RECONSTRUCTING SPATIOTEMPORAL GENE
EXPRESSION DATA FROM PARTIAL OBSERVATIONS

DUSTIN A. CARTWRIGHT, SIOBHAN M. BRADY, DAVID A. ORLANDO,
BERND STURMFELS, AND PHILIP N. BENFEY

ABSTRACT. Developmental transcriptional networks in plants and animals operate in both space and time. To understand these transcriptional networks it is essential to obtain whole-genome expression data at high spatiotemporal resolution. Substantial amounts of spatial and temporal microarray expression data previously have been obtained for the Arabidopsis root; however, these two dimensions of data have not been integrated thoroughly. Complicating this integration is the fact that these data are heterogeneous and incomplete, with observed expression levels representing complex spatial or temporal mixtures. Given these partial observations, we present a novel method for reconstructing integrated high resolution spatiotemporal data. Our method is based on a new iterative algorithm for finding approximate roots to systems of bilinear equations.

1. INTRODUCTION

Transcriptional regulation plays an important role in orchestrating a host of biological processes, particularly during development (reviewed in [9][13]). Advances in microarray and sequencing technologies have allowed biologists to capture genome-wide gene expression data; the output of this transcriptional regulation. This expression data can then be used to identify genes whose expression is correlated with a particular biological process, and to identify transcriptional regulators that coordinate the expression of groups of genes that are important for the same biological process.

The identification of such genes and transcriptional regulators is complicated by the complex heterogeneous mixture of cell types and developmental stages that comprise each organ of an organism. Expression patterns that are found only in a subset of cell types within an organ will be diluted and may not be detectable in the collection of expression patterns obtained from RNA isolated from samples of an entire organ. Therefore techniques have been developed to enrich samples for specific cell types or developmental stages, especially for studies in plants [5]. In the model plant, Arabidopsis thaliana, several features of the root organ reduce its developmental complexity and facilitate analysis. Specifically, most root cell types are found within concentric cylinders moving from the outside of the root to the inside.
of the root (Figure 1). These cell type layers display rotational symmetry thus simplifying the spatial features of development. This feature has been exploited in the development of a cell type enrichment method. This enrichment method uses green fluorescent protein (GFP)-marked transgenic lines and fluorescently-activated cell sorting (FACS) to collect cell type enriched samples and has allowed for the identification of cell type-specific expression patterns [1, 2]. Using this technique, high resolution expression data have been obtained for nearly all cell types in the Arabidopsis root (herein called the marker-line dataset) [4, 10].

Another feature that makes the Arabidopsis root a tractable developmental model is that cell types are constrained in files along the root’s longitudinal axis and most of these cells are produced from a stem cell population found at the apex of the root. This feature allows a cell’s developmental timeline to be represented by its position along the length of the root. To obtain a developmental time-series expression dataset individual Arabidopsis roots were sectioned into thirteen pieces, each piece representing a developmental time point (herein called the longitudinal dataset) [4]. Each of these sections, however, contains a mixture of cell types, and the microarray expression values obtained are therefore the average of the expression levels over multiple cell types present at these specific developmental time points.

While the 19 fluorescently marked lines in Brady et al. [4] cover expression in nearly all cell types, they do not comprehensively mark all developmental stages of these cell types. Also, the procambium cell type was not measured, as a fluorescent marker-line that marks that cell type did not exist at the time. However, expression from the longitudinal dataset, does contain averaged expression of all cell types, and may be used to infer the missing cell type data.

Previous studies have looked at separating expression data from the heterogeneous cell populations that make up tumors into the contributions of their constituent cell types [8, 16]. However, in that context, the difficulty comes from the fact that the mixture of cell types in each sample is unknown, whereas within our experimental context, the cell type mixture of each sample is known. Two computational methods have been developed to combine the Arabidopsis longitudinal and marker-line datasets as experimentally resolving this expression with marker lines is nearly impossible [4, 6]. However, neither method takes all data into account when reconstructing expression. In [4], only high relative gene expression is considered, and in [6], no attempt is made to infer expression for cells not covered by any marker-line.

In this work we formulate a model for expression levels in Arabidopsis roots in which cell type and developmental stage are independent sources of variation. The microarray data specifying overall expression levels for certain mixtures of cells lead to an overconstrained system of bilinear equations. Moreover, due to the nature of the problem, we are exclusively interested in positive real solutions. We present a new method for finding non-negative real approximate solutions to bilinear equations, based on the techniques
of expectation maximization (EM) [15, Sec. 1.3] and iterative proportional fitting (IPF) [7] from likelihood maximization in statistics. Earlier work has used expectation maximization to find non-negative matrix factorizations [11], and our method is a generalization of that work.

We applied our method to estimate spatiotemporal subregion expression patterns for 20,872 Arabidopsis transcripts. These patterns have identified gene expression in cell types and developmental stages which were previously unknown. A searchable database of verlays of these patterns on a schematic Arabidopsis root is under development and will be made publicly available at http://www.arexdb.org.

2. Methods

2.1. Expression data. Our method uses the normalized expression data collected in [4]. Expression levels were measured across 13 longitudinal sections in a single root (longitudinal dataset) and across 19 different markers (marker-line dataset). For simplicity, the J2501 line was removed from further analysis as it is redundant with the WOODEN-LEG marker-line. The APL marker-line was also removed, as it contains domains of expression marked by both the S32 and the SUC2 marker-lines and adds no extra information. The remaining 17 markers covering 14 cell types are listed in the second column of Table 1.

Due to computational constraints, the original normalization of this data was performed for the longitudinal and the marker-line datasets independently [4]. In order to account for differences caused by these separate normalization procedures, we adjusted the marker-line data by a global factor of 0.92. This factor was calculated by comparing the expression values of ubiquitous, evenly expressed probe sets between the two datasets. We assume that by comparing these probe sets, any true expression differences due to cell type and longitudinal section specificity should be minimal and thus any differences in expression level is a byproduct of the separate normalization procedures. A set of 43 probesets were identified which were expressed ubiquitously (above a normalized value of 1.0 in all samples) and whose expression did not vary significantly among samples within a dataset (ratio of min/max expression within a dataset is at most 0.5). The scaling factor necessary to make the mean expression within the marker-line dataset equal to the mean expression within the longitudinal dataset was calculated for each probe set in this set. The median value of these 43 scaling factors was 0.92, which was used as the global adjustment factor (Table 5).

2.2. Model. To model the transcript expression level of an individual cell we assume that the effects of its cell type and its section on its expression level are independent of each other. More precisely, we assume that the transcript expression level of a cell of type $j$ in section $i$ is equal to the product $x_i \cdot y_j$, where $x_i$ depends only on the section and $y_j$ depends only on the cell type. In other words, for each transcript, there is an idealized profile
of expression over different cell types, and an idealized profile of expression over different sections. Within a given section, our assumption is that the transcript expression level varies proportionally to its cell type profile, and within a given cell type, proportionally to its longitudinal profile.

Each microarray sample in the two datasets (described in Section 2.1), is composed of a distinct mixture of cell types and sections. Within each sample, the measured transcript expression level is a convex linear combination of the expression levels of its constituent cells. Under the above assumptions, these measurements constitute a system of bilinear equations,

$$\sum_{i=1}^{13} \sum_{j=1}^{14} a_{ijk} x_i y_j = b_k \quad \text{for } k = 1, \ldots, 30$$

where $x_i$ and $y_j$ are the model parameters for the 13 sections and 14 cell types respectively, and $b_k$ is the measured expression level as $k$ ranges over the 30 measured samples (13 longitudinal sections and 17 markers).

The coefficients $a_{ijk}$ are obtained by combining the cell type by marker line data (Table 1) with the cell-count matrix (Table 2). For $k$ at most 13, the measurement with index $k$ comes from the $k$th longitudinal section. We will use $a_{***k}$ to denote the corresponding matrix, with $a_{ijk}$ in the $i$th row and $j$th column. We set this matrix to be zero everywhere except the $k$th row, where it is proportional to the $k$th row of the cell-count matrix, but rescaled to sum to 1. For $k$ greater than 13, the measurements come from one of the 17 marker-lines. The matrix $a_{***k}$ is likewise zero except for those

| Cell type                  | Marker-lines                      |
|----------------------------|-----------------------------------|
| Quiescent center           | AGL42, RM1000, SCR5               |
| Columella                  | PET111                            |
| Lateral root cap           | LRC                               |
| Hair cell                  | COBL9 (8-13)                      |
| Non-hair cell              | GL2                               |
| Cortex                     | J0571, CORTEX (7-13)              |
| Endodermis                 | J0571, SCR5                       |
| Xylem pole pericycle       | WOL (2-9), JO121 (9-13), J2661 (13)|
| Phloem pole pericycle      | WOL (2-9), S17 (8-13), J2661 (13) |
| Phloem                     | S32, WOL (2-9)                    |
| Phloem companion cells     | SUC2 (10-19), WOL (2-9)           |
| Xylem                      | S4 (2-7), S18 (8-13), WOL (2-9)   |
| Lateral root primordia     | RM1000                            |
| Procambium                 | WOL (2-9)                         |

Table 1. The 14 cell types in the *Arabidopsis* root and the 17 marker-lines which mark them [4]. For markers that only mark the cell type in some of the sections, these sections are indicated by the range in parenthesis.
Table 2. The cell count matrix gives the number of cells in each spatiotemporal subregion. The 13 rows correspond to longitudinal sections 1 through 13. From left to right, the 14 columns correspond to the following spatiotemporal subregions: quiescent center, columella, lateral root cap, hair cell, non-hair cell, cortex, endodermis, xylem pole pericycle, phloem pole pericycle, phloem, phloem companion cells, xylem, lateral root primordia, and procambium.

|          | 0  | 24 | 51 | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
|----------|----|----|----|---|----|----|----|----|----|----|----|----|----|
|          | 4  | 12 | 152| 24| 48 | 12 | 12 | 12 | 22 | 0  | 0  | 12 | 0  |
|          | 0  | 0  | 280| 40| 80 | 40 | 40 | 20 | 45 | 20 | 20 | 25 | 0  |
|          | 0  | 0  | 210| 40| 80 | 40 | 40 | 20 | 45 | 20 | 20 | 25 | 0  |
|          | 0  | 0  | 210| 40| 80 | 40 | 40 | 20 | 45 | 20 | 20 | 25 | 0  |
|          | 0  | 0  | 210| 40| 80 | 40 | 40 | 20 | 45 | 20 | 20 | 25 | 0  |
|          | 0  | 0  | 210| 40| 80 | 40 | 40 | 20 | 45 | 20 | 20 | 25 | 0  |
|          | 0  | 0  | 210| 40| 80 | 40 | 40 | 20 | 45 | 20 | 20 | 25 | 0  |
|          | 0  | 0  | 210| 40| 80 | 40 | 40 | 20 | 45 | 20 | 20 | 25 | 130|
|          | 4  | 0  | 0  | 40| 80 | 40 | 40 | 20 | 45 | 20 | 20 | 25 | 0  |
|          | 0  | 0  | 0  | 40| 80 | 40 | 40 | 20 | 45 | 20 | 20 | 25 | 0  |
|          | 0  | 0  | 0  | 40| 80 | 40 | 40 | 20 | 45 | 20 | 20 | 25 | 0  |

spatiotemporal subregions marked by that marker as indicated in Table 1. Note that the non-zero entries of $a_{s+k}$ may span multiple columns for those markers which are listed in multiple rows of Table 1. The non-zero entries of $a_{s+k}$ are proportional to the corresponding entries of the cell matrix, but rescaled to sum to 1.

2.3. Cell matrix. As described in the previous section, the coefficients $a_{ijk}$ in our model depend on the number of cells in each spatiotemporal subregion. These cell number estimates were generated by visual inspection of successive optical cross-sections of *Arabidopsis* roots along the longitudinal axis using confocal laser scanning microscopy. For the xylem, phloem and procambium cell types, cell counts were obtained from earlier experiments [3, 14]. What follows is a detailed description of this visual and literature analysis. These results are also summarized in Table 2.

Longitudinal section 1 encompasses two tiers of 12 columella cells, and three tiers of lateral root cap cells (15, 18 and 18 moving up from the tip).

Longitudinal section 2 contains one tier of 12 columella cells and six tiers of lateral root cap cells (20, 20, 28, 28, 28 and 28 moving up from the tip). For all other cell types in longitudinal section 2, three tiers of cells are present. Eight trichoblast (hair cell precursor) cells and 16 atrichoblast (non-hair cell precursor) cells are present circumferentially throughout the
root, resulting in 24 and 48 cells respectively in the hair cell and non-hair cell precursor files in longitudinal section 2. Throughout the root, eight cortex and eight endodermis cells are present circumferentially. However in longitudinal section 2, the cortex/endodermis initial is undergoing asymmetric periclinal divisions to produce the cortex and endodermis cell files, so we consider there to be approximately 0.5 cells of the cortex and endodermis type, resulting in 12 cells of each type in longitudinal section 2. When the Arabidopsis root is seven days old, each longitudinal section from 3–13 contains approximately five cells of each type along the root’s longitudinal axis.

In longitudinal section 2, the tangential and periclinal divisions that give rise to phloem cell files do not occur, but do occur in longitudinal section 3 [3]. Three cells are present in the main xylem axis in the first tier of cells, four cells in the second tier, and five cells in the third tier [14]. Eight procambial cells are present in the first cell tier, 12 procambial cells in the second tier, and 18 cells in the third tier resulting in 28 procambial cells in longitudinal section 2 [14]. For all sections xylem pole pericycle cells are the two cells that flank the xylem axis on either end, and phloem pole pericycle cells are considered the intervening cells. Four pericycle cells can be identified as flanking xylem cells in all three tiers of cells present in longitudinal section 2 [14]. Seven intervening phloem pole pericycle cells can be found in tier one, and eight intervening cells can be identified in the third tier [14], resulting in 22 procambial cells in longitudinal section 2.

In a seven day old root, each of the longitudinal sections 3–13 contains approximately five tiers of cells. In longitudinal section 3, columella cells can no longer be identified, and 10 tiers of lateral root cap cells exist containing 28 cells each. In sections 4–6, a lateral root cap cell is twice the length and half the width of an epidermal cell. Eighty-four cells were identified in each tier, and two and a half tiers of cells exist each for longitudinal sections 4–6 resulting in 210 cells for each longitudinal section. All other cell types have undergone the appropriate tangential and periclinal divisions to establish their respective cell files by longitudinal section 3. Two protophloem cells, two metaphloem cells and four accompanying companion cells are present in the phloem tissue [3]. With the combination of protophloem and metaphloem cells, 20 phloem cells and 20 companion cells exist in each longitudinal section. Approximately 40 procambial cells exist in each longitudinal section. Secondary cell growth does not occur in the developmental stages sampled, therefore, this number remains fixed throughout all developmental stages. In longitudinal section 12, a non-emerged lateral root is hypothesized to be present based on microarray expression data [4]. This lateral root is estimated to be approximately 130 cells, or one tier of cells in longitudinal section 2.

In our modelling the distinct vasculature, protophloem and metaphloem cell types were treated as a single cell type, as no marker-line was specific enough to differentiate clearly between these cell types. Also, the metaxylem and protoxylem were considered as a single cell type by the same rationale.
2.4. Solving Bilinear Equations. In this section we present our method for solving the system of bilinear equations given by (1). More generally, we have a system

\[(2) \quad f(x, y) := \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ijk} x_i y_j = b_k \quad \text{for } k = 1, \ldots, \ell.\]

In our application, we have \(n = 13\), \(m = 14\), and \(\ell = 30\). Unlike other numerical methods for solving systems of polynomial equations, our algorithm has the advantage that it finds only non-negative, real solutions. Moreover, even in systems where there are no exact solutions, as will generally be the case in an overconstrained system of equations, our method will find approximate solutions. A more detailed, technical mathematical study of the method will be available in a forthcoming paper by the first author.

Our method is based on the Expectation-Maximization (EM) [15, Sec. 1.3] and Iterative Proportional Fitting (IPF) [7] algorithms used for maximum likelihood estimation in statistics. These are iterative algorithms which reduce the modified Kullback-Leibler divergence at each step:

\[(3) \quad D(f(x, y) \parallel b) = \sum_{k=1}^{\ell} \left( b_k \log \left( \frac{b_k}{f_k(x, y)} \right) - b_k + f_k(x, y) \right).\]

The traditional Kullback-Leibler divergence consists only of the first term in the summation. The other two terms are a natural generalization, which is necessary only when the vectors \(f(x, y)\) and \(b\) do not sum to one [11].

Our method begins with an arbitrarily chosen starting point \((x^{(0)}, y^{(0)})\) in \(\mathbb{R}^{m+n}_>\). In each iteration \(s\), the expectation step computes the quantities:

\[(4) \quad w_{ijk}^{(s)} := \frac{a_{ijk} x_i^{(s)} y_j^{(s)}}{\sum_{i'=1}^{m} \sum_{j'=1}^{n} a_{i'j'k} x_{i'}^{(s)} y_{j'}^{(s)}}\]

for all \(i, j,\) and \(k\). This quantity \(w_{ijk}^{(s)}\) is an estimate of the contribution of the \((i, j)\) term in the \(k\)th equation in (2). The maximization step is an analogue of the IPF algorithm, and itself consists of an iteration. We first compute the analogues of the sufficient statistics:

\[
X_i^{(s)} = \sum_{j=1}^{n} \sum_{k=1}^{\ell} w_{ijk}^{(s)} \\
Y_j^{(s)} = \sum_{i=1}^{m} \sum_{k=1}^{\ell} w_{ijk}^{(s)}.
\]
Then we perform an iteration beginning with \( x_i^{(s,0)} = x_i^{(s)} \) and \( y_j^{(s,0)} = y_j^{(s)} \) and the update rules

\[
x_i^{(s,t+1)} := x_i^{(s,t)} \frac{X_i^{(s)}}{\sum_{j=1}^{m} \sum_{k=1}^{\ell} a_{ijk} x_i^{(s,t)} y_j^{(s,t)}}
\]

\[
y_j^{(s,t+1)} := y_j^{(s,t)} \frac{Y_j^{(s)}}{\sum_{i=1}^{n} \sum_{k=1}^{\ell} a_{ijk} x_i^{(s,t+1)} y_j^{(s,t)}}
\]

until the parameters converge. We then re-normalize and use the values from the last index \( t \) for the next step of the EM algorithm:

\[
x_i^{(s+1)} := \frac{x_i^{(s,t)}}{\sum_{i'=1}^{n} x_i^{(s,t)}}
\]

\[
y_j^{(s+1)} := y_j^{(s,t)} \sum_{i'=1}^{m} x_i^{(s,t)}
\]

At each step of each of these algorithms, the Kullback-Leibler divergence defined in (3) decreases. In the statistics literature, the convergence of the EM and IPF algorithms is known under the additional assumptions that

\[
\sum_{i=1}^{n} x_i = \sum_{j=1}^{m} y_j = \sum_{k=1}^{\ell} b_k = 1
\]

However, relaxing these conditions does not change the convergence proof.

We repeatedly ran our EM algorithm beginning with 20 different randomly chosen starting points. For each transcript in the data, all 20 runs of the algorithm converged to the same solution, strongly suggesting that we have found a global minimum to the modified Kullback-Leibler divergence.

2.5. Computational validation methodology. In order to validate our method, we simulated expression profiles according to various models and tested our method’s ability to reconstruct the underlying parameters. First, we simulated data according to the same independence model defined in Section 2.2. The underlying spatiotemporal subregion expression levels were sampled from a log-normal distribution with standard deviation 0.5. The simulated measurements \( b_k \) were computed from these subregion levels according to our model of the Arabidopsis root in (1). Finally, multiplicative error was added, distributed according to a log-normal distribution with standard deviation 0.03 to simulate measurement noise. This procedure created expression data with varying but comparable expression levels, which we will call the “uniform” dataset. However, since we are particularly interested in genes for which the expression levels are not uniform, we also produced simulations with the expression level for a given section or cell type raised by a factor of 10, which we will call the “elevated” dataset. In this dataset, we only measured the error for the same section or cell type which
was elevated. These simulations measure our ability to detect a dominant expression pattern.

In addition, we designed simulations that test the robustness of the algorithm to failures of the bilinear model for root expression levels. For each section and cell type, we simulated data in which the expression levels for cells in that section or cell type did not follow the bilinear model, and call these the “section” and “cell type” datasets respectively. Instead, the expression levels in the given section or cell type were chosen independently according to a log-normal distribution with standard deviation $0.5\sqrt{2}$. The factor of $\sqrt{2}$ was introduced because the product of two log-normally distributed numbers with standard deviation $0.5$ is distributed log-normally with standard deviation $0.5\sqrt{2}$.

The predictions were compared to the true expression levels across the spatiotemporal subregions within each section and each cell type. For each section and each cell type, the expression levels in its spatiotemporal subregions were averaged, ignoring those combinations which are not physically present in the root, (i.e. those whose entry in Table 2 is 0). The difference between the predicted and true average expressions was computed as a proportion of the true average expression. We then computed the root mean square of the proportional error over 500 simulations.

### 2.6. Visualization of predicted expression patterns

Predicted expression values were colored according to an *Arabidopsis* root template (Figure 1). The green channel of each cell was set according to a linear mapping between the expression range shown in the template $[1,10]$ or $[1,5]$ to the range $[0,255]$. Expression values above or below that range are given values of 255 or 0 respectively. The mapping is also shown to the right of the false color image in the form of a gradient key. Phloem cells by longitudinal section are visualized separately on the right hand side of the root as they are physically occluded by other cells in the left hand side representation. The minimum and maximum range of expression value visualized can also be adjusted by the user.

### 2.7. In vivo validation methodology

To validate predicted expression values, we used transgenic *Arabidopsis thaliana* lines containing transcriptional GFP fusions in the Columbia ecotype [12]. For each gene being validated, six plants from at least two insertion lines previously described as expressing GFP were characterized. All plants were grown vertically on 1X Murashige and Skoog salt mixture, 1% sucrose and 2.3 mM 2-(N-morpholino)ethanesulfonic acid (pH 5.7) in 1% agar. Seedlings were prepared for microscopy at 5 days of age. Confocal images were obtained using a 25x water-immersion lens on a Zeiss LSM-510 confocal laser-scanning microscope using the 488-nm laser for excitation. Roots were stained with 10 $\mu$g/mL propidium iodide for 0.5 to 2 minutes and mounted in water. GFP was rendered in green and propidium iodide in red. Images were saved in TIFF format. Images were manually stitched together in Adobe Photoshop.
CS2 using the Photomerge command. The black background surrounding the root was modified to ensure uniformity across figures. No other image enhancement was performed.

3. Results

3.1. Computation validation. The root mean square percentage errors in the reconstruction of each parameter are shown in Table 3. In the first two columns, where the data were generated according to the bilinear model, the error rate is generally no greater than the simulated measurement error. In most cases, elevated expression led to a lower error rate. In particular, reconstruction of expression in procambium was much more accurate in the elevated dataset.

The last two columns show that the algorithm is robust to violations of the bilinear model. Also, the predicted expression level in each cell type is generally not greatly affected by the failure of the model in other cell types, and similarly with sections.

3.2. In vivo validation. To determine whether our algorithm is able to accurately resolve spatiotemporal subregion-level transcript expression values, it would be ideal to compare the predictions to measured microarray expression values of the same spatiotemporal subregion. However, due to technical constraints, it is not possible to measure mRNA expression to such a degree of specificity and thus we cannot validate the estimates directly. Instead, we validated the method by visually comparing the predicted pattern of expression to patterns obtained from transcriptional GFP fusions using laser scanning confocal microscopy, as described in [12].

For each gene validated, a false-colored root image was generated by coloring each spatiotemporal subregion of an annotated Arabidopsis root template (Figure 1) according to the expression level in that subregion as predicted by our method. This false-colored image was then visually compared against the actual pattern of fluorescence observed in plants expressing a transcriptional GFP fusion specific for the promoter of that gene. These transcriptional GFP fusions contain up to 3 kb of regulatory sequence upstream of the translational start site of the respective gene. In many cases, this sequence is sufficient to recapitulate endogenous mRNA expression patterns as defined by cell type resolution microarray data [12]. This comparative method of validation allows us to assess the accuracy of spatiotemporal subregion expression predictions in an efficient and technically feasible way.

As a benchmark validation test, a set of three transcriptional fusions which were used to obtain some of the marker-line dataset were examined: S18 (AT5G12870), S4 (AT3G25710), and S32 (AT2G18380). These fusions were originally selected for use in profiling because they exhibited enriched cell type expression as observed by laser scanning confocal microscopy and subsequently confirmed in the microarray expression data. The expression
predictions from our method accurately recapitulated the observed pattern of all three benchmark genes (Figure 2 and data not shown).

To assess the novel predictive ability of our method to reconstruct \textit{in vivo} expression patterns given missing data, we selected transcriptional fusions
for genes for which our method predicts expression in cell types or in spatiotemporal subregions that were not marked by fluorescent marker-lines in the original dataset. At least two lines per transcriptional fusion were monitored. With respect to an unmarked cell type, we selected a candidate gene predicted by our model to be highly expressed in the columella and developing procambium. Imaging of a transcriptional fusion of this gene confirmed this expression (Figure 3).

We next determined if our method could correctly differentiate expression in a specific developmental stage of a cell type. The collection of marker-lines used to generate the original dataset included a marker for all developmental stages of non-hair cells, composed of their precursors (atrichoblasts) and fully developed non-hair cells. However, the marker-line used for hair cells only marks mature hair cells, and not their precursors (trichoblasts). Using predictions from our method we tested a candidate gene with predicted expression throughout the epidermis—in mature hair cell, trichoblast, mature non-hair cell and atrichoblast cell files—with higher expression predicted in non-hair cells than in hair cells. This differential expression was validated using a transcriptional fusion (Figure 4) demonstrating that our method is not only able to identify expression in a developmental stage of a cell type not marked by the marker-line data, but also to accurately differentiate relative levels of a transcript. However, it should be noted that expression in the transcriptional fusion did not fully corroborate the expression predicted by our algorithm—specifically, expression was found in the lateral root cap which was not predicted by our algorithm.

Examination of the raw microarray expression data revealed that expression was not elevated in the lateral root cap in the input microarray data. Most likely, the presence of GFP is not indicative of erroneous reconstruction of gene expression in this case. Instead, the transcriptional fusion does not contain sufficient regulatory elements to direct the appropriate expression as described in [12], perhaps within downstream sequences. For this reason, a comparison of the ratio between raw marker line and section expression data can be obtained as a link for each gene so that the user can simultaneously assess raw expression data with the reconstructed expression patterns.

4. Discussion

We have shown that spatiotemporal patterns of gene expression in the Arabidopsis root can be reconstructed using information from the marker-line and longitudinal datasets. Current experimental techniques are limited in their ability to rapidly and accurately microdissect organs into all component cell types at all developmental stages. Our computational technique helps to overcome these limitations. We fully integrate the marker-line and longitudinal data sets into a comprehensive expression pattern, across both space and time. In particular, this method has enabled the identification
of Arabidopsis root procambium and trichoblast-specific genes, which have been previously experimentally intractable cell types.

Our high-resolution expression patterns will allow us to better understand the regulatory logic that controls developmental processes of the Arabidopsis root. These transcriptional regulatory networks are key to understanding developmental processes and environmental responses. With only a portion of these genes and fewer cell types, high-resolution spatiotemporal data has been used to identify transcriptional regulatory modules [4]. Our more accurate and complete dataset will allow a more comprehensive discovery of regulatory networks across additional cell types.

Moreover, we expect that our algorithm and the model which underlies it are applicable to time course experiments on other heterogeneous cell mixtures. Measurements in multicellular organisms are taken from complex cell mixtures of organs, tissues, heterogeneous cell lines, or cancerous samples. When precise histological characterization of these samples can estimate underlying cell type composition, our method can be used to reconstruct the underlying cell type-specific gene expression patterns or any other type of quantitative data, such as high-throughput protein abundance measurements. Theoretically, this algorithm can be applied to identify missing data in any experimental system that captures data in two or more dimensions which are assumed to be independent of one another.

5. Acknowledgments

The authors acknowledge the support of the DARPA project “Microstates to Macrodynamics: A New Mathematics of Biology.” DAC and BS were supported by the U.S. National Science Foundation (DMS-0354321, DMS-0456960, and DMS-0757236). SMB was an NSERC postdoctoral scholar. SMB, DAO and PNB acknowledge support through the NSF AT2010 program. The authors would like to thank Wolfgang Busch, Anjali Iyer-Pascuzzi, Terri Long, Anne Shiu, and Rossangela Sozzani for critical review of the manuscript.

References

[1] K. Birnbaum, J.W. Jung, J.Y. Wang, G.M. Lambert, J.A. Hirst, D.W. Galbraith, and P.N. Benfey. Cell type-specific expression profiling in plants via cell sorting of protoplasts from fluorescent reporter lines. *Nature Methods*, 2(8):615, 2005.
[2] K. Birnbaum, D.E. Shasha, J.Y. Wang, J.W. Jung, G.M. Lambert, D.W. Galbraith, and P.N. Benfey. A gene expression map of the Arabidopsis root. *Science’s STKE*, 302(5652):1956, 2003.
[3] M. Bonke, S. Thitamadee, A.P. Mähön, M.T. Wauser, and Y. Helariutta. APL regulates vascular tissue identity in Arabidopsis. *Nature*(London), 426(6963):181–186, 2003.
[4] Siobhan M. Brady, David A. Orlando, J. Y. Lee, J. Y. Wang, J. Koch, J. R. Dimmey, D. Mace, U. Ohler, and Philip N. Benfey. A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*, 318(5851):801–806, 2007.
[5] W. Busch and J.U. Lohmann. Profiling a plant: expression analysis in Arabidopsis. Current Opinion in Plant Biology, 10(2):136–141, 2007.

[6] B. Chaudhuri, F. Hörmann, S. Lalonde, S.M. Brady, D.A. Orlando, P.N. Benfey, and W.B. Frommer. Protonophore- and pH-insensitive glucose and sucrose accumulation detected by FRET nanosensors in Arabidopsis root tips. Plant J., Epub ahead of print, 2008.

[7] J. N. Darroch and D. Ratchiff. Generalized iterative scaling for log-linear models. Annals of Math. Stat., 43(5):1470–1480, 1972.

[8] Debashis Ghosh. Mixture models for assessing differential expression in complex tissues using microarray data. Bioinformatics, 20:1663–1669, 2004.

[9] Anjali S. Iyer-Pascuzzi and Philip N. Benfey. Transcriptional networks in root cell fate specification. Biochim. Biophys. Acta., 2008.

[10] Yuling Jiao, S Lori Tausta, Neeru Gandotra, Ning Sun, Tie Liu, Nicole K Clay, Teresa Ceserani, Meiqin Chen, Ligeng Ma, Matthew Holford, Hui yong Zhang, Hongyu Zhao, Xing-Wang Deng, and Timothy Nelson. A transcriptome atlas of rice cell types uncovers cellular, functional, and developmental hierarchies. Nature Genetics, 41:258–263, 2009.

[11] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. Adv. Neural Info. Proc. Syst., 13:556–562, 2001.

[12] J.Y. Lee, J. Colinas, J.Y. Wang, D. Mace, U. Ohler, and P.N. Benfey. Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots. Proceedings of the National Academy of Sciences, 103(15):6055–6060, 2006.

[13] M. Levine and EH Davidson. Gene regulatory networks for development. Proc Natl Acad Sci USA, 102(14):4936–42, 2005.

[14] A.P. Mahonen, M. Bonke, L. Kauppinen, M. Riikonen, P.N. Benfey, and Y. Helariutta. A novel two-component hybrid molecule regulates vascular morphogenesis of the Arabidopsis root. Genes & Development, 14(23):2938–2943, 2000.

[15] Lior Pachter and Bernd Sturmfels. Algebraic Statistics for Computational Biology. Cambridge University Press, 2005.

[16] D. Venet, F. Pecasse, C. Maenhaut, and H. Bersini. Separation of samples into their constituents using gene expression data. Bioinformatics, 17:S279–S287, 2001.

(Dustin A. Cartwright) DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, BERKELEY, CA
E-mail address: dustin@math.berkeley.edu

(Siobhan M. Brady) DEPARTMENT OF PLANT BIOLOGY AND GENOME CENTER, UNIVERSITY OF CALIFORNIA, DAVIS, CA
E-mail address: sbrady@ucdavis.edu

(David A. Orlando) DEPARTMENT OF BIOLOGY AND PROGRAM IN COMPUTATIONAL BIOLOGY AND BIOINFOMATICS, DUKE UNIVERSITY, DURHAM, NC
E-mail address: david.orlando@duke.edu

(Bernd Sturmfels) DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, BERKELEY, CA
E-mail address: bernd@math.berkeley.edu

(Philip N. Benfey) DEPARTMENT OF BIOLOGY AND DUKE CENTER FOR SYSTEMS BIOLOGY, DUKE UNIVERSITY, DURHAM, NC
E-mail address: benfeyp@duke.edu
| Probe      | Gene(s)                             | Longitudinal | Marker-line | Ratio  |
|------------|-------------------------------------|--------------|-------------|--------|
| 246630     | AT1G50730                           | 1.073        | 1.239       | 0.866  |
| 246980     | AT5G67530                           | 1.676        | 1.605       | 1.045  |
| 247163     | AT5G65685                           | 1.423        | 1.508       | 0.944  |
| 247286     | AT5G64280                           | 1.146        | 1.292       | 0.887  |
| 247334     | AT5G63610                           | 1.75         | 1.497       | 1.169  |
| 247363     | AT5G63200                           | 1.225        | 1.305       | 0.938  |
| 248266     | AT5G53440                           | 2.514        | 1.867       | 1.347  |
| 250524     | AT5G08520                           | 1.691        | 1.54        | 1.099  |
| 250791     | AT5G05610                           | 1.151        | 1.334       | 0.863  |
| 251104     | AT5G01720;AT5G01715                 | 1.214        | 1.257       | 0.966  |
| 251233     | AT3G62800                           | 1.171        | 1.204       | 0.973  |
| 252157     | AT3G50430                           | 1.398        | 1.21        | 1.155  |
| 253409     | AT4G32960                           | 1.406        | 1.572       | 0.894  |
| 253565     | AT4G31200                           | 1.461        | 1.326       | 1.101  |
| 253826     | AT4G27960;AT5G53300                 | 26.884       | 26.78       | 1.004  |
| 255253     | AT4G05000                           | 1.167        | 1.696       | 0.688  |
| 255704     | AT4G00170                           | 1.35         | 1.853       | 0.729  |
| 255725     | AT1G25540                           | 1.507        | 1.636       | 0.921  |
| 255838     | AT2G33490                           | 1.691        | 1.473       | 1.148  |
| 255946     | AT1G22020                           | 1.335        | 1.481       | 0.901  |
| 256236     | AT3G12350                           | 1.293        | 1.978       | 0.654  |
| 256907     | AT3G24030                           | 1.601        | 1.751       | 0.915  |
| 256961     | AT3G13445                           | 1.383        | 1.441       | 0.96   |
| 258269     | AT3G15690                           | 1.794        | 1.489       | 1.205  |
| 259243     | AT3G07565                           | 1.678        | 1.768       | 0.949  |
| 259280     | AT3G01150                           | 1.302        | 1.941       | 0.671  |
| 259313     | AT3G05090                           | 1.275        | 1.642       | 0.776  |
| 259341     | AT3G03740                           | 1.342        | 1.606       | 0.835  |
| 259800     | AT1G72175                           | 1.159        | 1.4         | 0.828  |
| 260133     | AT1G66340                           | 1.278        | 1.636       | 0.781  |
| 261348     | AT1G79810                           | 1.16         | 1.464       | 0.792  |
| 261515     | AT1G71800                           | 1.236        | 1.366       | 0.905  |
| 261634     | AT1G49970                           | 1.468        | 1.732       | 0.847  |
| 261666     | AT1G18440                           | 1.258        | 1.297       | 0.97   |
| 261744     | AT1G08490                           | 1.265        | 1.384       | 0.914  |
| 262089     | AT1G56000;AT1G55980                 | 1.554        | 1.194       | 1.301  |
| 262379     | AT1G73020                           | 1.15         | 1.398       | 0.823  |
| 262672     | AT1G76050                           | 1.306        | 1.356       | 0.963  |
| 262860     | AT1G64810                           | 1.412        | 1.39        | 1.016  |
| 263984     | AT2G42670                           | 1.327        | 1.314       | 1.01   |
| 264307     | AT1G61900                           | 1.273        | 1.718       | 0.741  |
| 265129     | AT1G30970                           | 1.333        | 1.411       | 0.945  |
| 267401     | AT2G26210                           | 1.198        | 1.476       | 0.812  |

| Table 4. | Mean expression values and scaling factors of ubiquitously, evenly expressed probesets across longitudinal and marker-lines |
Figure 1. *Arabidopsis* root template used for expression pattern overlays.
Figure 2. (A) Expression of AT2G18380 in all developmental stages of the phloem was predicted by our method and visualized in a representation of the Arabidopsis root. Phloem cells are shown external to the root. (B) GFP expression in the longitudinal axis and (C) expression in cross-section of expression driven by the AT2G18380 promoter validate the prediction.
Figure 3. (A) Our method correctly predicts specific expression of a candidate gene in a cell type, procambium, that is only covered by a general tissue marker, WOL. Expression conferred by the candidate gene promoter fused to GFP as a reporter was visualized in the columella (B) and in the procambium by a longitudinal section (B) and a cross section (C). The label X indicates the xylem axis. The expression also validates a maximal peak in the meristematic zone.
Figure 4. (A) Our method correctly predicts candidate gene expression in trichoblast cells in the meristematic zone, which are not currently covered by any marker-lines. Furthermore, the algorithm was able to predict differential expression within epidermal tissue with high expression in non-hair cells and atrichoblasts (immature non-hair cells in the meristematic and elongation zone) and decreased expression in hair cells or trichoblasts (immature hair cells in the meristematic and elongation zone). Expression conferred by the candidate gene promoter fused GFP as a reporter was visualized in the epidermis in a longitudinal section (B) and was specifically identified as high in atrichoblasts, and lower in trichoblasts (marked with an asterisk) in cross-section (C). Trichoblasts or hair cells differentiate at the junction between two underlying cortical cells.