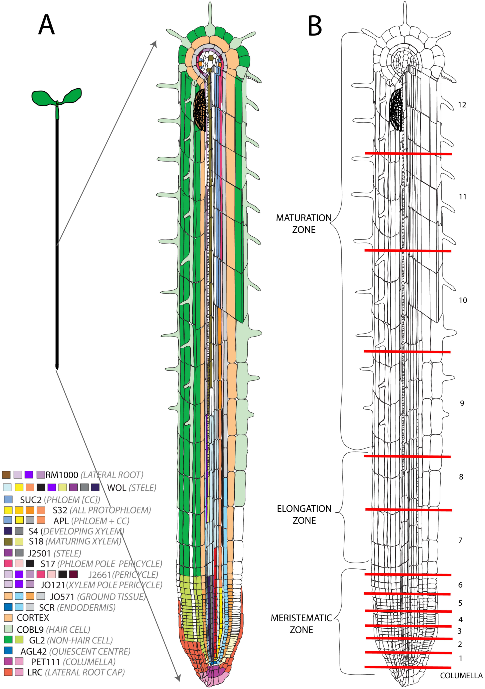
Reconstruction Spatiotemporal Gene Expression from Partial Observations

Dustin Cartwright [1]

April 7, 2010

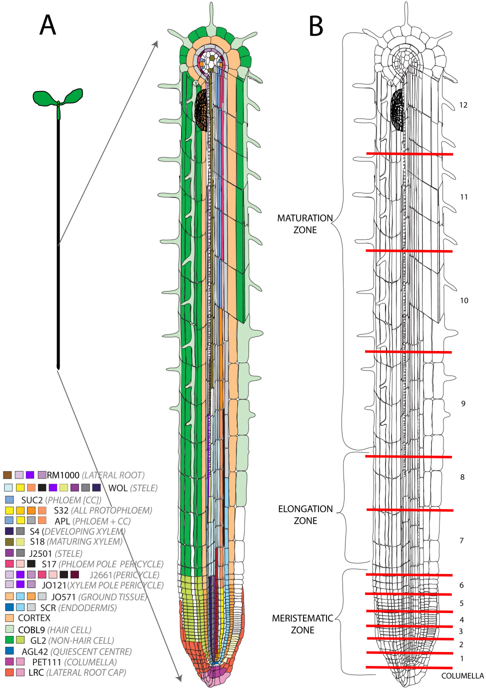
[1] Joint with David Orlando, Siobhan Brady, Bernd Sturmfels, and Philip Benfey. Research supported by the DARPA project Fundamental Laws of Biology

Arabidopsis root



Arabidopsis root

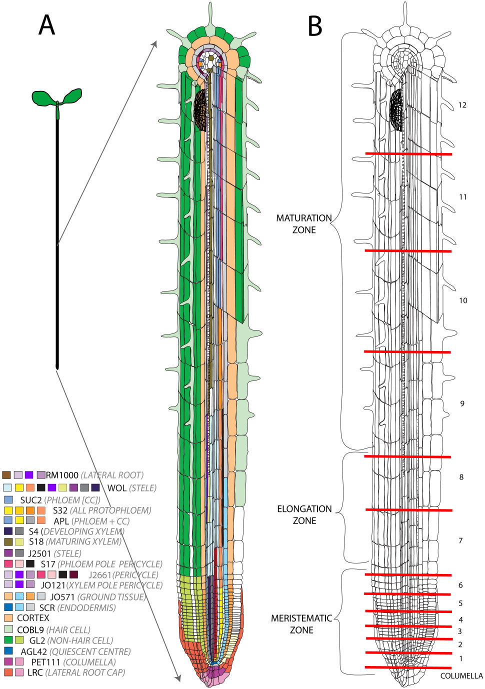
Gene expression microarrays are a tool to understand dynamics and regulatory processes.



Arabidopsis root

Gene expression microarrays are a tool to understand dynamics and regulatory processes. Two ways of separating cells in the lab:

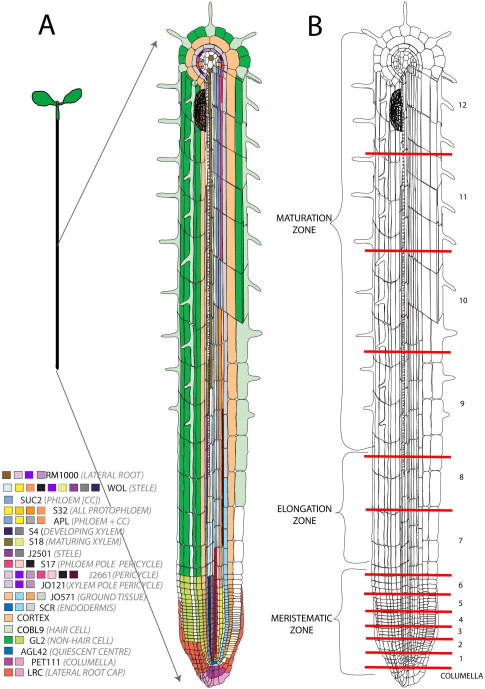
- Chemically, using 18 markers (colors in diagram A)



Arabidopsis root

Gene expression microarrays are a tool to understand dynamics and regulatory processes. Two ways of separating cells in the lab:

- Chemically, using 18 markers (colors in diagram A)
- Physically, using 13 longitudinal sections (red lines in diagram B)



Measurement along two axes

- Markers measure variation among cell types.

Measurement along two axes

- Markers measure variation among cell types.
- Longitudinal sections measure variation along developmental stage.

Measurement along two axes

- Markers measure variation among cell types.
- Longitudinal sections measure variation along developmental stage.

Naïve approach would use variation among each set of experiments as proxies for variation along each of the two axes.

Problem with naïve approach

Correspondence between markers and cell types is imperfect.

Problem with naïve approach

Correspondence between markers and cell types is imperfect.
For example, the sample labelled APL consists of mixture of two cell types:

	cell type	
section	phloem	phloem companion cells
12	$\frac{1}{16}$	$\frac{1}{16}$
$\vdots$	$\vdots$	$\vdots$
7	$\frac{1}{16}$	$\frac{1}{16}$
6	$\frac{1}{16}$	0
$\vdots$	$\vdots$	$\vdots$
3	$\frac{1}{16}$	0
2	0	0
1	0	0
columella	0	0

Problem with naïve approach

Similarly, the longitudinal sections do not have the same mixture of cells. For example:

- In each of sections 1-5, 30-50% of the cells are lateral root cap cells.

Problem with naïve approach

Similarly, the longitudinal sections do not have the same mixture of cells. For example:

- In each of sections 1-5, 30-50% of the cells are lateral root cap cells.
- In sections 6-12, there are no lateral root cap cells.

Problem with naïve approach

Similarly, the longitudinal sections do not have the same mixture of cells. For example:

- In each of sections 1-5, 30-50% of the cells are lateral root cap cells.
- In sections 6-12, there are no lateral root cap cells.

Conclusion: Need to analyze each transcript across all 31 $(= 13 + 18)$ experiments to model the expression pattern in the whole root.

Model

- Expression level for each combination of a cell type and a section.

Model

- Expression level for each combination of a cell type and a section.
- Each marker and longitudinal section measures a linear combination of these expression levels.
- The coefficients of these linear combinations are determined by:
  - Numbers of cells present in each section
  - Marker selection patterns

Model

- Expression level for each combination of a cell type and a section.
- Each marker and longitudinal section measures a linear combination of these expression levels.
- The coefficients of these linear combinations are determined by:
  - Numbers of cells present in each section
  - Marker selection patterns

Under-constrained system: 31 $(= 13 + 18)$ measurements and 129 expression levels.

Assumption

Since the system is under-constrained, we make the following assumption:

Assumption

Since the system is under-constrained, we make the following assumption:

- The dependence on the expression level on the section is independent of the dependence on the cell type.

Assumption

Since the system is under-constrained, we make the following assumption:

- The dependence on the expression level on the section is independent of the dependence on the cell type.
- More precisely, the expression level in section i and type j is $x_i y_j$ for some vectors x and y.

Assumption

Since the system is under-constrained, we make the following assumption:

- The dependence on the expression level on the section is independent of the dependence on the cell type.
- More precisely, the expression level in section i and type j is $x_i y_j$ for some vectors x and y.

Example

If the expression level is either 0 or 1 (off or on), then our assumption says that it is 1 for the combination of some subset of the sections and some subset of the cell types.

Non-negative bilinear equations

Equating the expression levels from the above model with actual observations gives a system of bilinear equations:

Non-negative bilinear equations

Equating the expression levels from the above model with actual observations gives a system of bilinear equations:

$$
\begin{aligned}
x^t A^{(1)} y &= o_1 \\
&\vdots \\
x^t A^{(k)} y &= o_k \\
x_1 + \cdots + x_n &= 1 \quad \text{(normalization)}
\end{aligned}
$$

where

$A^{(1)}, \ldots, A^{(k)}$	$n \times m$ non-negative matrices (cell mixture)
$o_1, \ldots, o_k$	positive scalars (expression levels)

Non-negative bilinear equations

Equating the expression levels from the above model with actual observations gives a system of bilinear equations:

$$\begin{aligned} x^t A^{(1)} y &= o_1 \\ &\vdots \\ x^t A^{(k)} y &= o_k \\ x_1 + \cdots + x_n &= 1 \quad \text{(normalization)} \end{aligned}$$

where

$A^{(1)}, \ldots, A^{(k)}$	$n \times m$ non-negative matrices (cell mixture)
$o_1, \ldots, o_k$	positive scalars (expression levels)

We want approximate solutions with x and y non-negative vectors of dimensions $n \times 1$ and $m \times 1$ respectively.

Kullback-Leibler divergence

Maximum likelihood estimation: Given a model (function $f : \Theta \to \mathbb{R}^k$) and empirical counts for each of the k events, determine the parameters which maximize the probability of the counts given the model.

Kullback-Leibler divergence

Maximum likelihood estimation: Given a model (function $f : \Theta \to \mathbb{R}^k$) and empirical counts for each of the k events, determine the parameters which maximize the probability of the counts given the model.
Equivalently, maximum likelihood parameters minimize the Kullback-Leibler divergence between the predicted distribution and the empirical distribution ($=$ normalized counts):

$$D(o \| f(\theta)) := \sum_{\ell=1}^{k} o_\ell \log\left(\frac{o_\ell}{f_\ell(\theta)}\right)$$

Kullback-Leibler divergence

Maximum likelihood estimation: Given a model (function $f : \Theta \to \mathbb{R}^k$) and empirical counts for each of the k events, determine the parameters which maximize the probability of the counts given the model.
Equivalently, maximum likelihood parameters minimize the Kullback-Leibler divergence between the predicted distribution and the empirical distribution ($=$ normalized counts):

$$D(o \| f(\theta)) := \sum_{\ell=1}^{k} o_\ell \log \left(\frac{o_\ell}{f_\ell(\theta)} \right) - o_\ell + f_\ell(\theta)$$

With two additional terms, the generalized Kullback-Leibler divergence provides a measurement of the difference between any two positive vectors.

Finding maximum likelihood parameters

Two statistical methods for finding maximum likelihood parameters:

- Expectation Maximization: reduce solving mixture model (summation) to solving underlying equations.
- Iterative Proportional Fitting: solving log-linear (monomial) equations.

Expectation Maximization

Want to solve:

$$\sum_{i,j} A_{ij}^{(\ell)} x_i y_j = o_\ell \text{ for } \ell = 1, \ldots, k \tag{1}$$

Expectation Maximization

Want to solve:

$$\sum_{i,j} A_{ij}^{(\ell)} x_i y_j = o_\ell \text{ for } \ell = 1, \ldots, k \tag{1}$$

- Start with guesses $\tilde{x}$, $\tilde{y}$

Expectation Maximization

Want to solve:

$$\sum_{i,j} A_{ij}^{(\ell)} x_i y_j = o_\ell \text{ for } \ell = 1, \ldots, k \tag{1}$$

- Start with guesses $\tilde{x}$, $\tilde{y}$
- Estimate contribution of (i,j) term of left side of equation 1 needed to obtain equality:

$$e_{ij\ell} := \frac{A_{ij}^{(\ell)} \tilde{x}_i \tilde{y}_j}{\sum_{i'j'} A_{i'j'}^{(\ell)} \tilde{x}_i \tilde{y}_j} o_\ell$$

Expectation Maximization

Want to solve:

$$\sum_{i,j} A_{ij}^{(\ell)} x_i y_j = o_\ell \text{ for } \ell = 1, \ldots, k \tag{1}$$

- Start with guesses $\tilde{x}$, $\tilde{y}$
- Estimate contribution of (i, j) term of left side of equation 1 needed to obtain equality:

$$e_{ij\ell} := \frac{A_{ij}^{(\ell)} \tilde{x}_i \tilde{y}_j}{\sum_{i'j'} A_{i'j'}^{(\ell)} \tilde{x}_i \tilde{y}_j} o_\ell$$

- Find approximate solution to system:

$$\left(\sum_{\ell} A_{ij}^{(\ell)} \right) x_i y_j \approx \sum_{\ell} e_{ij\ell}$$

Expectation Maximization

Want to solve:

$$\sum_{i,j} A_{ij}^{(\ell)} x_i y_j = o_\ell \text{ for } \ell = 1, \ldots, k \tag{1}$$

- Start with guesses $\tilde{x}$, $\tilde{y}$
- Estimate contribution of (i, j) term of left side of equation 1 needed to obtain equality:

$$e_{ij\ell} := \frac{A_{ij}^{(\ell)} \tilde{x}_i \tilde{y}_j}{\sum_{i'j'} A_{i'j'}^{(\ell)} \tilde{x}_i \tilde{y}_j} o_\ell$$

- Find approximate solution to system:

$$\left(\sum_{\ell} A_{ij}^{(\ell)} \right) x_i y_j \approx \sum_{\ell} e_{ij\ell}$$

- Repeat until convergence

Iterative Proportional Fitting

Want to minimize Kullback-Leibler divergence of:

$$\left(\sum_{\ell} A_{ij}^{(\ell)} \right) x_i y_j \approx \sum_{\ell} e_{ij\ell}$$

Iterative Proportional Fitting

Want to minimize Kullback-Leibler divergence of:

$$\left(\sum_{\ell} A_{ij}^{(\ell)}\right) x_i y_j \approx \sum_{\ell} e_{ij\ell}$$

Simplify:

$$A_{ij} x_i y_j \approx e_{ij} \quad \text{for } 1 \leq i \leq n, 1 \leq j \leq m.$$

Iterative Proportional Fitting

Want to minimize Kullback-Leibler divergence of:

$$\left(\sum_{\ell} A_{ij}^{(\ell)}\right) x_i y_j \approx \sum_{\ell} e_{ij\ell}$$

Simplify:

$$A_{ij} x_i y_j \approx e_{ij} \quad \text{for } 1 \leq i \leq n, 1 \leq j \leq m.$$

Algorithm:

- Adjust $\tilde{x}_i$:

$$\tilde{x}_i \leftarrow \tilde{x}_i \frac{\sum_j e_{ij}}{\sum_j A_{ij} \tilde{x}_i \tilde{y}_j}$$

- Adjust $\tilde{y}_i$:

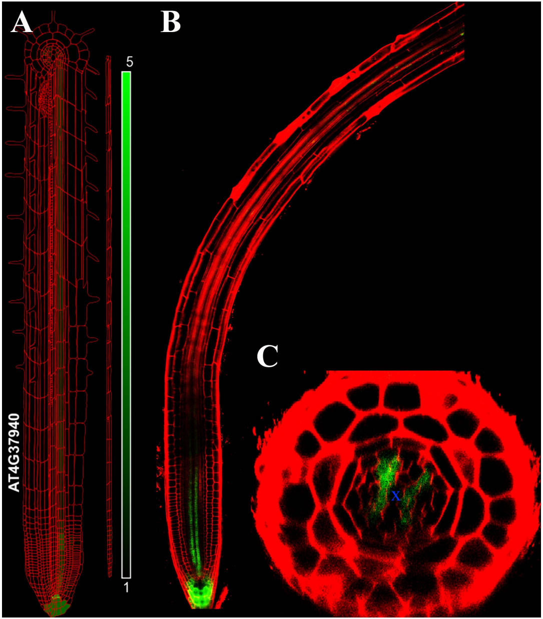
$$\tilde{y}_j \leftarrow \tilde{y}_j \frac{\sum_i e_{ij}}{\sum_i A_{ij} \tilde{x}_i \tilde{y}_j}$$

- Iterate until convergence

Back to Arabidopsis root

Using this algorithm, we estimated the expression profiles of $30,000$ transcripts in several hours.

Validation



A: reconstructed expression levels.
B and C: same transcript visualized using green fluorescent protein (GFP).

Generalization: positive root finding

The EM/IPF-based algorithm can be generalized to find exact or approximate positive solutions to polynomial systems of equations:

$$\sum_{\alpha \in S} a_{\ell\alpha} x^{\alpha} = o_{\ell} \quad \text{for } \ell = 1, \ldots, k,$$

where

- S is a finite set of exponent vectors,
- coefficients $a_{\ell\alpha}$ are all non-negative,
- the o_{ℓ} are positive, and
- a technical condition on the exponents (sufficient to be homogeneous or multi-homogeneous).