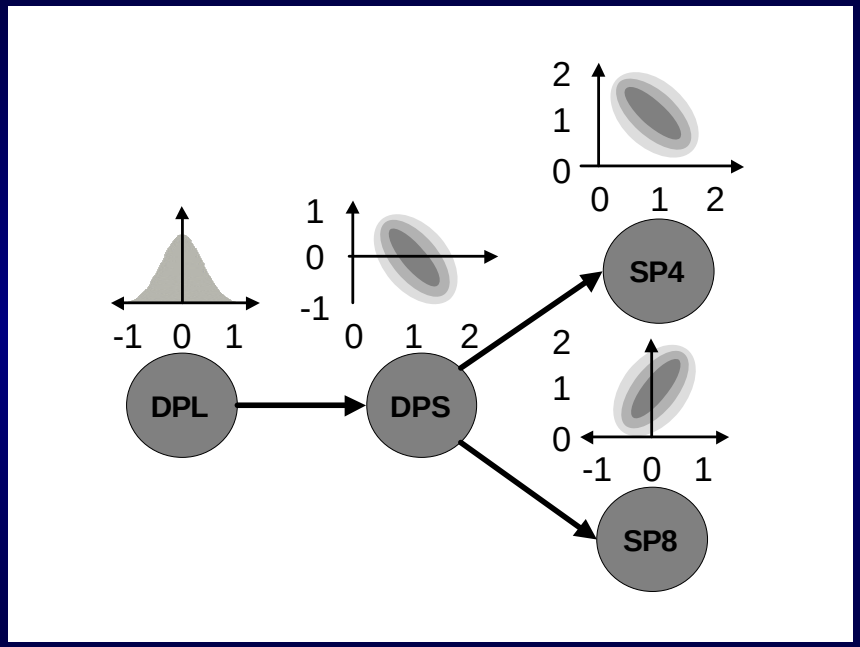
Hematopoiesis

case study

Hematopoiesis

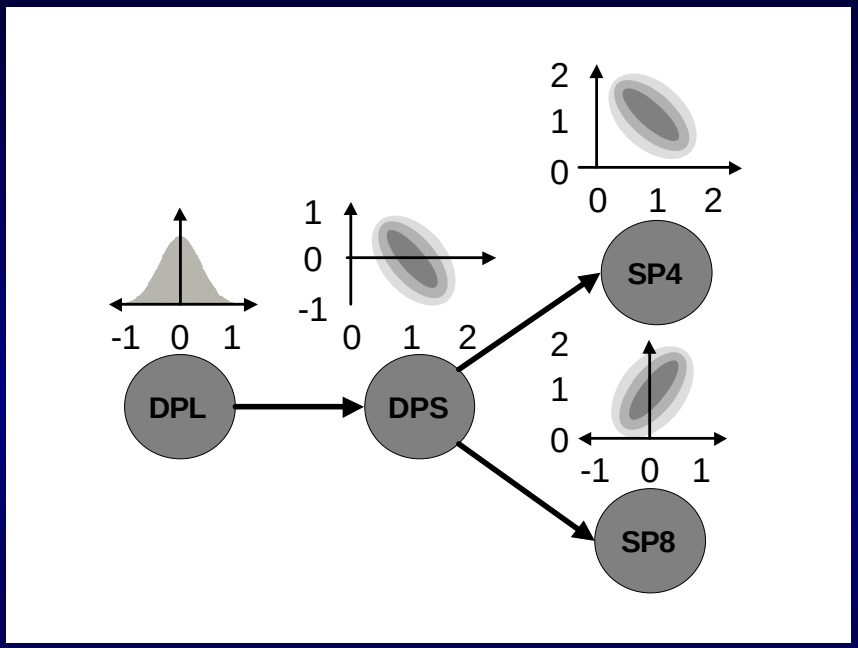


DTree



- Data: for all genes i expression levels x_i in four stages $(x_{iDPL}, x_{iDPS}, x_{iSP4}, x_{iSP8})$
- Model assumption: immediate predecessor matters most

DTree



- Data: for all genes i expression levels x_i in four stages $(x_{iDPL}, x_{iDPS}, x_{iSP4}, x_{iSP8})$
- Model assumption: immediate predecessor matters most

Hence:

$$f(x_i) = f(x_{iSP8}|x_{iDPS}) f(x_{iSP4}|x_{iDPS}) f(x_{iDPS}|x_{iDPL}) f(x_{iDPL})$$

Conditional Gaussian

Definition

Let $u, v \in \{1, \ldots, d\}$ and $\theta_{u|v} = (\mu_{u|v}, w_{u|v}, \sigma^2_{u|v})$. Then the conditional Gaussian density is

$$f(x_u | x_v, \theta_{u|v}) = \frac{1}{\sqrt{2\pi}\sigma_{u|v}} \exp\left(\frac{-(x_u - \mu_{u|v} - w_{u|v} x_v)^2 /}{2\sigma^2_{u|v}} \right).$$

Remark

- *If* $w_{u|v} = 0$ *then* $f(x_u | x_v, \theta_{u|v})$ *is Gaussian* $N(\mu_{u|v}, \sigma^2_{u|v})$
- $f(\cdot|\cdot)$ *maximal for* $x_u = w_{u|v} x_v - \mu_{u|v}$

DTree

Definition

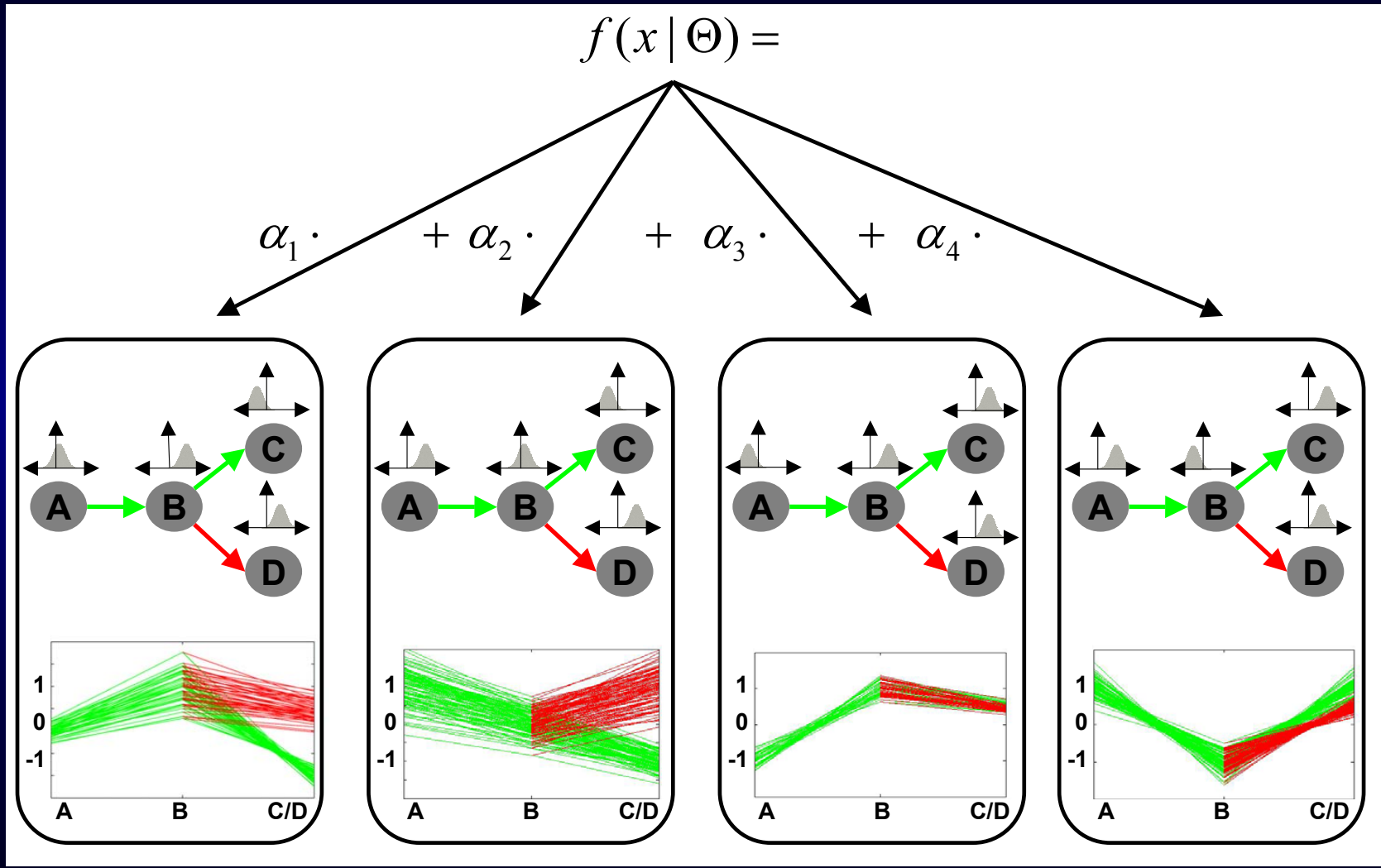
Given a tree in vertices $\{1,\ldots,d\}$ by parent map pa and $\theta = (pa, \theta_{1|pa[1]}, \ldots, \theta_{d|pa[d]})$. Then a dependence tree density is

$$f(x|\theta) = \prod_{u=1}^{d} f(x_u | x_{pa(u)}, \theta_{u|pa(u)})$$

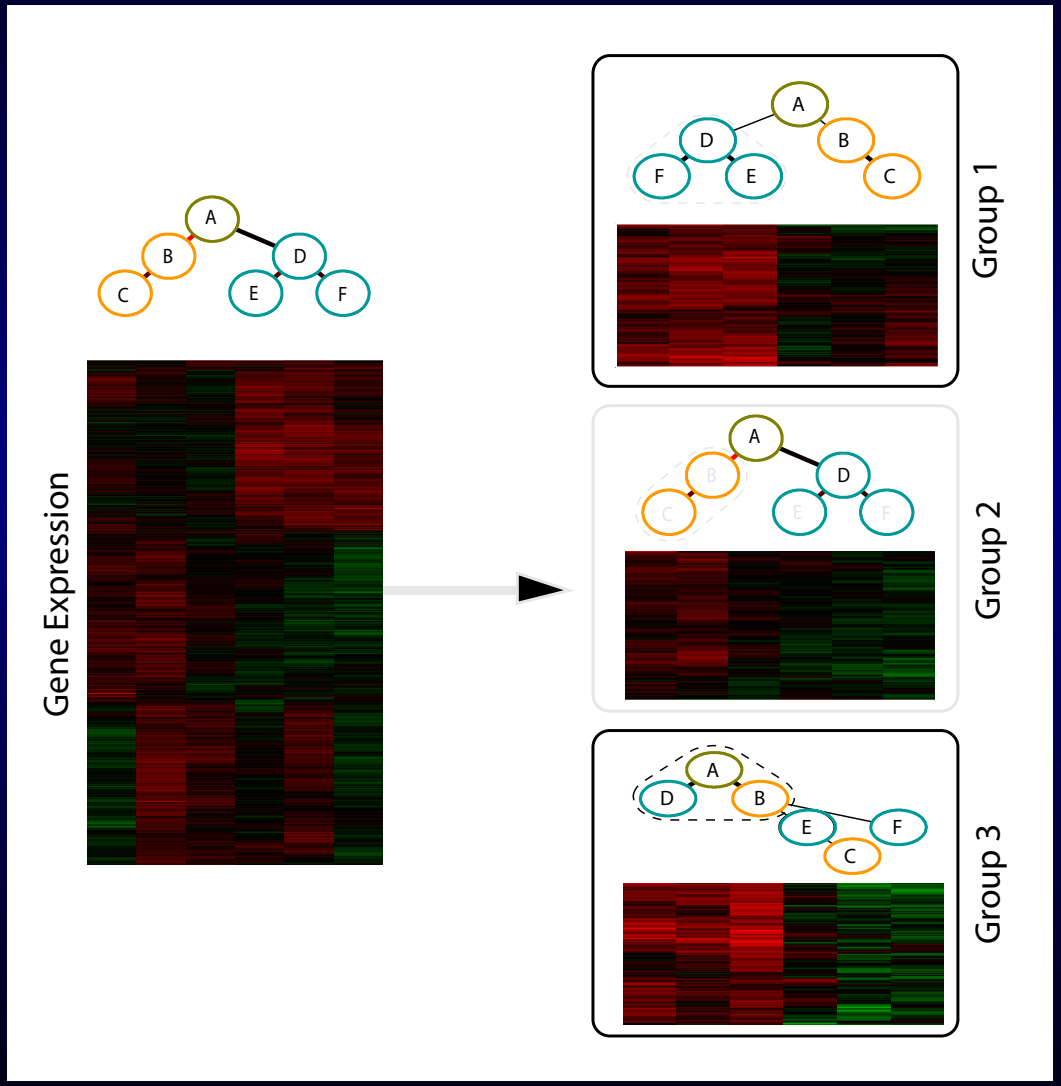
Remark

Every DTree is a multi-variate Gaussian. Number of parameters $O(d^2)$ for d-variate Gaussian and $O(d)$ for DTree

Mixture of DTrees



Module-specific structures



Inferring tree structure

Problem

Let $t(x)$ be the unknown true density and $f_{pa}(x) = f(x|\theta)$. Find

$$pa^\star = \arg\min_{pa} D(t||f_{pa}).$$

$D(t||f_{pa})$ is the relative entropy $\int_{\text{supp}(t)} t(x) \log \frac{t(x)}{f_{pa}(x)} dx$.

Lemma (Chow, Liu 1968)

Mutual information $I(X,Y) = \int_X p(x,y) \log \frac{p(x,y)}{p_x(x)p_y(y)} dxdy$.
Then

$$pa^\star = \arg\max_{pa} \sum_{i=1}^{d} I(x_u, x_{pa(u)})$$

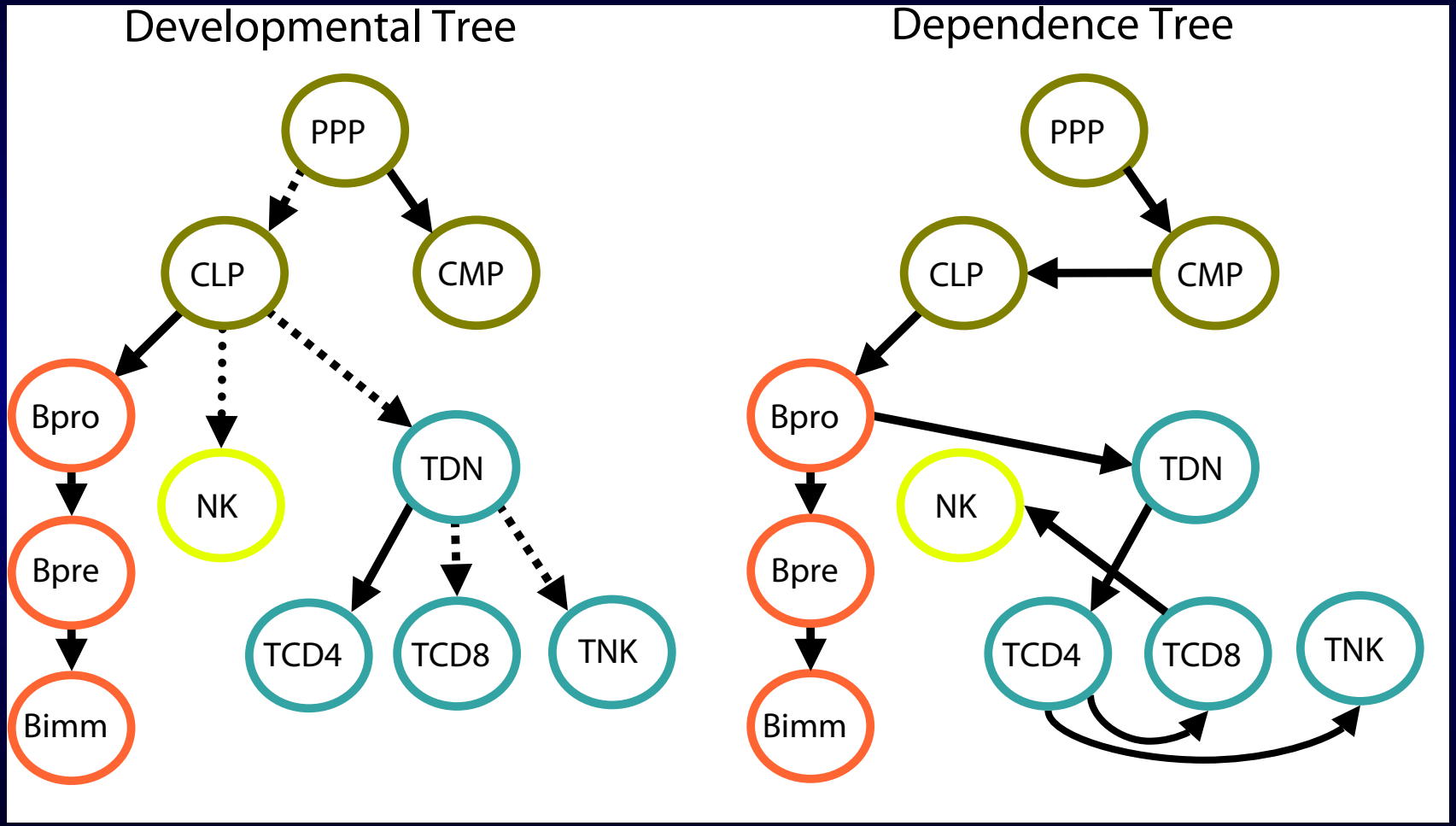
also minimizes $D(t||f_{pa})$.

Inferring tree structure

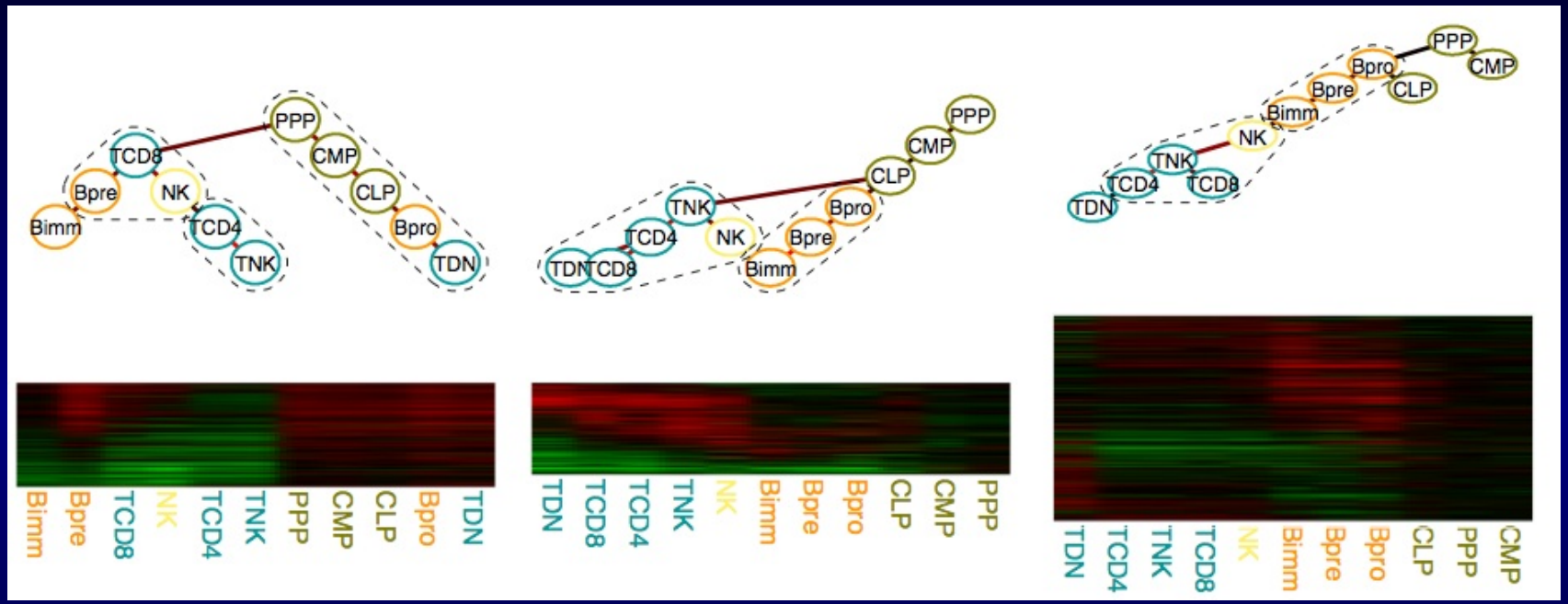
Algorithm

1. *Compute pair-wise correlation coefficients for all pairs* $(u,v) \in \{1,\ldots,d\}^2$
2. *Interpret as weighted graph*
3. *Compute a maximal spanning tree with greedy algorithm*

Recovering known developmental tree



Results: Module-specific structures

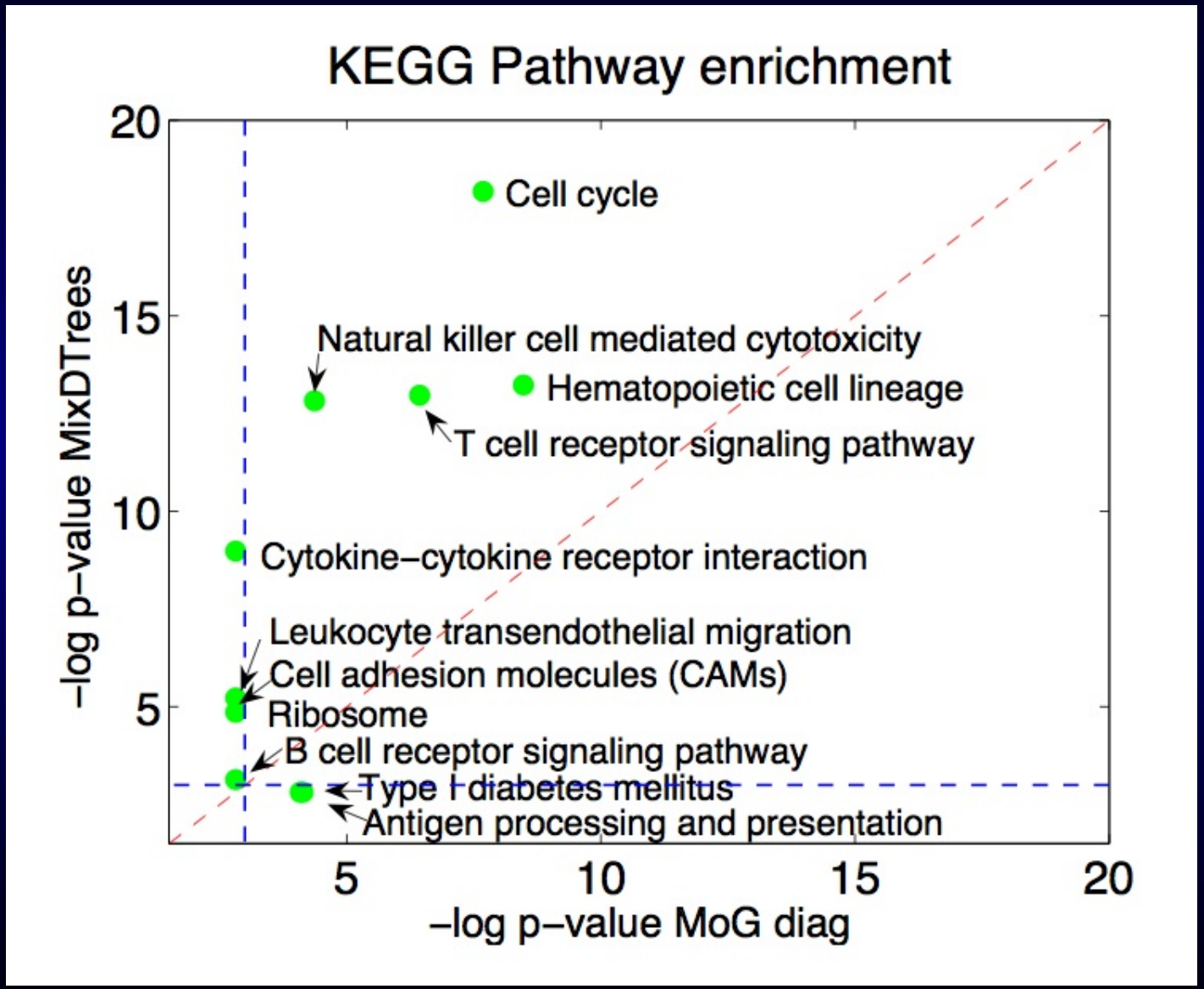


Genes activated in proliferating cells

Genes activated in T-cells

Genes activated in B-cells

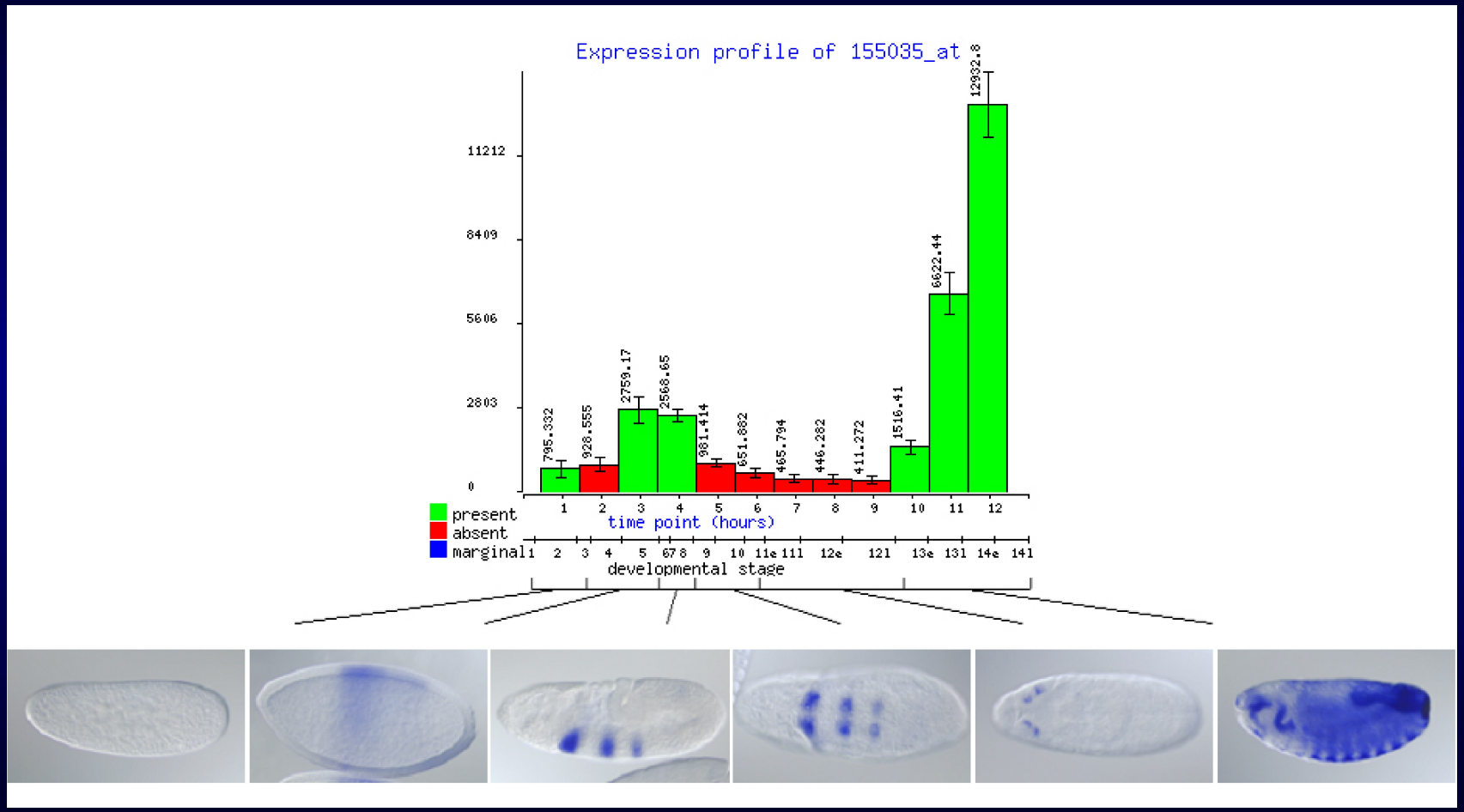
Results: External validation



Syn-expression

case study

Spatio-temporal data

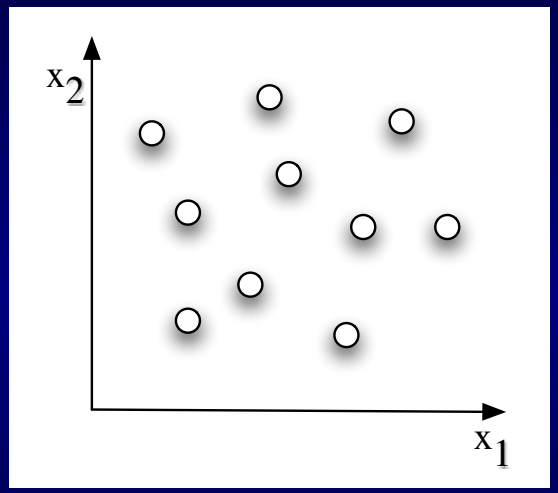


Problem

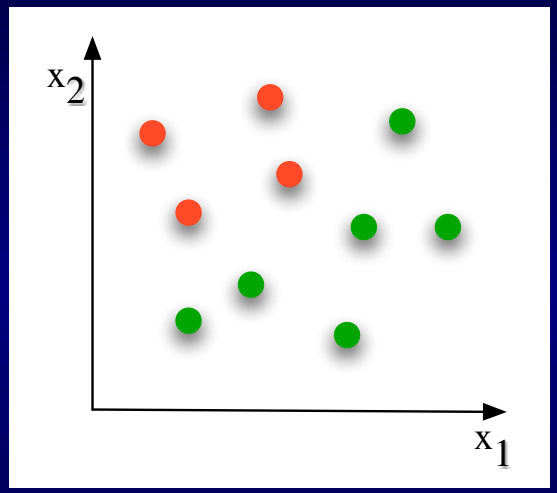
Syn-expression: temporal *and* spatial co-expression
Fuse two or more sources of heterogeneous data

- Distinct abundance?
- Data quality?
- Confidence?

Semi-supervised learning

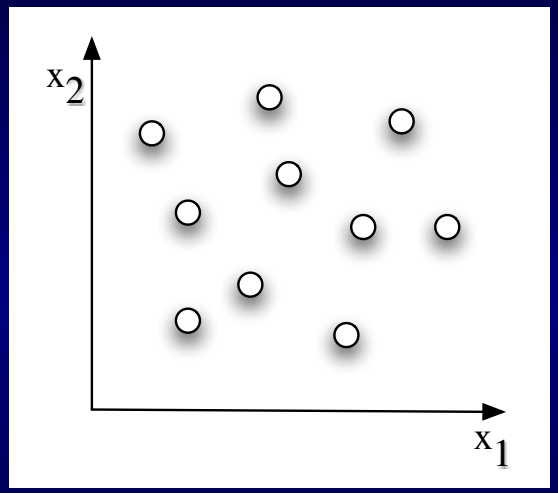


Unsupervised

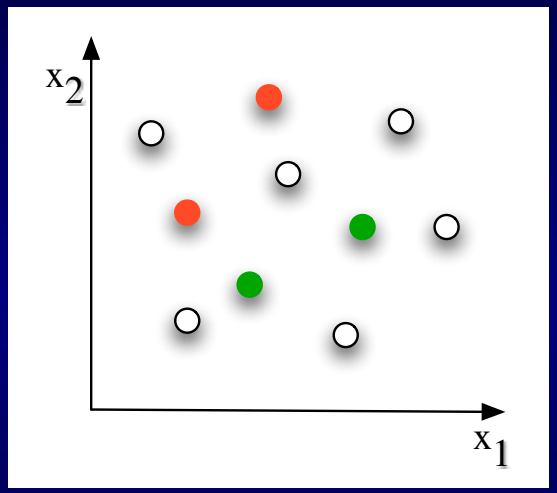


Supervised

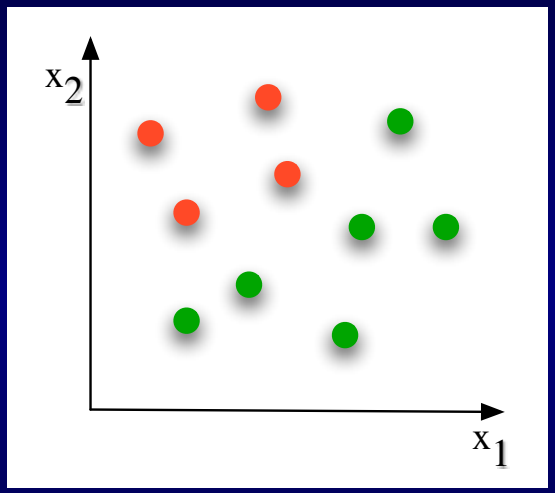
Semi-supervised learning



Unsupervised

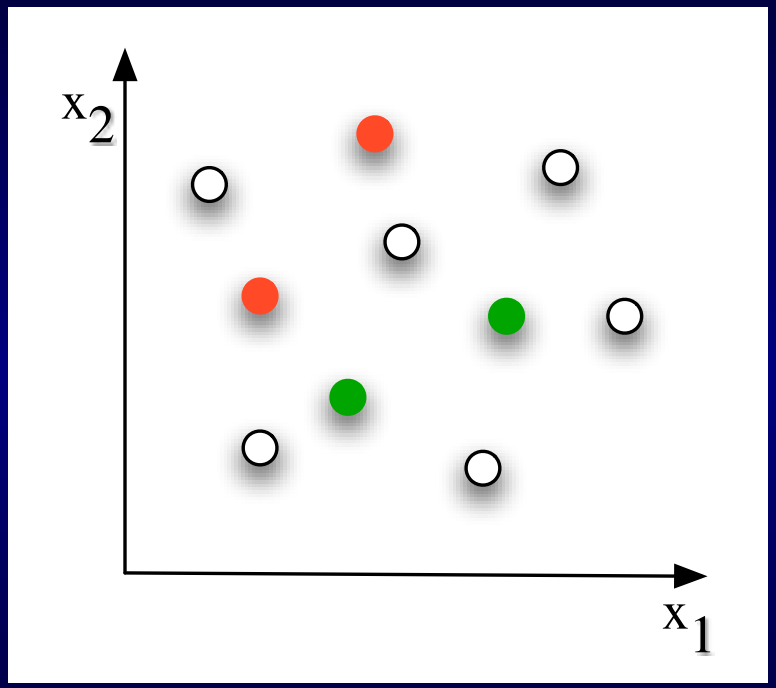


Semi-supervised

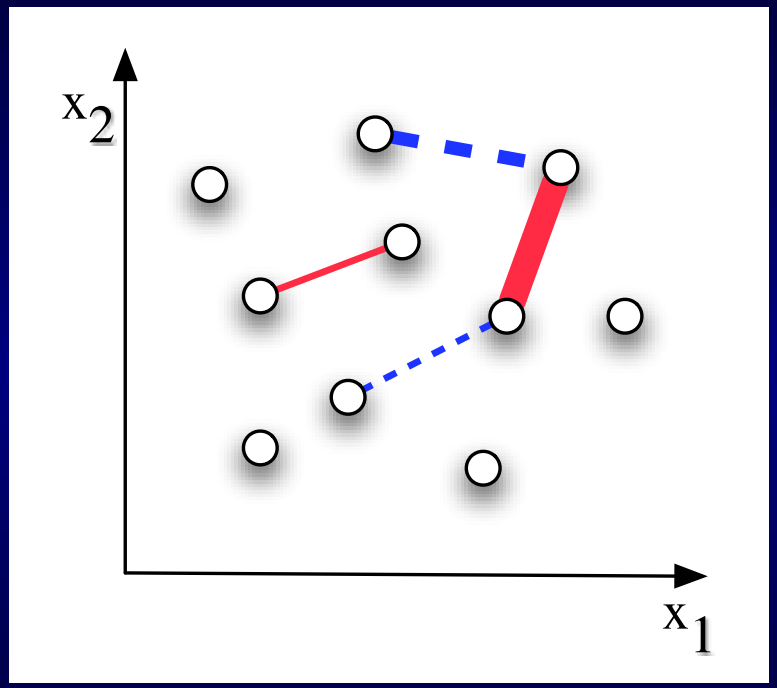


Supervised

Semi-supervised learning

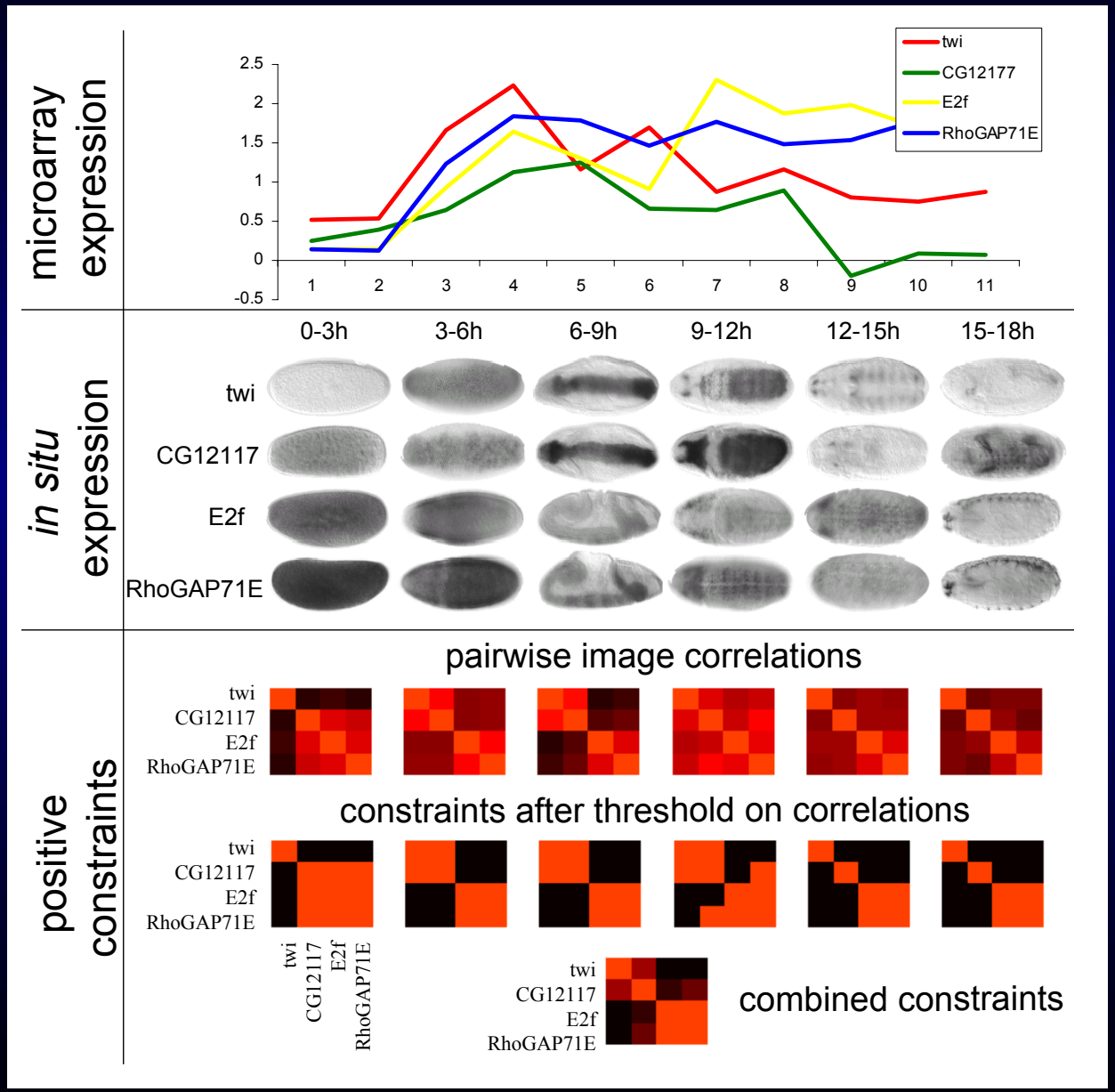


Label data points

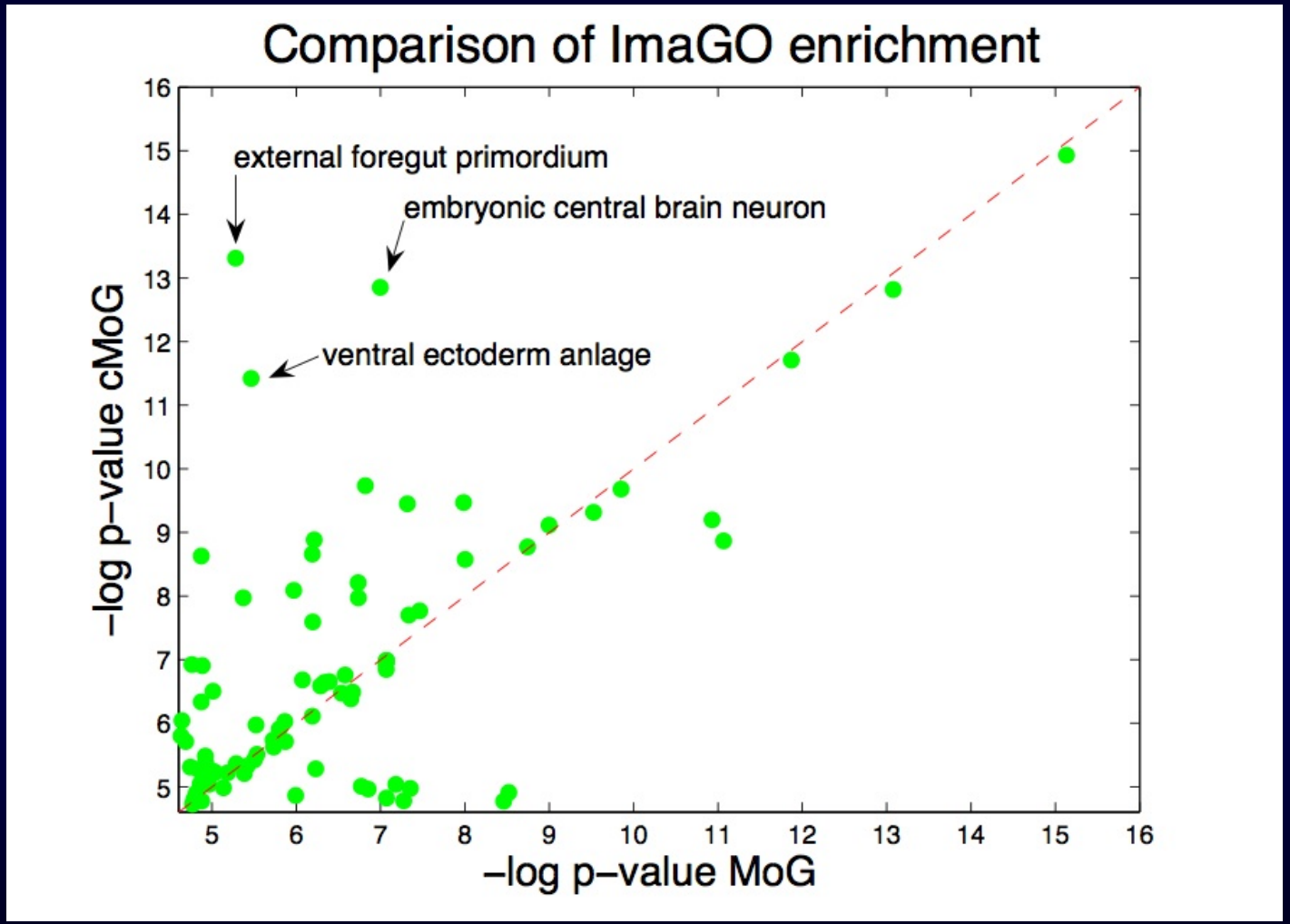


Constrain pairs

Spatio-temporal data



Results: External validation



Discussion

Conclusion

- Mixture models effective framework for detection of functional modules
- Biological meaningful parameterization of multi-variates circumvents dimensionality problem
- Semi-supervised learning of mixtures allows use of heterogeneous data of imbalanced abundance
- Molecular imaging (in situ expression) provides important discrimination

Acknowledgments

- Ivan G. Costa (University of Pernambuco, Brazil), Benjamin Georgi (UPenn), Ruben Schilling, Christoph Hafemeister (MPI Molecular Genetics)
- Stefan Röpcke (Altana Pharma)
- Fritz Melchers (MPI Infection Biology)
- Christoph Dieterich (Max Delbrück Zentrum)

Thanks.

http://schliep.org