Transcriptional modules in *A. thaliana* stress responses: an application of Bayesian data fusion

David L. Wild

Systems Biology Centre, University of Warwick

April 15, 2009

Outline

1. Background and Motivation
2. Bayesian Hierarchical Clustering
3. Integrating Transcription Factor Binding Data
4. Results

Outline

1. Background and Motivation
2. Bayesian Hierarchical Clustering
3. Integrating Transcription Factor Binding Data
4. Results

Arabidopsis thaliana

Scientific Motivation

- Can we find sensible sub-sets of genes from which to infer regulatory networks?
- Does co-expression equal co-regulation?
- Can we identify “transcriptional modules" (sets of gene regulated by a common set of transcription factors)?

Scientific Motivation

- Can we find sensible sub-sets of genes from which to infer regulatory networks?
- Does co-expression equal co-regulation?
- Can we identify “transcriptional modules" (sets of gene regulated by a common set of transcription factors)?

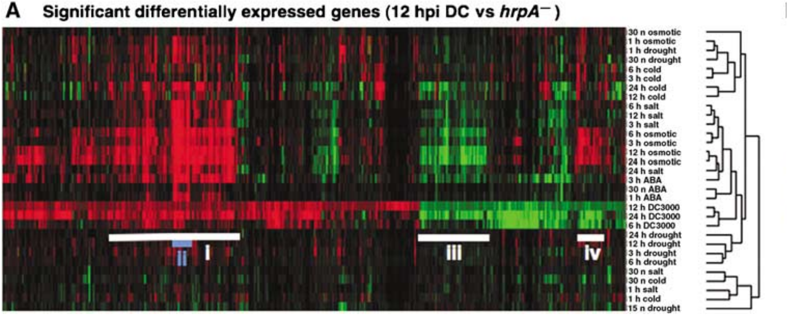
Scientific Motivation

- Can we find sensible sub-sets of genes from which to infer regulatory networks?
- Does co-expression equal co-regulation?
- Can we identify “transcriptional modules" (sets of gene regulated by a common set of transcription factors)?

Scientific Motivation

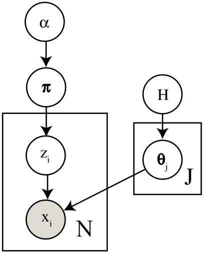
- Can we find sensible sub-sets of genes from which to infer regulatory networks?
- Does co-expression equal co-regulation?
- Can we identify “transcriptional modules" (sets of gene regulated by a common set of transcription factors)?

Agglomerative Hierarchical Clustering



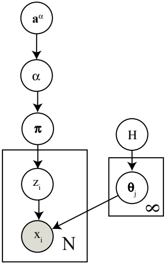
Torres-Zabala et al. EMBO Journal (2007) 26, 1434–1443

Finite Mixture Model



$$
\begin{gathered}
\pi|\alpha, J \sim \mathit{Dirichlet}(\cdot|\alpha/J) \\
\theta_j|H \sim H(\cdot) \\
z_i|\pi \sim \mathit{Multinomial}(\cdot|\pi) \\
x_i|\mathbf{z}_i = j, \theta \sim F(\cdot|\theta_j)
\end{gathered}
$$

Infinite Mixture Model



$$\pi|\alpha \sim \mathit{Stick}(\alpha)$$
$$\theta_j|H \sim H(\cdot)$$
$$z_i|\pi \sim \mathit{Multinomial}(\cdot|\pi)$$
$$x_i|\mathbf{z}_i = j, \theta \sim F(\cdot|\theta_j)$$

Rasmussen, *Advances in Neural Information Processing Systems 12*. 2000:554–560.

Infinite Mixture Model as Dirichlet Process Mixture

The limit of an infinite number of components corresponds to a Dirichlet Process Prior

$$\mathrm{G}(\phi) = \sum_{j=1}^{\infty} \pi_j \delta(\phi - \theta_j)$$

$$\mathrm{G} \mid \alpha, H \sim DP(\alpha, H)$$

Used as clustering model for gene expression profiles:

Medvedovic and Sivaganesan, *Bioinformatics*, vol. 18, 1194–1206, 2002.

Wild et al. *3rd International Conference on Systems Biology, Stockholm, Sweden*, 2002.

Rasmussen et al. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2007. [http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70269].

Outline

1. Background and Motivation
2. Bayesian Hierarchical Clustering
3. Integrating Transcription Factor Binding Data
4. Results

Bayesian Hierarchical Clustering (Heller and Ghahramani 2005)

- A Bayesian way to do hierarchical clustering where marginal likelihoods are used to decide which merges are advantageous
- A novel fast bottom-up way of doing approximate inference in a Dirichlet Process mixture model (e.g. an infinite Gaussians mixture model)
- BHC is virtually identical to traditional hierarchical clustering except that instead of distance it uses marginal likelihoods to decide on merges.
- R/Bioconductor implementation forthcoming (Savage et al., submitted)

Bayesian Hierarchical Clustering (Heller and Ghahramani 2005)

- A Bayesian way to do hierarchical clustering where marginal likelihoods are used to decide which merges are advantageous
- A novel fast bottom-up way of doing approximate inference in a Dirichlet Process mixture model (e.g. an infinite Gaussians mixture model)
- BHC is virtually identical to traditional hierarchical clustering except that instead of distance it uses marginal likelihoods to decide on merges.
- R/Bioconductor implementation forthcoming (Savage et al., submitted)

Bayesian Hierarchical Clustering (Heller and Ghahramani 2005)

- A Bayesian way to do hierarchical clustering where marginal likelihoods are used to decide which merges are advantageous
- A novel fast bottom-up way of doing approximate inference in a Dirichlet Process mixture model (e.g. an infinite Gaussians mixture model)
- BHC is virtually identical to traditional hierarchical clustering except that instead of distance it uses marginal likelihoods to decide on merges.
- R/Bioconductor implementation forthcoming (Savage et al., submitted)

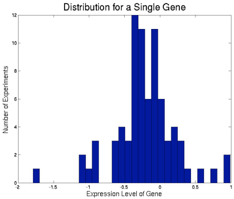
Bayesian Hierarchical Clustering (Heller and Ghahramani 2005)

- A Bayesian way to do hierarchical clustering where marginal likelihoods are used to decide which merges are advantageous
- A novel fast bottom-up way of doing approximate inference in a Dirichlet Process mixture model (e.g. an infinite Gaussians mixture model)
- BHC is virtually identical to traditional hierarchical clustering except that instead of distance it uses marginal likelihoods to decide on merges.
- R/Bioconductor implementation forthcoming (Savage et al., submitted)

Gene Expression Data

$$\begin{aligned} x_j^{(i)} &\sim \mathit{Multinomial}(\cdot|\theta_j) \\ \theta_j &\sim \mathit{Dirichlet}(\cdot|\alpha_j) \end{aligned} \tag{-2}$$

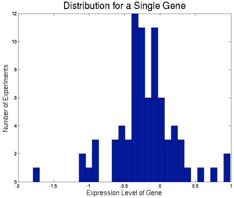
1. Underexpressed - negative tail
2. Unchanged
3. Overexpressed - positive tail



Gene Expression Data

$$\begin{aligned} x_j^{(i)} &\sim \mathit{Multinomial}(\cdot|\theta_j) \\ \theta_j &\sim \mathit{Dirichlet}(\cdot|\alpha_j) \end{aligned} \tag{-2}$$

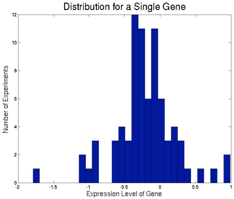
1. Underexpressed - negative tail
2. Unchanged
3. Overexpressed - positive tail



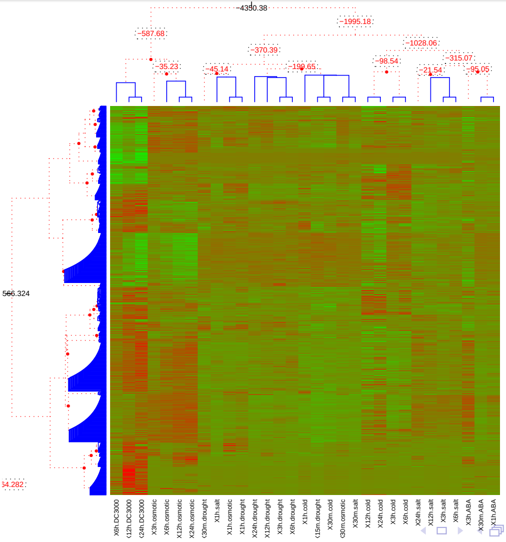
Gene Expression Data

$$\begin{aligned} x_j^{(i)} &\sim \mathit{Multinomial}(\cdot|\theta_j) \\ \theta_j &\sim \mathit{Dirichlet}(\cdot|\alpha_j) \end{aligned} \tag{-2}$$

1. Underexpressed - negative tail
2. Unchanged
3. Overexpressed - positive tail



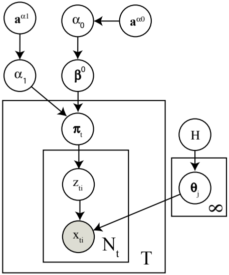
BHC clustering of *A. thaliana* expression data



Outline

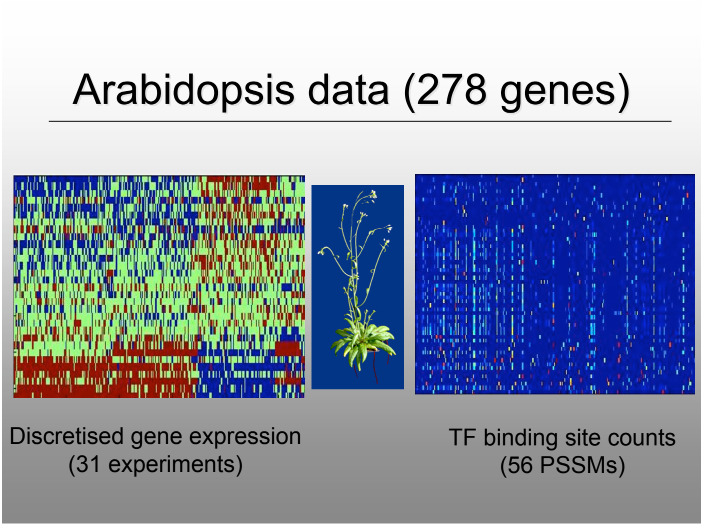
1. Background and Motivation
2. Bayesian Hierarchical Clustering
3. Integrating Transcription Factor Binding Data
4. Results

Hierarchical Dirichlet Process



$$G_0 | \alpha_0, H \sim DP(\alpha_0, H)$$
$$G_j | \alpha_0, G_0 \sim DP(\alpha_0, G_0)$$

Teh et al. *JASA*, 2006.



Arabidopsis data (278 genes)

Discretised gene expression
(31 experiments)

TF binding site counts
(56 PSSMs)

Gene Expression Data Model

Marginal likelihood for a single mixture component:

$$P(D|M) = \prod_i \frac{\Gamma(B_i)}{\Gamma(N_i + B_i)} \prod_k \frac{\Gamma(n_{ik} + \beta_{ik})}{\Gamma(\beta_{ik})}$$

$$B_i = \sum_k \beta_{ik}, N_i = \sum_k n_{ik}$$

i indexes over features (experiments)
k indexes over discrete data categories
β_{ik} are the Dirichlet prior hyperparameters
(naïve Bayes data model with optional hyperparameter fitting)

Transcription Factor Binding Site Data Model

Marginal likelihood for a single mixture component:

$$P(D|M) = \frac{\Gamma(B)}{\Gamma(N+B)} \prod_i \frac{\Gamma(n_i + \beta_i)}{\Gamma(\beta_i)}$$

$$B = \sum_i \beta_i, N = \sum_i n_i$$

i indexes over features (TF binding motifs)
β_i are the Dirichlet prior hyperparameters
(*'bag of words'* data model (Teh et al.) with optional hyperparameter fitting)

HDP-like prior

Let x_{ji} be the observed response for i-th gene in the j-th context. We introduce an extra latent variable r_i for each gene with

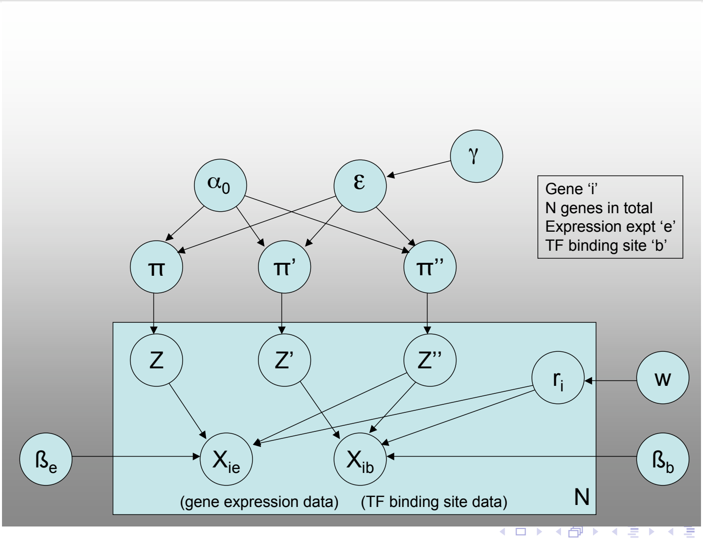
$$p(r_i = 1) = w, \qquad p(r_i = 0) = 1 - w.$$

If $r_i = 1$ then $\theta_i = (\theta_{1i}, \theta_{2i}) \sim G_3$ (fused)
If $r_i = 0$ then $\theta_{1i} \sim G_1$ and $\theta_{2i} \sim G_2$ are conditionally independent (unfused)
This defines 3 contexts. Unlike the HDP, we have

$$G_1 \sim \mathrm{DP}(\alpha_0, G_0^{(1)}), G_2 \sim \mathrm{DP}(\alpha_0, G_0^{(2)}), G_3 \sim \mathrm{DP}(\alpha_0, G_0)$$

where $G_0^{(j)}$ represents the marginal distribution of ϕ_j under G_0.

Graphical Model



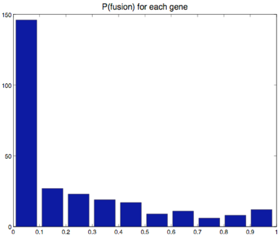
Gene ‘i’
N genes in total
Expression expt ‘e’
TF binding site ‘b’

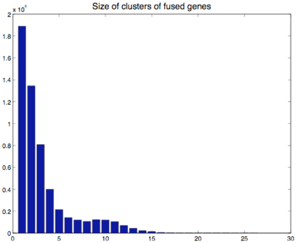
(gene expression data) (TF binding site data)

Outline

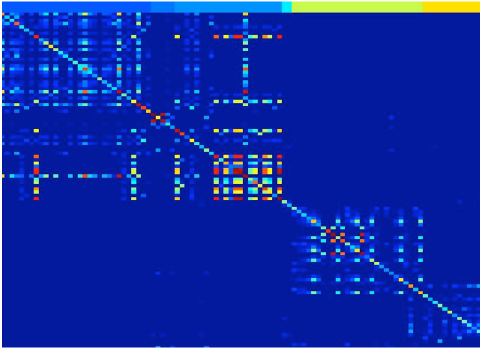
1. Background and Motivation
2. Bayesian Hierarchical Clustering
3. Integrating Transcription Factor Binding Data
4. Results

Results (I) - Fused Genes



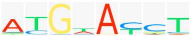


Results (II) Gene grouping Matrix

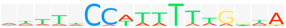


Results (III) - Genes which ‘move’

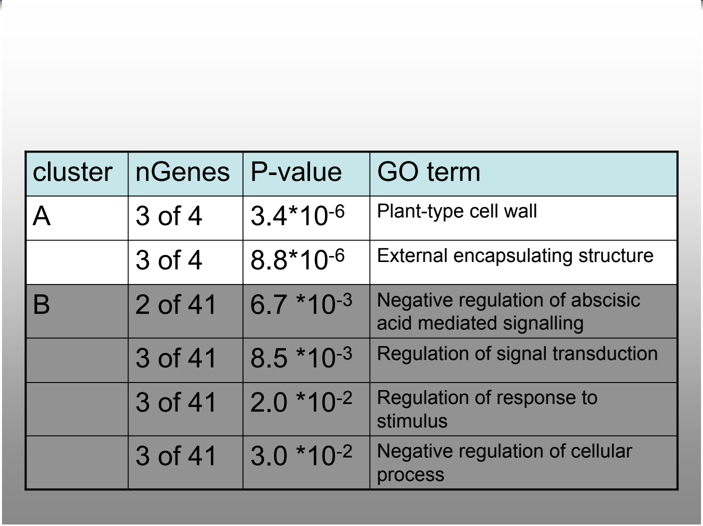
At1g53580: hydroxyacylglutathione hydrolase
tobacco EIN3-like motif -
2 other genes (of 13) have this motif



At5g53870: plastocyanin-like domain
(copper ion binding)
Arabidopsis thaliana AG motif -
4 other genes (of 13) have this motif



Results (IV) Gene Ontology Term Enrichment



cluster	nGenes	P-value	GO term
A	3 of 4	$3.4*10^{-6}$	Plant-type cell wall
	3 of 4	$8.8*10^{-6}$	External encapsulating structure
B	2 of 41	$6.7 *10^{-3}$	Negative regulation of abscisic acid mediated signalling
	3 of 41	$8.5 *10^{-3}$	Regulation of signal transduction
	3 of 41	$2.0 *10^{-2}$	Regulation of response to stimulus
	3 of 41	$3.0 *10^{-2}$	Negative regulation of cellular process

Conclusions

- HDP-based models useful for biological data integration
- New biological insights ?
- Extensions to graphical model formalism needed for 'conditional' graphical models?

Conclusions

- HDP-based models useful for biological data integration
- New biological insights ?
- Extensions to graphical model formalism needed for ‘conditional’ graphical models?

Conclusions

- HDP-based models useful for biological data integration
- New biological insights ?
- Extensions to graphical model formalism needed for ‘conditional’ graphical models?

Conclusions

- HDP-based models useful for biological data integration
- New biological insights ?
- Extensions to graphical model formalism needed for ‘conditional’ graphical models?

Acknowledgments

- Richard Savage
- Jim Griffin (Kent) and Zoubin Ghahramani (Cambridge)
- Sascha Ott and Richard Hickman (Warwick)
- Murray Grant and Bill Truman (Exeter)
- Funded by EPSRC EP/F027400/1 (Life Science Interface)

Acknowledgments

- Richard Savage
- Jim Griffin (Kent) and Zoubin Ghahramani (Cambridge)
- Sascha Ott and Richard Hickman (Warwick)
- Murray Grant and Bill Truman (Exeter)
- Funded by EPSRC EP/F027400/1 (Life Science Interface)

Acknowledgments

- Richard Savage
- Jim Griffin (Kent) and Zoubin Ghahramani (Cambridge)
- Sascha Ott and Richard Hickman (Warwick)
- Murray Grant and Bill Truman (Exeter)
- Funded by EPSRC EP/F027400/1 (Life Science Interface)

Acknowledgments

- Richard Savage
- Jim Griffin (Kent) and Zoubin Ghahramani (Cambridge)
- Sascha Ott and Richard Hickman (Warwick)
- Murray Grant and Bill Truman (Exeter)
- Funded by EPSRC EP/F027400/1 (Life Science Interface)

Acknowledgments

- Richard Savage
- Jim Griffin (Kent) and Zoubin Ghahramani (Cambridge)
- Sascha Ott and Richard Hickman (Warwick)
- Murray Grant and Bill Truman (Exeter)
- Funded by EPSRC EP/F027400/1 (Life Science Interface)

Acknowledgments

- Richard Savage
- Jim Griffin (Kent) and Zoubin Ghahramani (Cambridge)
- Sascha Ott and Richard Hickman (Warwick)
- Murray Grant and Bill Truman (Exeter)
- Funded by EPSRC EP/F027400/1 (Life Science Interface)