BIOINFORMATICS

Vol. 29 ISMB/ECCB 2013, pages i27–i35
doi:10.1093/bioinformatics/btt206

Automated annotation of gene expression image sequences via non-parametric factor analysis and conditional random fields

Iulian Pruteanu-Malinici[1], William H. Majoros[1] and Uwe Ohler[1,2,3,*]

[1]Institute for Genome Sciences & Policy, Duke University, Durham, NC 27708, USA, [2]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27708, USA and [3]Berlin Institute for Medical Systems Biology, Max Delbruck Center for Molecular Medicine, Berlin, Germany

ABSTRACT

Motivation: Computational approaches for the annotation of phenotypes from image data have shown promising results across many applications, and provide rich and valuable information for studying gene function and interactions. While data are often available both at high spatial resolution and across multiple time points, phenotypes are frequently annotated independently, for individual time points only. In particular, for the analysis of developmental gene expression patterns, it is biologically sensible when images across multiple time points are jointly accounted for, such that spatial and temporal dependencies are captured simultaneously.

Methods: We describe a discriminative undirected graphical model to label gene-expression time-series image data, with an efficient training and decoding method based on the junction tree algorithm. The approach is based on an effective feature selection technique, consisting of a non-parametric sparse Bayesian factor analysis model. The result is a flexible framework, which can handle large-scale data with noisy incomplete samples, i.e. it can tolerate data missing from individual time points.

Results: Using the annotation of gene expression patterns across stages of *Drosophila* embryonic development as an example, we demonstrate that our method achieves superior accuracy, gained by jointly annotating phenotype sequences, when compared with previous models that annotate each stage in isolation. The experimental results on missing data indicate that our joint learning method successfully annotates genes for which no expression data are available for one or more stages.

Contact: uwe.ohler@duke.edu

1 INTRODUCTION

The use of high-throughput image acquisition, such as in phenotypic screens, has been quickly increasing and thus provides a new source of data for computational biologists. Microscopy of colored or fluorescent probes, followed by imaging, is able to deliver spatial and temporal quantitative phenotype information such as gene expression at high resolution (Busch *et al*., 2012; Ljosa *et al*., 2009; Walter *et al*., 2010). In addition, expression patterns can be documented and distributed over the internet as a valuable resource to the research community. Recent advances in throughput, or long-term investment in specific projects, have by now generated large collections of images. Such image databases are traditionally analyzed through direct inspection by human curators; an example is the Berkeley *Drosophila* Genome Project (BDGP) gene expression pattern database (Tomancak *et al*., 2002, 2007). In this dataset, images are assigned to stage ranges within the 17 embryonic stages defined by developmental features, and annotated collectively in small groups using a controlled vocabulary (CV), i.e. annotation terms. This allows researchers to search image databases and compare spatial and temporal embryonic development.

Given the very diverse nature of imaging technology, samples and biological questions, computational approaches have often been tailored to a specific dataset. For example, the image-based profiling of gene expression patterns via *in situ* hybridization (ISH) requires the development of accurate and automatic image analysis systems for using such data, to understand regulatory networks and development of multicellular organisms. Images are affected by multiple sources of noise due to experiments or microscopy (incomplete or multiple embryos, variance of probes across genes, illumination artifacts), making the extraction and registration of embryos non-trivial (Fowlkes *et al*., 2005, 2008; Harmon *et al*., 2007; Keranen *et al*., 2006; Kumar *et al*., 2002; Mace *et al*., 2010; Puniyani *et al*., 2010;). Peng and Myers (2004) and Peng *et al*. (2007) introduced an automatic image annotation framework using various high-dimensional feature representations and classifying frameworks: Principal Component Analysis (PCA), wavelets, Gaussian mixture models, Support Vector Machines (SVM), Quadratic Discriminant Analysis. Each image may show the embryo under different views: lateral, dorsal or ventral; this is a challenge for gene annotation, as embryonic structures may be visible in only certain views. Yet, recent studies have shown that incorporating images from multiple views could consistently improve the annotation accuracy (Ji *et al*., 2009; Pruteanu-Malinici *et al*., 2011).

It is desirable to represent images in a way that takes advantage of image features and offers robustness to image distortions. In contrast to such large feature sets prone to high redundancy and high computational costs, Frise *et al*. (2010) identified a set of basic expression patterns in *Drosophila*. A set of 39 well-defined clusters describing specific regions of embryo expression were determined from 2693 lateral views of early development. As with the majority of described approaches, this study involved a high level of human intervention in selecting 'good' images for training/testing purposes—a potential drawback, considering the rapid increase in the size of ISH image collections. In contrast, Pruteanu-Malinici *et al*. (2011) proposed a new approach for automatic annotation of spatial expression patterns using a 'vocabulary' of basic patterns that involved little to no human intervention. This work provided a flexible unsupervised framework in competitively predicting gene annotation terms, while using only a small set of features.

*To whom correspondence should be addressed.

A particular aspect that has been largely neglected by computational approaches so far is that data acquired from such experiments often span multiple time points or conditions. Phenotypes are typically annotated stage-by-stage, without jointly learning the salient temporal dependencies across multiple time points, which should allow for an overall higher accuracy; e.g. the annotation terms predicted for earlier stages should inform the prediction at later stages. Furthermore, many genes are annotated with more than one term from the vocabulary, creating an additional dependency structure between annotations within the same stage range.

In this article, we address the automatic annotation of *Drosophila* embryo gene expression *sequences*, building on state-of-the-art models from computer vision and machine learning. There are several challenges that need to be addressed when approaching this problem through computational methods. As we mentioned previously, the image acquisition process results in embryonic structures with multiple perspectives, shapes and locations. Moreover, the shape/position of the same embryonic structure may vary from image to image: 'variation in morphology and incomplete knowledge of the shape and position of various embryonic structures' have made the gene annotation task more prohibitive (Ji *et al.*, 2008).

We first show that a non-parametric Bayesian factor analysis (BFA) approach, the infinite factor model, allows for an efficient and sparse feature representation of the *Drosophila* gene expression dataset. Then, we propose a conditional random field (CRF) to tackle the time-evolving annotation task. Experiments show that the exploitation of dependencies across adjacent developmental stages leads to annotation accuracy superior to existing *Drosophila* gene expression annotation approaches. The proposed framework also tackles the missing data scenario: for many genes, one or more stage ranges are absent from the image collection; in such cases, human annotators would take into account the entire set of expression data from adjacent stages to successfully annotate the available images. The challenge to automatize this process is novel and represents a step closer toward a fully automatic gene annotation pipeline. These predictions can be later analyzed by biologists to assess the correctness of the image acquisition and the level of interest for that particular gene. Finally, for a given gene, the described framework predicts the entire *set* of annotation terms simultaneously, taking full advantage of the term dependencies that exist at the stage-range level.

The rest of this article is organized as follows: in Section 2, we focus on data description and introduce the sparse BFA-CRF framework. Experimental results are reported in Section 3, followed by conclusions and future work in Section 4.

2 MATERIALS AND METHODS

2.1 Data description

One of the most popular large-image expression datasets is the BDGP collection of embryonic expression patterns. The project started with a first release of images for 2000 genes; the second release was in 2007 with 6000 genes. Release number 3 came in 2010 bringing the total to 7500 genes, including 97% of the sequence-specific transcription factor genes. As of today, the collection consists of over 105 000 images, which document patterns of embryonic gene expression for over 7400 of the 13 659 protein-coding genes identified in the *Drosophila melanogaster* genome. Expression is visualized by RNA ISH, which provides an effective way of locating specific mRNA sequences by hybridizing complementary mRNA-binding oligonucleotides and a suitable dye (Tautz *et al.*, 1989).

The mRNA expression apparent in the captured *in situ* images was verified by independently derived microarray time-course analysis using Affymetrix GeneChip technology (more details can be found at http://insitu.fruitfly.org, and in Tomancak *et al.*, 2002). Gene expression patterns were documented by taking low (2×) and high (20×) magnification images, at multiple developmental stages. The low-magnification digital images were taken to capture groups of embryos, to provide a permanent record of the hybridization in each well. Each slide was then further examined under higher magnification using a Zeiss Axiophot optical microscope. Images were assigned to developmental stage ranges following the sequence of events taking place at specific times after fertilization, using the 17 stages defined in (Campos-Ortega, 1985). In this analysis, we focused on the first 15 hr of *Drosophila* development, spanning embryonic stages 4–6, 7–8, 9–10, 11–12 and 13–16. Developmental stages 1–3 were skipped owing to predominant ubiquitous expression patterns not of interest to our analysis.

Any gene is represented, on average, by approximately 12 images; however, the number of images per gene varies from 1 to 80. This variability reflects the BDGP strategy to document highly dynamic, complex and novel patterns, while lowering the number of images documenting common expression patterns. Among those, there are images with non-informative patterns due to poor-quality staining/washing or non-tissue–specific expression (maternal or ubiquitous). Images within the same window can show different patterns due to embryo orientation or the relatively long developmental time spanned by a stage range. Images are annotated with ontology terms from a CV describing developmental expression patterns (Fig. 1). This vocabulary has been developed and refined by FlyBase (The FlyBase Consortium, 2002) over the past few years, allowing human curators to compare their findings with expression data assembled from the literature, expansion of annotations to greater detail and thorough searches of datasets based on Gene Ontology schema. The annotations used throughout this project consisted of a subset of about 300 of the 5800 annotation terms in the FlyBase CV, many of which only apply to later stages of development.

As mentioned previously, we use all available images in our approach, i.e. including those taken with any embryo orientation: lateral, dorsal and ventral. Before extracting features, we segmented and registered images using a previously described probabilistic segmentation approach based on statistical shape models (Mace *et al.*, 2010). This provides us with 240×120 pixel images, mostly containing a single embryo in a standard dorsal(up)/anterior(down) orientation and no background. In Figure 1, we show a particular gene expression pattern across five developmental stage ranges of interest. The complexity and variability of the image data led to competitive but not perfect results, in terms of precise embryo extraction as well as embryo orientation, which increased the challenge of automatic gene annotation.

We here use the average intensities in a down-sampled fixed grid size of 80×40 pixels as input features for the entire collection of images within the BDGP dataset.

2.2 Feature extraction—sparse BFA

Sparse Bayesian factor analysis (sBFA) is a statistical method for modeling many dependent random variables through linear combinations of a few hidden variables (Gorsuch, 1983). This model is designed to address the high-dimensional setting where hundreds or thousands of genes are simultaneously examined. The sparsity assumption is the key feature in our model that allows us to scale stable and accurate inference to a large number of images/genes represented by many image input features.

For high-dimensional models, sparsity helps prevent sampling errors from swamping out the true signal in data, leading to stable parameter estimates. In our framework, sparsity implies that each image/gene is

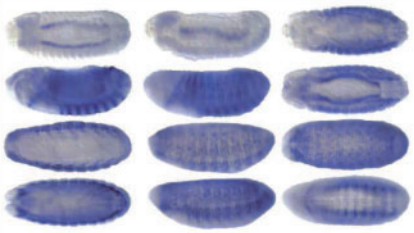
Developmental stage range	Gene expression	Annotation terms
4-6		cellular blastoderm, dorsal ectoderm anlage in statu nascendi, endoderm anlage in statu nascendi, yolk, ventral ectoderm **anlage in statu nascendi**, yolk nuclei, procephalic ectoderm anlage in statu nascendi, visual anlage in statu nascendi, amnioserosa anlage in statu nascendi, foregut anlage in statu nascendi
7-8		procephalic ectoderm anlage, ventral ectoderm primordium P2, trunk mesoderm primordium P2, hindgut **anlage**, anterior endoderm anlage, dorsal ectoderm primordium, posterior endoderm primordium P2, pole cell, head mesoderm primordium P4, yolk nuclei
9-10		procephalic ectoderm primordium, anterior endoderm primordium, dorsal ectoderm primordium, trunk mesoderm primordium, ventral nerve cord **primordium** P3, head mesoderm primordium P2, visual primordium, ventral ectoderm primordium, posterior endoderm primordium, foregut primordium
11-12		embryonic central brain neuron, yolk nuclei, anterior midgut primordium, brain primordium, mesectoderm **primordium**, foregut primordium, salivary gland body primordium, germ cell, dorsal epidermis primordium, somatic muscle primordium
13-16		embryonic central nervous system, embryonic foregut, embryonic head epidermis, embryonic midgut, embryonic **salivary gland**, sensory system **head**, embryonic/larval muscle system, embryonic central brain neuron, embryonic ventral epidermis, embryonic **optic lobe**

Fig. 1. Examples of *Drosophila* embryo ISH images and associated annotation terms (BDGP database) for gene Actn (FBgn0000667), across five developmental stage ranges. The dark blue stained regions highlight areas where genes are expressed; darker colors correspond to higher gene expression levels

affected only by a few underlying estimated factors, and as a result, many of the mixing weights in the model will be (near) zero.

The sparse Bayesian factor model was derived using the following matrix representation:

$$\mathbf{X} = \mathbf{AS} + \mathbf{E} \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_n]$ is a $p \times n$ dimensional data matrix, with n the number of features, quantifying the associated gene-expression values for p images (genes) under consideration. Each row of $\mathbf{X}$ is called a *gene pattern* with dimension $1 \times n$. Here, we assume that each gene pattern is already normalized to zero mean. $\mathbf{A}$ is the factor loading matrix with dimension $p \times k$, which contains the linear weights. $\mathbf{S}$ is the factor matrix with dimension $k \times n$, with each element modeled by a standard normal distribution. Each *column* of $\mathbf{S}$ is the factor score for feature i ($i = 1, 2, \ldots, n$) and each *row* is called a factor. $\mathbf{E}$ is the additive Gaussian noise with dimension $p \times n$. Both $\mathbf{A}$ and $\mathbf{S}$ are inferred by the model simultaneously.

From the model we can see that each row of $\mathbf{X}$ is modeled by a linear combination of the factors (rows of $\mathbf{S}$), indicating that the variability of the original p feature patterns can be explained by only k latent factors. The model can also be written in vector form as follows:

$$\mathbf{x}_j = \mathbf{a}_j\mathbf{S} + \varepsilon_j \ (j = 1, 2, ..., p) \tag{2}$$

where $\mathbf{x}_j$ and $\mathbf{a}_j$ denote the j^{th} row of $\mathbf{X}$ and $\mathbf{A}$, respectively, and the basis matrix $\mathbf{S}$ is shared across all samples. Indeed, factor analysis is an unsupervised dimensionality reduction method used widely in data analysis and signal processing (Prince *et al.*, 2008).

To impose the sparsity required by the underlying biological assumption where spatial gene expression patterns are modeled only by a few domains (factors), we used the Student-t distribution, which consists of a Gaussian distribution and a Gamma prior on the precision parameter. The sparseness is directly controlled by the precision parameter $\alpha_{j,m}$; the objective of imposing sparseness is to automatically shrink most elements in $\mathbf{A}$ near zero. The updating equations, along with a full description of the sparse factor model used in this manuscript can be found in Pruteanu-Malinici *et al.* (2011).

For an extension of our previous work, which largely focused on two developmental stage ranges only, the number of factors (k) for every developmental stage range needed to be determined in an ideally unbiased fashion. For this, we used an adaptive Gibbs sampler, which automatically truncated the loading and factor matrices through an adaptive selection of the number of important factors. This sparse Bayesian infinite factor model, first introduced by Bhattacharya and Dunson (2011), obviates the need for pre-specifying the number of factors; the effective number of factors (here denoted by k^*) is chosen such that the contribution from adding additional factors is negligible. This approach has been shown to produce accurate estimates of the true effective number of factors k^*; the adaptation of the Gibbs sampler occurs every 10 iterations at the beginning of the Markov chain but decreases in frequency exponentially fast, so as to satisfy the diminishing adaptation condition in Theorem 5 of Roberts and Rosenthal (2007). More specifically, the decreasing frequency is modeled as an exponential

$$p(t) = \exp(\alpha_0 + \alpha_1 t) \tag{3}$$

at the t^{th} Gibbs iteration with α_0 and α_1 chosen so that adaptation occurs every 10 iterations initially but then decreases in frequency exponentially fast. The loadings matrix is adaptively modified by monitoring the

columns with all elements within some pre-specified small neighborhood of zero. For some iterations, columns may be discarded or a new column could be simply added to the matrix; the remaining parameters of the model are modified accordingly. The parameters of the factors (in the case of adding some) are estimated from their prior distribution to fill in the necessary values.

2.3 Conditional random fields

Probabilistic graphical models such as Bayesian networks and random fields are increasingly popular choices for statistical modeling of complex biological relationships. Although Bayesian networks provide a viable solution for directed acyclic relationships where the direction of causality can be easily identified, undirected graphical models offer a clear advantage for highly connected relational structures that are not simple chains or trees. Among undirected models, CRFs (Lafferty *et al*., 2001) have proven to be among the most powerful predictors owing to their inherently discriminative (rather than generative) nature.

In a CRF, the observable variables ($\mathbf{X}=\{X_i\}$) and unobservable variables ($\mathbf{Y}=\{Y_i\}$) are treated separately, with the unobservables globally conditioned on the observables. The relationships among the unobservables are modeled via an undirected graph $G=(\mathbf{Y},\mathbf{E})$, in which the Y_is are the nodes (vertices), and edges $\mathbf{E}\subseteq\mathbf{Y}\times\mathbf{Y}$ are pairs (Y_i, Y_j); the edges serve to denote direct non-independence relations between pairs of Y_is. In particular, a node Y_i is taken to be conditionally independent of all other nodes Y_j, given the immediate neighbors N_G of Y_i in the graph:

$$P(Y_i|\{Y_{k\neq i}\}) = P(Y_i|N_G(Y_i)) \tag{4}$$

for $N_G(Y_i)=\{Y_{j\neq i}|(Y_i, Y_j)\in\mathbf{E}\}$.

The well-known Hammersley–Clifford theorem (Hammersley and Clifford, 1971) provides a means of computing conditional densities via decomposition of the graph into cliques. In particular,

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})} e^{\sum_{I\in cliques(G)} \lambda_I \Phi_I(\mathbf{Y}_I,\mathbf{X})} \tag{5}$$

where $\mathbf{Y}_I$ denotes the nodes in clique I, and $Z(\mathbf{X})$ is a normalizing constant; it is assumed that $P(\mathbf{Y})>0$ for all possible joint assignments to $\mathbf{Y}$. Φ_I is called the *potential function for clique I*; in practice, these are often pooled among like-sized cliques. Because cliques larger than some reasonable size N are typically ignored, modeling is accomplished by choosing a suitable set $\{\Phi_1, \Phi_2, \ldots \Phi_N\}$ of potential functions for different clique sizes; the λ_Is and any additional parameters of the Φ's can be trained discriminatively via cross validation.

Exact inference with a CRF is tractable if the graph can be converted into a chain or a tree. To this end, a *junction tree* can be obtained by collapsing tight clusters of nodes into meta-nodes and extracting a maximal spanning tree from the resulting structure (Jensen *et al*., 1990; Lauritzen and Spiegelhalter, 1988). The *sum–product algorithm* (Pearl, 1988) can then be applied to propagate local densities across the tree, permitting exact computation of posterior probabilities for joint or individual value assignments to nodes in the graph, or identification of the *maximum a posteriori* assignment; for linear-chain CRFs, these are analogous to the well-known Forward–Backward and Viterbi algorithms for hidden Markov models (Rabiner, 1989).

To infer the presence or absence of specific annotation terms for individual embryo images, we constructed a CRF structured as shown in Figure 2. Each node Y_i denotes the status of an annotation term: $Y_i=1$ means *present* (the annotation term applies to the image), $Y_i=0$ means *absent* (the annotation term does not apply). Columns correspond to developmental stages. All of the nodes in a column are directly connected via an edge to all nodes in adjacent columns (blue lines). Within a column, the nodes are connected in a linear chain (i.e. each node has exactly 1 or 2 neighbors within the column), with the order of the chain chosen so as to maximize the total mutual information between all adjacent pairs in the chain; this maximization was carried out via a standard traveling-salesman heuristic in Matlab. Each column was constrained to include only the most popular annotation terms in the training partition (for more details, see Results). The sparse image factors (previous section) were included as observables X_i; the X_is were specific to each column, and numbered from $k=57$ to $k=160$, depending on developmental stage, with later stages having more factors.

We defined potential functions for cliques of size up to 2: $\Phi_1(Y_i, \mathbf{X})=\lambda_1 \log P(\mathbf{X}|Y_i)$, and $\Phi_2(Y_i, Y_j, \mathbf{X})=\lambda_2 \log P(Y_i, Y_j)$, where $P(Y_i, Y_j)$ is a multinomial and $P(\mathbf{X}|Y_i)$ is a multivariate Gaussian with diagonal covariance, both trained by simple counts (maximum likelihood) from the training partition during cross validation (see Results). Coefficients λ_1 and λ_2 were estimated by maximizing the training-partition classification accuracy via simple hill climbing. Φ_1 we refer to as the *node potential*, as it is associated with single-node cliques, and Φ_2 we refer to as the *edge potential*, as it is associated with two-node cliques (individual edges in the graph).

3 RESULTS

In this section, we describe the application of a sparse BFA-CRF framework for automatic time-course gene expression pattern annotation. Our procedure starts by extracting sparse meaningful features (sBFA) from the entire collection of *Drosophila* embryos, suitable for downstream temporal analysis based on a conditional-random-fields approach. We then use the estimated factor loadings as observed variables in the CRF framework, so as to infer most likely annotation terms for previously unseen images.

3.1 Factor inference/decomposition of expression patterns

The BDGP collection divides early embryogenesis of *Drosophila* into six developmental stage ranges, 1–3, 4–6, 7–8, 9–10, 11–12, 13–16, and most of the CV terms are stage-range specific. As mentioned previously, we skipped stage range 1–3 owing to lack of informative images, as well as a low number of annotation terms associated to it. We applied the sBFA model to the entire set of images from the five stage ranges of interest. These spanned thousands of images (Table 1), with each stage being annotated by a set of 40–150 annotation terms. To illustrate the potential of the sBFA for decomposing expression patterns into meaningful features, we show selective estimated factors from developmental stages 9–10. The model began with the set of 5929 embryo images and estimated the loadings and factor matrices while having full control of the degree of sparsity (throughout our analysis, the sparsity on the factor loading matrix $\mathbf{A}$ was controlled by a scale parameter of the Gamma prior distribution on the precision parameter α, $h_0=10^{-6}$). Figure 3 illustrates a selection of the estimated factors that, per ensemble, correspond to lateral, dorsal and ventral views, demonstrating the ability of the model to automatically extract distinct patterns for different embryo orientations. As mentioned in 'Materials and Methods' section, the number of factors was determined in an unsupervised fashion, for every developmental stage range, using the sparse Bayesian infinite factor model. The estimated number of factors can be found in Table 2; in addition, we compared these findings with an empirically determined estimation akin to Pruteanu-Malinici *et al*., 2011. As illustrated in Table 2, the range of factors is similar for both scenarios: fully unsupervised (infinite factor models) or estimated by underlying

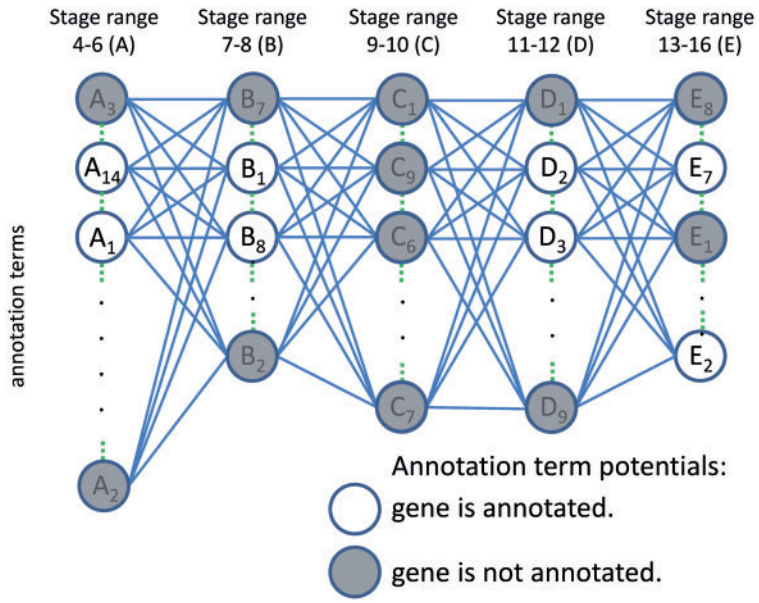


Fig. 2. CRF framework used for the automated annotation of time-course *Drosophila* embryo ISH images. Nodes correspond to annotation terms, and edges denote relationships. The order of the annotation terms within a given stage range was determined using a standard traveling–salesman heuristic in Matlab

Table 1. Statistics of the images from the BDGP database before and after the filtering process

BDGP image statistics	Stage range 4–6	Stage range 7–8	Stage range 9–10	Stage range 11–12	Stage range 13–16
Original number of images	9484	5744	5929	13 737	19 784
Number of images after filtering process	8722	5227	5523	13 245	19 269
Number of images shared across 1807 genes in common	6610	4615	4468	9315	11 111
Number of annotation terms	14	8	9	9	8

Note: The annotation terms represent a fraction of the total CV; for any given stage range, they cover ~85% of the total number of genes of interest, being frequent enough to show statistical significance.

biological assumptions (generate and test). A convergence check on the estimated number of factors for two randomly chosen stage ranges is illustrated in Figure 4. Similar to the sBFA analysis, the Bayesian infinite factor model was run for 6000 Gibbs iterations, discarding the first 1000 and estimating the model parameters on the remaining 5000 iterations.

The feature extraction/selection process was followed by filtering non-informative (such as ubiquitous) gene expression patterns. Using Euclidean distances between estimated sparse factor analysis weights and a null vector as reference, we separated informative images from the non-informative ones (for a full description see Pruteanu-Malinici *et al*., 2011). We successfully removed 2.6–9% non-informative images (Table 1).

3.2 Large-scale annotation of time-course expression patterns

In evaluating the performance of the sBFA-CRF framework, we used the estimated sparse loadings/features only on the set of genes in common between all five stage ranges of interest and a repertoire of annotation terms from a CV. The most popular annotation terms were independently selected for each stage range, to cover ~85% of the entire set of genes. This resulted in a set of 1807 images and 48 annotation terms distributed as follows: 14 terms for stage range 4–6, 8 terms for stage range 7–8, 9 terms for stage range 9–10, 9 terms for stage range 11–12 and 8 terms for the last stage range 13–16 (Table 1).

To assess the relative utility of various parts of our model, we determined the prediction accuracy of the full model compared with versions of the model handicapped in various ways. In particular, we considered including (in separate experiments) the following sets of edges in the CRF:

- Relationships across adjacent stage ranges and within stage ranges (between annotation terms): blue and green lines in Figure 2 (full model, scenario A).
- Relationships across adjacent stage ranges only: blue lines in Figure 2 (scenario B).
- Relationships within stage ranges only: green lines in Figure 2 (baseline, scenario C).

For example, when images are annotated for individual stage ranges in isolation, the relationships indicated by the edges between adjacent stages are ignored. Two examples of *Drosophila* expression pattern images across time are shown in Figure 5: we are interested in modeling the edge potentials between developmental stage ranges, as well as those between annotation terms within each stage range.

Annotation accuracy was computed as a global measure, across all annotation terms and stage ranges, by comparing the ground truth (human curated labels) with the sBFA-CRF predictions. As previous methods had largely focused on the annotation of individual stage ranges, one term at a time, and are largely trained on heavily curated benchmark data rather than the whole BDGP dataset, a fair comparison to these approaches was not feasible. To put our approach into context, we therefore compared the results generated by the sBFA-CRF framework with our own recent binary Support Vector Machines (SVM)-based classification system described in Pruteanu-Malinici *et al.* (2011). We had previously shown that this approach provides competitive and often superior classification results compared with the best competing approaches, even when working with the full BDGP image data instead of 'cleaned' benchmarks.

Our previous method used independent SVM classifiers for each annotation term and stage range, disregarding relationships within and between adjacent stage ranges. This resulted in lower annotation accuracy, as shown in Table 3. The SVM results are comparable with the new CRF baseline scenario, which only considers edge potentials between annotation terms within the same stage range (both models simply ignore any temporal/transition information that might improve the overall accuracy). The advantage of using the edge potentials between adjacent stage ranges translates into an absolute increase of 3–4% in accuracy (i.e. a relative reduction of the error rate of >20%). All models were applied to the same set of 1807 images, using 10-fold cross validation; mean values across five runs are shown.

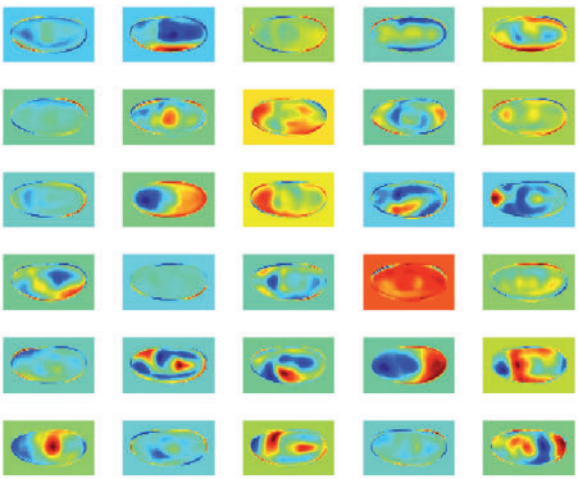


Fig. 3. Selected factors, estimated from a total of $k = 80$ factors and a grid size of 80×40 (developmental stage range 9–10). Different background colors are an artifact and not part of the model

3.3 Missing-data annotation analysis

In addition to improved gene annotation accuracy, the sBFA-CRF framework provides an elegant means of annotating images in missing-data scenarios. During CRF decoding, the most likely configuration of the model (i.e. values of the unobservables, **Y**) is computed using relationships between adjacent stage ranges, as well as within each stage range. In the case of missing data, the most likely state for a given node Y_i with no directly related observables **X** is estimated entirely through relationships in the random field. This allows us to infer annotation terms for missing images, which is of utmost importance in scenarios where, for a given gene, data have been collected for only a subset of the stage ranges.

In evaluating the performance of the sBFA-CRF model in annotating missing data, we manually removed 5–25% of images from the 1807 gene set, by randomly selecting genes and removing their corresponding images; for missing images, the node potentials were set to 1. Results are shown in Figure 6, where the model included edge potentials between adjacent stage ranges, as well as within stage ranges (CRF scenario A); we were able to annotate with 80% or better accuracy even for scenarios with 25% missing data. Remarkably, our model outperforms the SVM classification framework (which had access to full data) even when 15% of the data are withheld from the sBFA-CRF. Figure 7 illustrates particular cases with genes that are correctly annotated, for stages where their images were missing. As previously mentioned, this is of particular interest to biologists who require predictions for stages where images have not yet been collected. Our results confirm that the proposed time-course pipeline leads to highly successful expression pattern classification, despite the presence of uninformative images, registration errors and missing data in considerable amounts.

Lastly, we compared the sBFA-CRF predicted labels to the human curated ones (the ground truth), so as to identify genes and annotation terms for which the annotations were different but the outcome from our model appeared consistent. We

Table 2. Comparison of the number of estimated factors in the BDGP set

Method	Stage range 4–6	Stage range 7–8	Stage range 9–10	Stage range 11–12	Stage range 13–16
Generate and test	k = 60	k = 100	k = 100	k = 150	k = 150
Infinite factor models	k = 57	k = 60	k = 80	k = 140	k = 160

Note: First row corresponds to number selection based on biological prior knowledge followed by generate-and-test procedures. Second row shows the estimated number of factors, fully unsupervised (the infinite BFA).

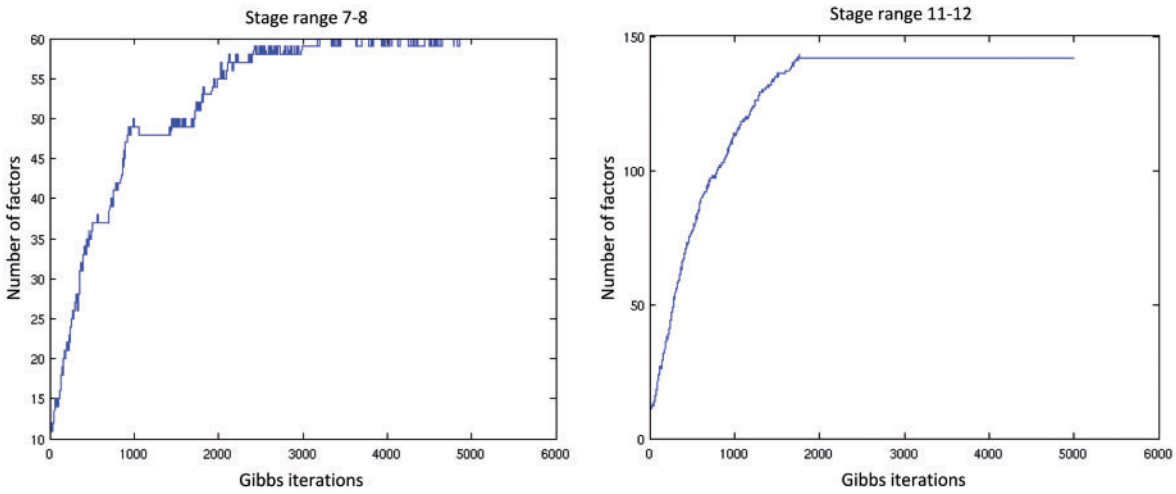


Fig. 4. Convergence of the estimated number of factors for two developmental stage ranges (7–8 and 11–12): 5000 Gibbs iterations

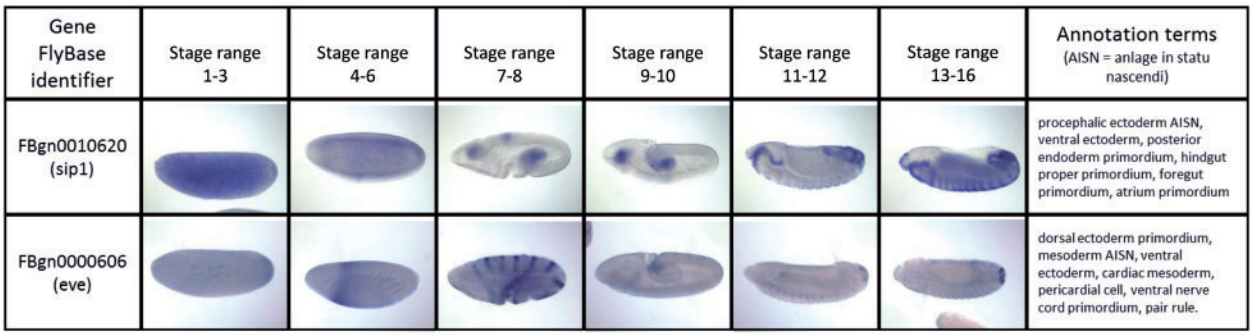


Fig. 5. *Drosophila* embryonic gene expression across six stage ranges. Images can display different embryo orientations due to the relatively long developmental time spanned by a stage range. Using the edge potentials between adjacent stage ranges, as well as within stage ranges, translates into a significant increase in annotation accuracy

Table 3. Summary of annotation accuracy

Annotation accuracy analysis	CRF scenario (A)	CRF scenario (B)	CRF scenario (C)	SVM
Mean annotation accuracy	86.75	85.69	82.93	83.32

Note: sBFA-CRF and SVM models: mean annotation accuracy, over 5 N-fold cross validation runs (N = 10). Scenario (**A**) includes relationships between adjacent stage ranges and within stage ranges; scenario (**B**) considers only relationships between adjacent stage ranges; scenario (**C**) models only relationships within stage ranges (baseline). For the SVM model, we used independent classifiers for each annotation term and stage range.

recognize that for the same annotation term, the corresponding regions in different images may have significant variations in visual appearances, which would lead to a difficult manual annotation task and could sometime generate ambiguous outcomes. We show three examples for which the sBFA-CRF annotations are opposite from the human curated ones and are likely correct given the full context (Fig. 8). For all three scenarios, we confirmed our findings with the BDGP human curators. Arguably, the model was correct in predicting different annotation terms in the following examples, including gene FBgn0003502, where human curators initially decided that expression in 'procephalic ectoderm AISN' is not detected for stage range 4–6; however, the sBFA-CRF predicted label, as well as a second careful visual inspection, would suggest the contrary. In addition, we identified another instance where 'ventral ectoderm anlage' should have been annotated for gene FBgn0022073 in stage range 7–8. The last scenario (gene FBgn0033988) corresponds to a case where all images are extremely difficult to annotate owing to out-of-focus staining issues or overall noise. On a second inspection, the model was arguably correct in labeling the annotations for both stage ranges 4–6 and 9–10.

4 CONCLUSIONS

We have described a novel sBFA-CRF model to automatically annotate *Drosophila* embryo gene-expression time-course data.

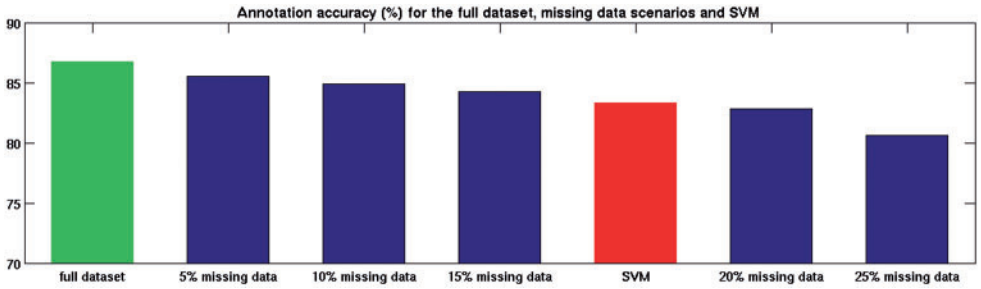


Fig. 6. Annotation accuracy results for missing data scenarios. The accuracy values were computed as global measures, across the entire set of 1807 genes. For each case, we randomly selected 5–25% of the complete gene set and removed their corresponding images, so as to simulate missing data scenarios. The green bar indicates the annotation accuracy for the full dataset scenario (previous analysis); the red bar corresponds to the SVM analysis

Gene FlyBase identifier	Stage range 4-6	Stage range 7-8	Stage range 9-10	Stage range 11-12	Stage range 13-16	Percentage of correctly annotated terms in the missing stage range	Correctly annotated terms (selection) (AISN = anlage in statu nascendi)	Incorrectly annotated terms (AISN = anlage in statu nascendi)
FBgn0015773 (NetA)						78.57	dorsal ectoderm AISN ventral ectoderm AISN procephalic ectoderm AISN posterior endoderm AISN mesoderm anlage AISN	cellular blastoderm anterior endoderm AISN trunk mesoderm AISN
FBgn0023533 (CG13372)						100	procephalic ectoderm primordium posterior endoderm primordium ventral nerve cord primordium	N/A
FBgn0015268 (Nap1)						77.77	ventral nerve cord primordium 2 brain primordium dorsal epidermis primordium	ventral epidermis primordium trunk mesoderm primordium
FBgn24150 (Ac78C)						87.5	ventral ectoderm primordium procephalic ectoderm anlage trunk mesoderm primordium head mesoderm primordium	dorsal ectoderm primordium
FBgn0004606 (zfh1)						100	ventral nerve cord embryonic dorsal epidermis embryonic ventral epidermis embryonic midgut	N/A

Fig. 7. Missing-data gene annotation analysis. Shaded boxes indicate the stage range for which data were missing (we manually removed those images). In two cases, the entire set of annotation terms is correctly annotated within the stage range despite the fact that no images were available (third and sixth rows). In these two scenarios, the CRF used the relationships across adjacent stages, to estimate the most likely configuration. A selection of the correctly and incorrectly annotated terms for the stage range with missing data are shown in the last two columns

The sparse BFA framework represents an efficient feature selection technique, which automatically determines the feature-space dimension, using a non-parametric implementation. The learned features are then used as observed variables in the CRF framework, so as to infer most likely annotation terms for previously unseen images. The CRF encodes temporal relationships between adjacent stage ranges throughout *Drosophila* development. By capturing the temporal sequence, the model is able to predict the entire collection of annotation terms in a single run and achieves superior performance when compared with highly competitive models that annotate stages in isolation. In addition to improved annotation accuracy, the experimental results demonstrate the success of the method in handling missing-data scenarios. This is extremely useful in real-life scenarios when estimates are needed over the ensemble of annotation terms, with only partial data being collected.

One promising extension to our approach would be to include 'latent' annotation terms in the CRF structure, to account for additional rare annotation terms for which we would not attempt to obtain a prediction. These latent terms could have limited connectivity in the graph, so as to allow large numbers of latencies to be included without compromising decoding efficiency. Such an extension may well improve prediction accuracy for the primary terms, even if the latent terms are themselves difficult to accurately predict (due to paucity of training data for those terms). It would also increase the flexibility of the resulting model: while we currently select the primary annotation terms manually based on their popularity among genes in a given stage range, a simple threshold on the number of genes being annotated, together with an appropriate means of ranking terms, would allow to automatically partition the primary versus latent sets. Based on our experience with the BFA-CRF model described here, additional work along these lines seems promising.

Finally, we are continuing to develop this approach in close collaboration with biologists so as to suggest outliers or interesting patterns during the anticipated expansion of the BDGP collection to the whole *Drosophila* genome. We plan to incorporate the sBFA-CRF framework into a Fiji plug in (Schindelin *et al.*,

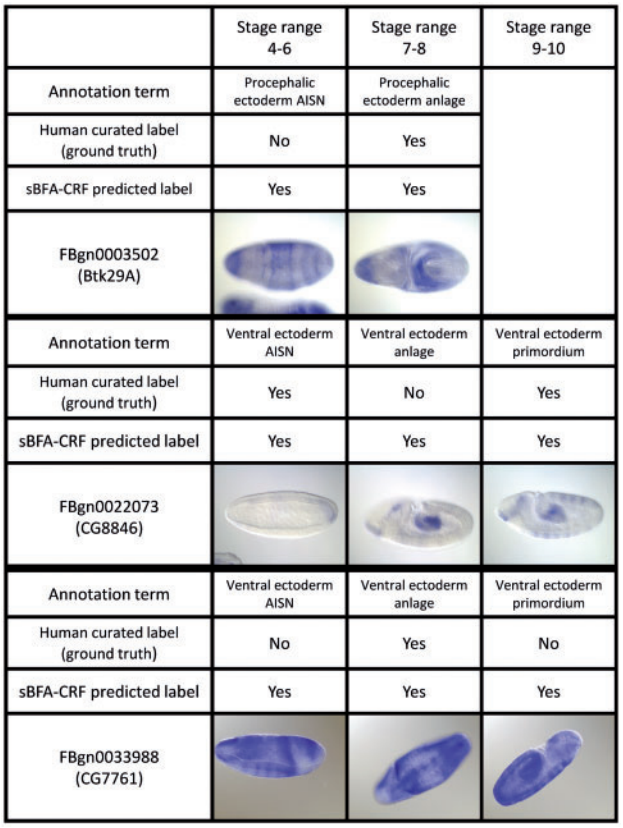


Fig. 8. Examples of genes/annotation terms for which the sBFA-CRF predictions differ from the human curated ones. Yes—gene is annotated; No—gene is not annotated. Note the consistency of model predictions within the context of annotations for neighboring stage ranges

2012)—a gene annotation tool that would accurately assign annotation terms to new/unseen images, in a timely manner.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Erwin Frise for his valuable help in data collection, as well as helpful comments throughout the model development and testing.

Funding: National Science Foundation CAREER award [DBI-0953184 to U.O.]

Conflict of Interest: none declared.

REFERENCES

Bhattacharya,A. and Dunson,D.B. (2011) Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291–306.

Busch,W. *et al.* (2012) A microfluidic device and computational platform for high-throughput live imaging of gene expression. *Nat. Methods*, **9**, 1101–1106.

Campos-Ortega,J.A. and Hartenstein,V. (1985) *The Embryonic Development of Drosophila Melanogaster*. Berlin, Springer-Verlag.

Frise,E. *et al.* (2010) Systematic image-driven analysis of the spatial Drosophila embryonic expression landscape. *Mol. Syst. Biol.*, **6**, 345.

Fowlkes,C.C. *et al.* (2005) Registering Drosophila embryos at cellular resolution to build quantitative 3D atlas of gene expression patterns and morphology. In: *Proceedings of the IEEE Computational Systems Bioinformatics Conference Workshops* (CSBW'05). Stanford, CA.

Fowlkes,C.C. *et al.* (2008) A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell*, **133**, 364–374.

Gorsuch,R.L. (1983) *Factor Analysis*. Lawrence Erlbaum Associates, Hillsdale, NJ (USA).

Hammersley,J.M. and Clifford,P. (1971) *Markov Fields on Finite Graphs and Lattices*. PHS Grant no. GM-10525-08, National Institute of Health, Public Health Service.

Harmon,C. *et al.* (2007) Comparative analysis of spatial patterns of gene expression in Drosophila melanogaster imaginal discs. *Res. Comput. Mol. Biol.*, **4453**, 533–547.

Jensen,F.V. *et al.* (1990) Bayesian updating in causal probabilistic networks by local computations. *Comput. Stat.*, **4**, 269–282.

Ji,S. *et al.* (2008) Automated annotation of Drosophila gene expression patterns using a controlled vocabulary. *Bioinformatics*, **24**, 1881–1888.

Ji,S. *et al.* (2009) A bag-of-words approach for Drosophila gene expression pattern annotation. *BMC Bioinformatics*, **10**, 119.

Keranen,S.V. *et al.* (2006) 3D morphology and gene expression in the Drosophila blastoderm at cellular resolution II: dynamics. *Genome Biol.*, **7**, R124.

Kumar,S. *et al.* (2002) BEST: a novel computational approach for comparing gene expression patterns from early stages of Drosophila melanogaster development. *Genetics*, **162**, 2037–2047.

Lafferty,J.D. *et al.* (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pp. 282–289. Williamstown, MA (USA).

Lauritzen,S.L. and Spiegelhalter,D.J. (1988) Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc. Series B Stat. Methodol.*, **50**, 157–224.

Ljosa,V. and Carpenter,A.E. (2009) Introduction to the quantitative analysis of two-dimensional fluorescence microscopy images for cell-based screening. *PLoS Comput. Biol.*, **5**, e1000603.

Mace,D.L. *et al.* (2010) Extraction and comparison of gene expression patterns from 2D RNA in situ hybridization images. *Bioinformatics*, **26**, 761–769.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. 2nd edn. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Peng,H. and Myers,E.W. (2004) Comparing in situ mRNA expression patterns of Drosophila embryos. In: *Proceedings of the 8th Conference on Research in Computational Molecular Biology*, pp. 157–166. San Diego, CA (USA).

Peng,H. *et al.* (2007) Automatic image analysis for gene expression patterns of fly embryos. *BMC Cell Biol.*, **8**, S7.

Prince,S.J. *et al.* (2008) Tied factor analysis for face recognition across large pose differences. *IEEE Trans. Pattern Anal. Mach. Intell.*, **30**, 970–984.

Pruteanu-Malinici,I. *et al.* (2011) Automatic annotation of spatial expression patterns via sparse Bayesian factor models. *PLoS Comput. Biol.*, **7**, e1002098.

Puniyani,K. *et al.* (2010) SPEX^2: automated concise extraction of spatial gene expression patterns from Fly embryo ISH images. *Bioinformatics*, **26**, i47–i56.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **7**, 257–286.

Roberts,G. and Rosenthal,J. (2007) Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.*, **44**, 458–475.

Schindelin,J. *et al.* (2012) Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, **9**, 676–682.

Tautz,D. and Pfeifle,C. (1989) A non-radioactive in situ hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene hunchback. *Chromosoma*, **98**, 81–85.

The FlyBase Consortium. (2002) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.*, **30**, 106–108.

Tomancak,P. *et al.* (2002) Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.*, **3**, 88.

Tomancak,P. *et al.* (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.*, **8**, R145.

Walter,T. *et al.* (2010) Visualization of image data from cells to organisms. *Nat. Methods*, **7**, 26–41.