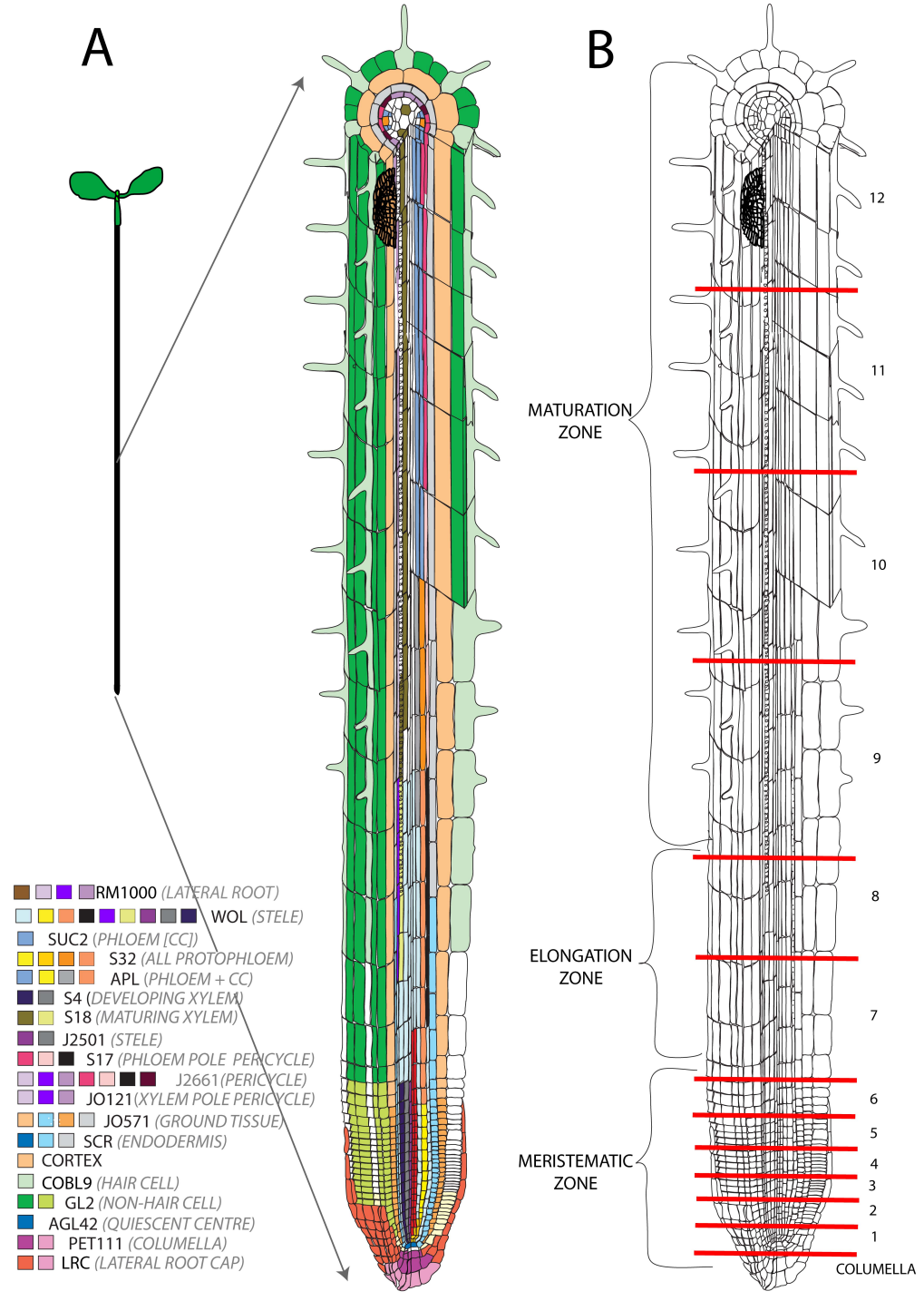
From Arabidopsis roots to bilinear equations

Dustin Cartwright [1]

October 22, 2008

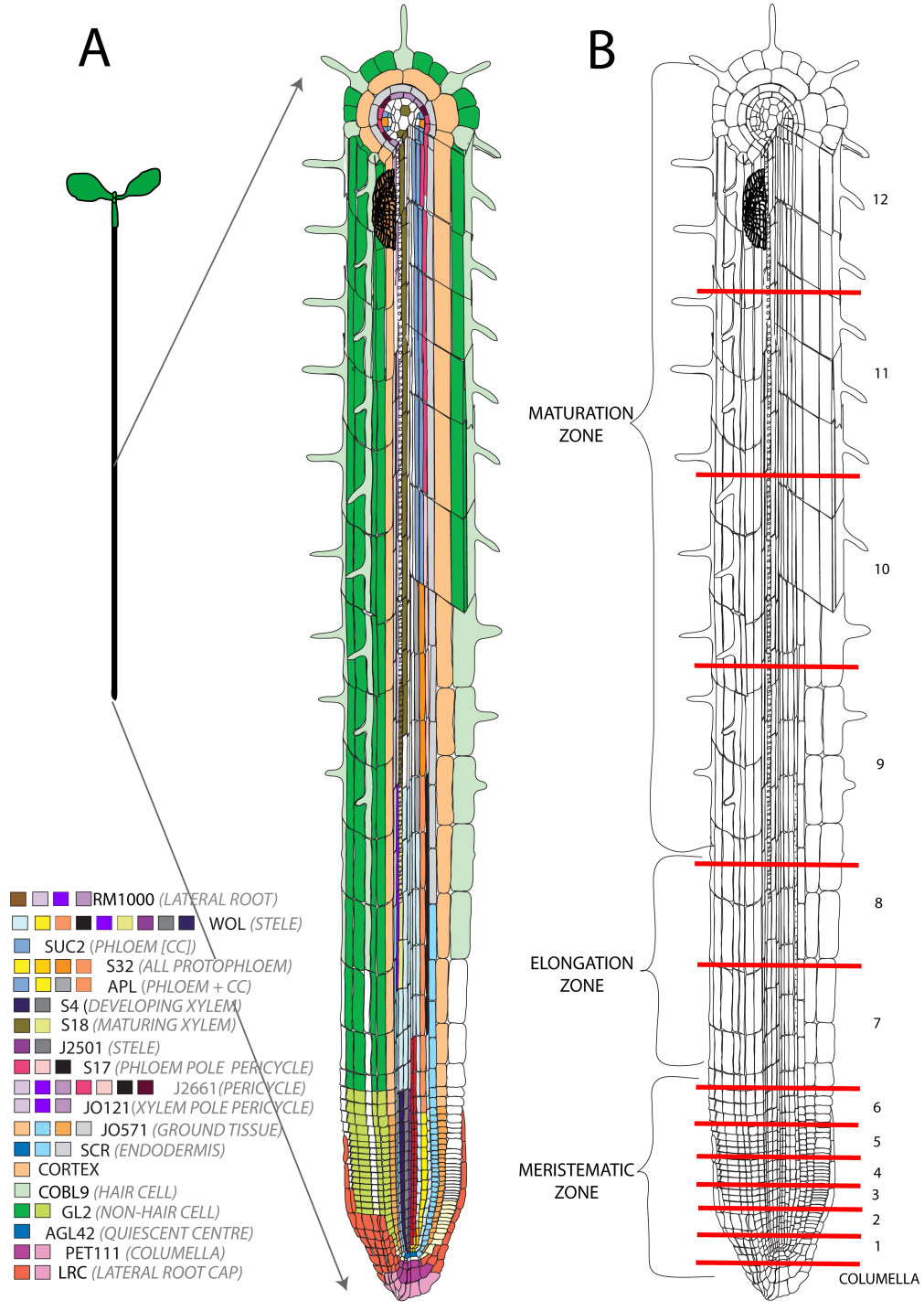
[1] joint with Philip Benfey, Siobhan Brady, David Orlando (Duke University) and Bernd Sturmfels (UC Berkeley), research supported by the DARPA project Fundamental Laws of Biology

Arabidopsis root



Arabidopsis root

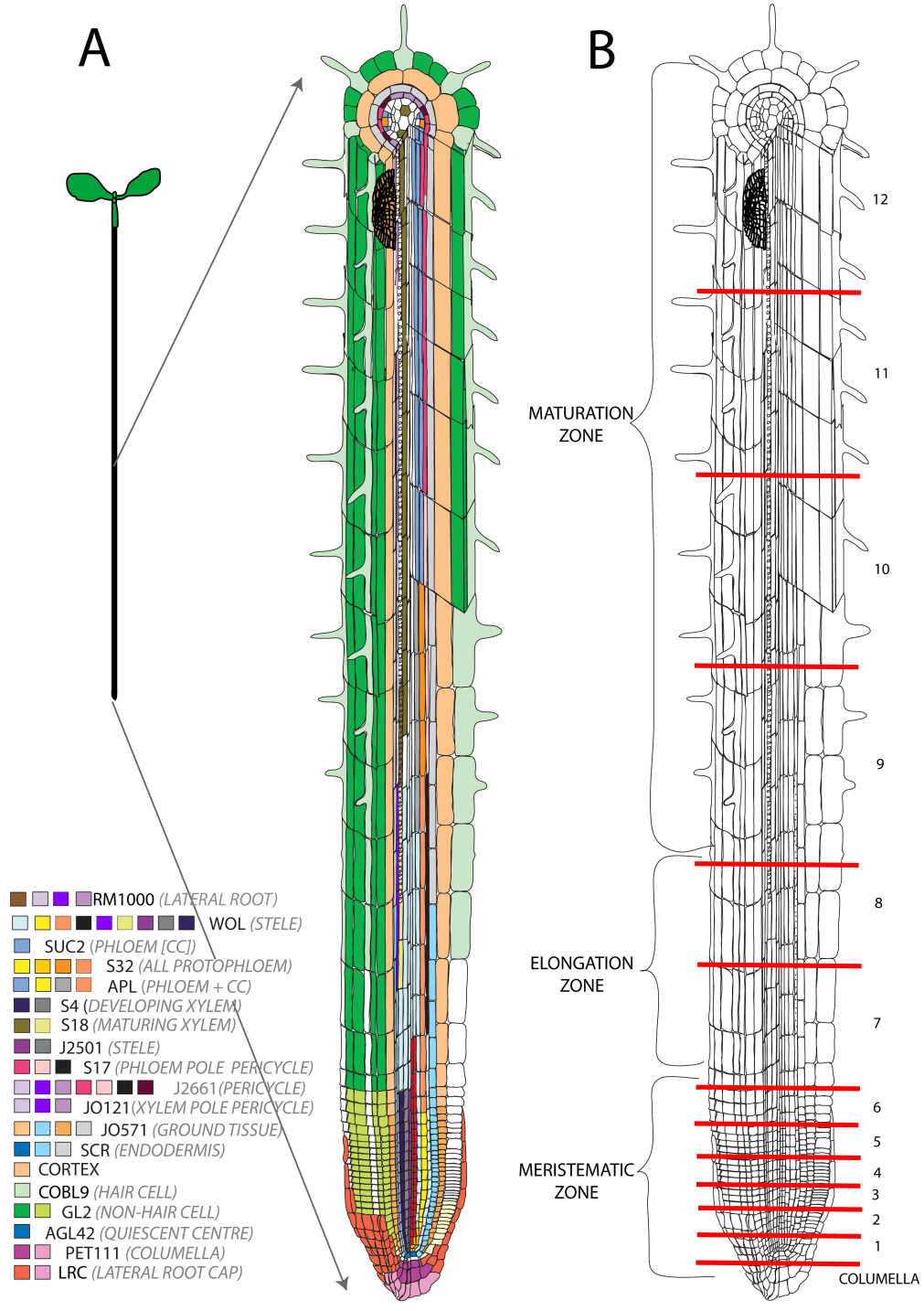
Gene expression microarrays are a tool to understand dynamics and regulatory processes.



Arabidopsis root

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

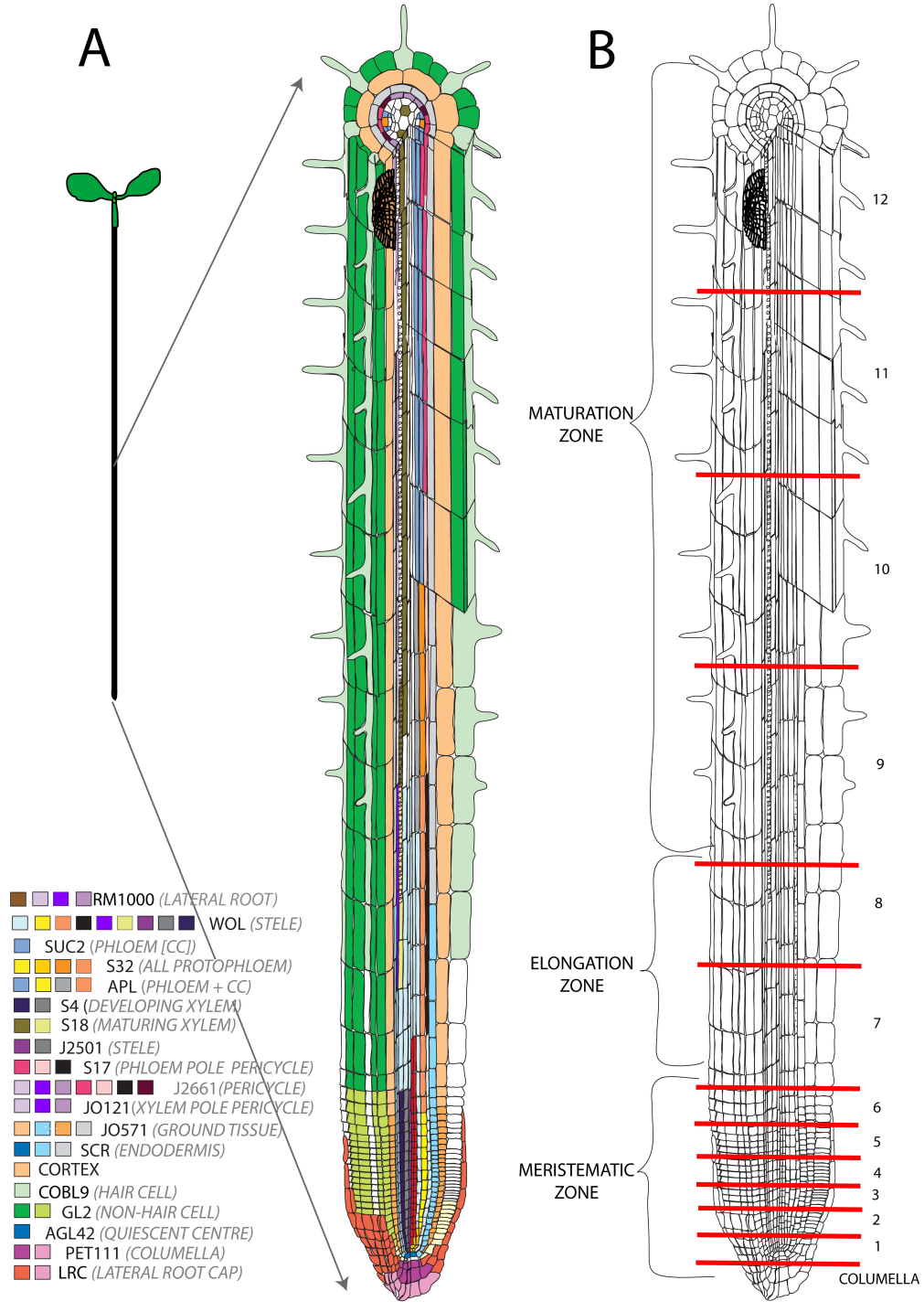
- Chemically, using 18 markers (colors in diagram A)



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	$\frac{1}{16}$	$\frac{1}{16}$
⋮	⋮	⋮
7	$\frac{1}{16}$	$\frac{1}{16}$
6	$\frac{1}{16}$	0
⋮	⋮	⋮
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

- A cluster consists of cells of the same type in the same section. Each cluster has an expression level.

Model

- A cluster consists of cells of the same type in the same section. Each cluster has an expression level.
- For each marker and each longitudinal section, we have a measurement functional, a linear combination of the expression levels in different clusters.

Model

- A cluster consists of cells of the same type in the same section. Each cluster has an expression level.
- For each marker and each longitudinal section, we have a measurement functional, a linear combination of the expression levels in different clusters. The coefficients of these functionals can be determined from:
  - Numbers of cells present in each section
  - Marker selection patterns

Model

- A cluster consists of cells of the same type in the same section. Each cluster has an expression level.
- For each marker and each longitudinal section, we have a measurement functional, a linear combination of the expression levels in different clusters. The coefficients of these functionals can be determined from:
  - Numbers of cells present in each section
  - Marker selection patterns

Under-constrained system: 31 $(= 13 + 18)$ functionals and 129 clusters.

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 of cluster 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 of cluster 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

$A^{(1)}, \ldots, A^{(k)}$ $n \times m$ non-negative matrices (cell mixture)

$o_1, \ldots, o_k$ non-negative scalars (expression levels)

Solve (approximately)

$$f_1(x, y) := x^t A^{(1)} y = o_1$$

$$\vdots$$

$$f_k(x, y) := x^t A^{(k)} y = o_k$$

$$x_1 + \cdots + x_n = 1$$

Non-negative bilinear equations

$A^{(1)}, \ldots, A^{(k)}$ $n \times m$ non-negative matrices (cell mixture)

$o_1, \ldots, o_k$ non-negative scalars (expression levels)

Solve (approximately)

$$f_1(x, y) := x^t A^{(1)} y = o_1$$

$$\vdots$$

$$f_k(x, y) := x^t A^{(k)} y = o_k$$

$$x_1 + \cdots + x_n = 1$$

for x and y non-negative vectors of dimensions $n \times 1$ and $m \times 1$ respectively.

Probabilistic interpretation

$$f_\ell(x, y) := \sum_{i,j} A_{ij}^{(\ell)} x_i y_j \text{ for } \ell = 1, \ldots, k$$

Up to scaling, this vector has the form of the family of probability distributions (depending on vectors x and y)

Probabilistic interpretation

$$f_\ell(x, y) := \sum_{i,j} A_{ij}^{(\ell)} x_i y_j \text{ for } \ell = 1, \ldots, k$$

Up to scaling, this vector has the form of the family of probability distributions (depending on vectors x and y) coming from the following process:

1. Pick a pair of integers from $\{1, \ldots, n\} \times \{1, \ldots, m\}$ with (i, j) having probability proportional to

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

Probabilistic interpretation

$$f_\ell(x,y) := \sum_{i,j} A_{ij}^{(\ell)} x_i y_j \text{ for } \ell = 1,\ldots,k$$

Up to scaling, this vector has the form of the family of probability distributions (depending on vectors x and y) coming from the following process:

1. Pick a pair of integers from $\{1,\ldots,n\} \times \{1,\ldots,m\}$ with (i,j) having probability proportional to
$$\left(\sum_\ell A_{ij}^{(\ell)}\right) x_i y_j$$
2. Output an integer from $\{1,\ldots,k\}$. Conditional on having picked i and j in the previous step, the probability of outputing ℓ is:
$$A_{ij}^{(\ell)} / \left(\sum_\ell A_{ij}^{(\ell)}\right)$$

Maximum Likelihood Estimation

Rescaling both sides of our system of equations:

$$\frac{f_\ell(x,y)}{\sum_{\ell'} f_{\ell'}(x,y)} = \frac{o_\ell}{\sum_{\ell'} o_{\ell'}} \text{ for } \ell = 1,\ldots,k$$

Maximum Likelihood Estimation

Rescaling both sides of our system of equations:

$$\frac{f_\ell(x,y)}{\sum_{\ell'} f_{\ell'}(x,y)} = \frac{o_\ell}{\sum_{\ell'} o_{\ell'}} \text{ for } \ell = 1,\ldots,k$$

Finding an approximate solution to these equations is known as Maximum Likelihood Estimation.

Kullback-Leibler divergence

Kullback-Leibler divergence gives a way of comparing two probability distributions:

$$D(z \| f(x,y)) := \sum_{\ell} z_\ell \log\left(\frac{z_\ell}{f_\ell(x)}\right)$$

Kullback-Leibler divergence

Kullback-Leibler divergence gives a way of comparing two probability distributions:

$$D(z\|f(x,y)) := \sum_{\ell} z_\ell \log\left(\frac{z_\ell}{f_\ell(x)}\right) - z_\ell + f_\ell(x,y)$$

We generalize divergence to any pair of non-negative vectors.

Kullback-Leibler divergence

Kullback-Leibler divergence gives a way of comparing two probability distributions:

$$D(z\|f(x,y)) := \sum_{\ell} z_\ell \log\left(\frac{z_\ell}{f_\ell(x)}\right) - z_\ell + f_\ell(x,y)$$

We generalize divergence to any pair of non-negative vectors.
By approximate solution to a system, we will mean the a solution which minimizes the Kullback-Leibler divergence.

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:

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

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:

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

- Find approximate solution to system:

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

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:

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

- Find approximate solution to system:

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

- Repeat until convergence

Likelihood maximization for monomial models

$$g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{nm}$$
$$(x_i), (y_j) \mapsto A_{ij} x_i y_j$$

where $A_{ij} = \sum_\ell A_{ij}^{(\ell)}$.

Likelihood maximization for monomial models

$$g\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{nm}$$
$$(x_i), (y_j) \mapsto A_{ij} x_i y_j$$

where $A_{ij} = \sum_\ell A_{ij}^{(\ell)}$.
Moment map (taking row sums and column sums):

$$\mu\colon \mathbb{R}^{nm} \to \mathbb{R}^n \times \mathbb{R}^m$$
$$b_{ij} \mapsto \Big(\sum_j b_{ij}\Big), \Big(\sum_i b_{ij}\Big)$$

Likelihood maximization for monomial models

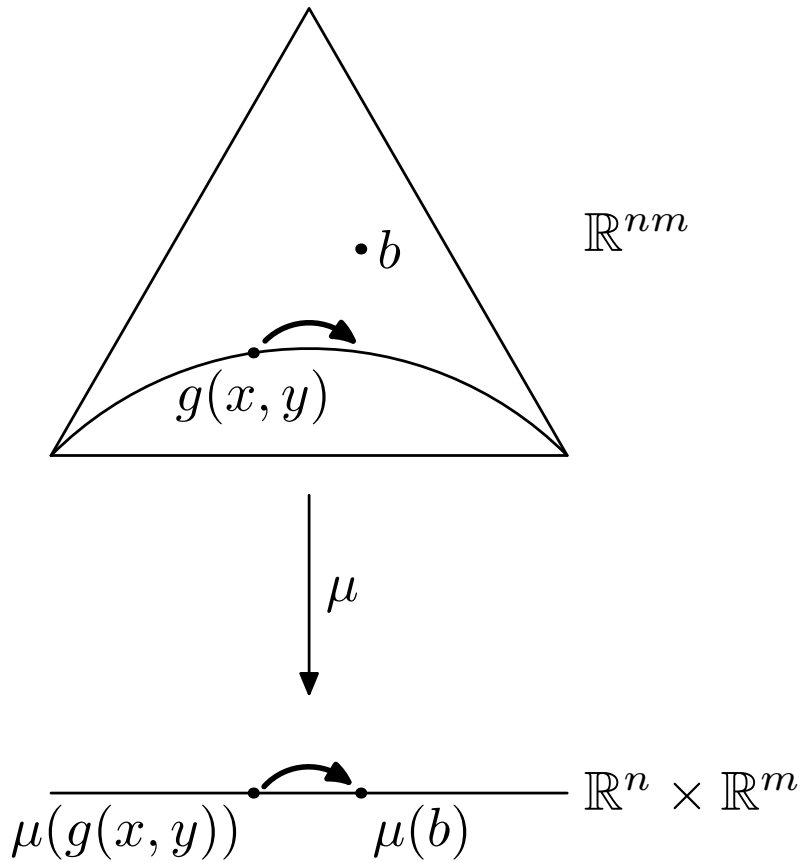
$$g\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{nm}$$
$$(x_i), (y_j) \mapsto A_{ij} x_i y_j$$

where $A_{ij} = \sum_\ell A_{ij}^{(\ell)}$.
Moment map (taking row sums and column sums):

$$\mu\colon \mathbb{R}^{nm} \to \mathbb{R}^n \times \mathbb{R}^m$$
$$b_{ij} \mapsto \Big(\sum_j b_{ij}\Big), \Big(\sum_i b_{ij}\Big)$$

Theorem
Kullback-Leibler divergence $D(z\|g(x,y))$ is minimized over all x and y when $\mu(z)$ equals $\mu(g(x,y))$.

Inverting the moment map: Iterative Proportional Fitting



Inverting the moment map: Iterative Proportional Fitting

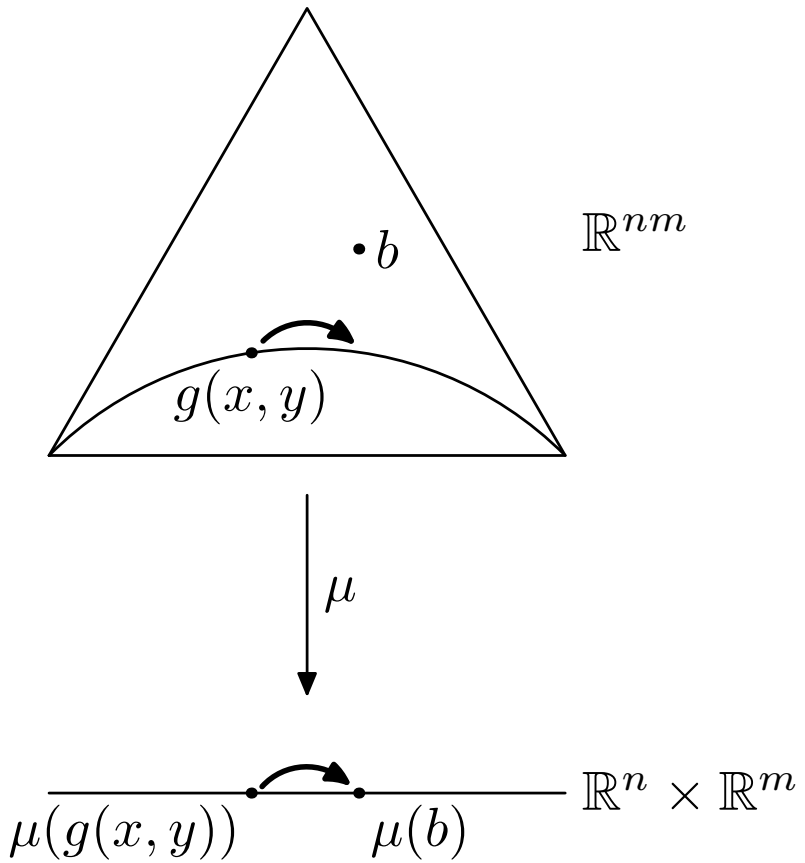
- Adjust $\tilde{x}_i$:

$$\tilde{x}_i \leftarrow \tilde{x}_i \frac{\sum_j b_{ij}}{\sum_j a_{ij} \tilde{x}_i \tilde{y}_j}$$

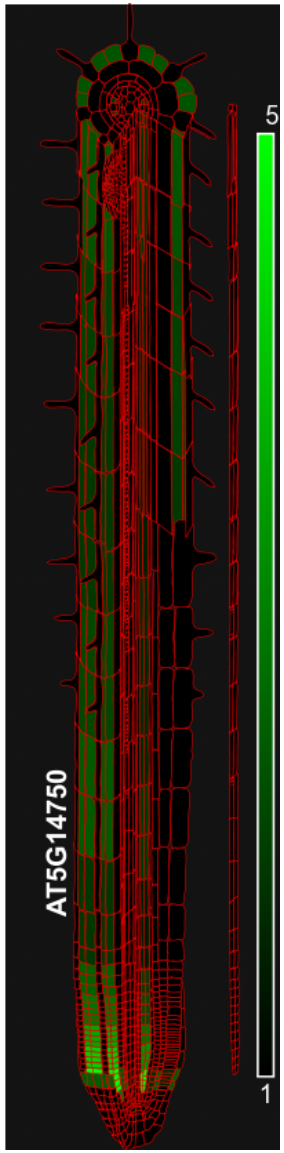
- Adjust $\tilde{y}_i$:

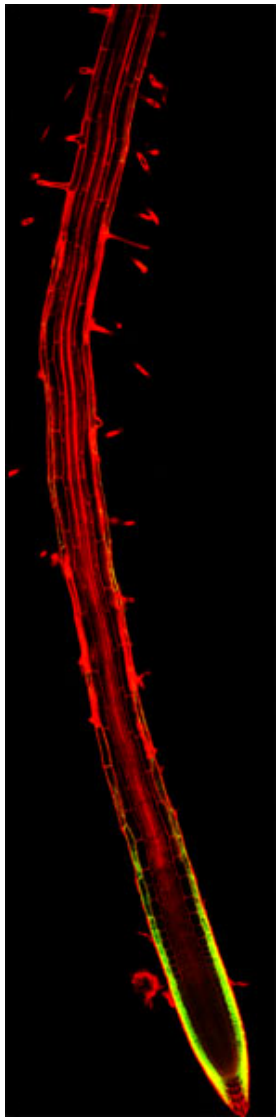
$$\tilde{y}_j \leftarrow \tilde{y}_j \frac{\sum_i b_{ij}}{\sum_i a_{ij} \tilde{x}_i \tilde{y}_j}$$

- Iterate until convergence



Validation: Preliminary results





On the left is a visual representation of the reconstructed expression levels.
On the right, the expression levels for the same transcript are visualized using GFP.