Temporal ordering and registration of images in studies of developmental dynamics

Carmeline J. Dsilva [*], Bomyi Lim [*], Hang Lu [†], Amit Singer [‡§], Ioannis G. Kevrekidis [*§], and Stanislav Y. Shvartsman [*¶]

Abstract

Dynamics of developmental progress is commonly reconstructed from imaging snapshots of chemical or mechanical processes in fixed embryos. As a first step in these reconstructions, snapshots must be spatially registered and ordered in time. Currently, image registration and ordering is often done manually, requiring a significant amount of expertise with a specific system. However, as the sizes of imaging data sets grow, these tasks become increasingly difficult, especially when the images are noisy and the examined developmental changes are subtle. To address these challenges, we present an automated approach to simultaneously register and temporally order imaging data sets. The approach is based on vector diffusion maps, a manifold learning technique that does not require *a priori* knowledge of image features or a parametric model of the developmental dynamics. We illustrate this approach by registering and ordering data from imaging studies of pattern formation and morphogenesis in three different model systems. We also provide software to aid in the application of our methodology to other experimental data sets.

KEY WORDS: temporal ordering, image registration, vector diffusion maps

Introduction

In one of the common approaches to studies of developmental dynamics, a group of embryos is fixed and stained to visualize a particular biochemical or morphological process within a developing tissue. The developmental dynamics must then be reconstructed from multiple embryos, each of which contributes only a snapshot of the relevant process along its developmental trajectory (Jaeger et al., 2004; Peter and Davidson, 2011; Fowlkes et al., 2008). Importantly, the "age" of any given embryo arrested in its development is often only approximately known. Typically, what is known is a certain time window to which a collection of embryos belongs (Ng et al., 2012; Richardson et al., 2014; Castro et al., 2009). Furthermore, images are often collected in different spatial orientations. In order to recover the developmental dynamics from such data sets, snapshots of different embryos must first be spatially aligned or *registered*, and then ordered in time.

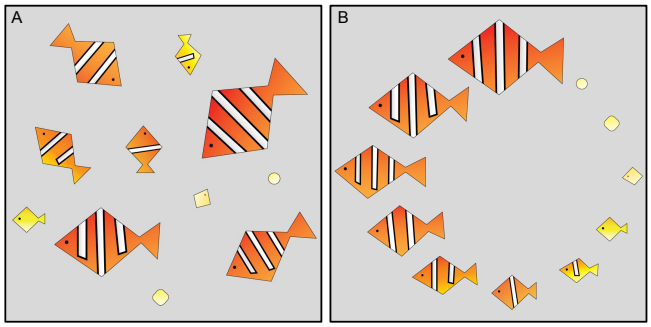


Fig. 1: Caricature illustrating the tasks of image registration and temporal ordering. (A) Images of "samples", each in a different orientation and a different stage of development. (B) Registered and ordered samples. For this caricature, the registration and ordering is straightforward because the data set is small, the landmarks are visually apparent, and the developmental changes are easy to recognize.

Temporal ordering and registration of images can be done manually when the number of images is small and the differences between them are visually apparent. Fig. 1 shows a caricature of fish development which illustrates the processes of growth and patterning. In this case, temporal ordering can be accomplished by arranging the fish by size, which is monotonic with the developmental progress. Image registration is based on obvious morphological landmarks, such as the positions of the head and the fins. In contrast to this example, real data pose nontrivial challenges for both registration and temporal ordering. In general, the landmarks needed for registration, as well as the attributes which

[*]Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey, USA
[†]School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA
[‡]Department of Mathematics, Princeton University, Princeton, New Jersey, USA
[§]Program in Applied and Computational Mathematics, Princeton University, Princeton, New Jersey, USA
[¶]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA

can be used to order the data, are not known *a priori*. Additional challenges arise from embryo-to-embryo variability, sample size, and measurement noise.

We present a robust algorithmic approach to simultaneous registration and temporal ordering. In contrast to a number of previous methodologies (Zitova and Flusser, 2003; Rowley et al., 1998; Hajnal and Hill, 2010; Greenspan et al., 1994; Zhao et al., 2003; Dubuis et al., 2013), our methodology does not rely on the *a priori* knowledge of landmarks for registration or markers of developmental progression. The approach is based on vector diffusion maps (Singer and Wu, 2012), a manifold learning algorithm which simultaneously addresses the problems of registration and temporal ordering. This algorithm is one of several nonlinear dimensionality reduction techniques that have been developed over the past decade (Belkin and Niyogi, 2003; Coifman et al., 2005; Coifman and Lafon, 2006; Tenenbaum et al., 2000; Roweis and Saul, 2000), for applications ranging from analysis of cryo-electron microscopy (cryo-EM) images of individual molecules (Zhao and Singer, 2014; Singer et al., 2011) to face recognition (Lafon et al., 2006) and classification of CT scans (Fernández et al., 2014).

Here, the vector diffusion maps algorithm is adapted for the analysis of images of developing tissues in studies of developmental dynamics, with the main objective of revealing stereotypic developmental trajectories from fixed images. To illustrate our approach, we analyze four experimental data sets. Our first two data sets come from live imaging studies of *Drosophila* and zebrafish embryogenesis. In both of these examples, the correct rotational orientation and temporal order are independently known, and these data sets will be used to validate our approach. Our third data set consists of images from fixed *Drosophila* embryos where the correct orientation and order is unknown; here, we will show how the algorithm can help uncover developmental dynamics which are not readily apparent. Our final data set consists of z-stacks of *Drosophila* wing discs, which we will use to illustrate how our methods can be used to analyze specific types of three-dimensional imaging data. We also show how to compute an average trajectory from a set of registered and ordered fixed images to remove noise due to intersample variability and obtain a smooth description of the underlying developmental dynamics.

Results

Vector diffusion maps for registration and temporal ordering

Vector diffusion maps (Singer and Wu, 2012) is a manifold learning technique developed for data sets which contain two sources of variability: geometric symmetries, such as rotations of the images, which one would like to factor out, and "additional" directions of variability, such as temporal dynamics, which one would like to uncover. Vector diffusion maps combine two algorithms, *angular synchronization* (Singer, 2011) for image registration and *diffusion maps* (Coifman et al., 2005) for extracting intrinsic low-dimensional structure in data, into a single computation. We will use the algorithm to register images of developing tissues with respect to planar rotations, as well as uncover the main direction of variability *after* removing rotational symmetries. Although in general, images may contain variations due to rotations, translations, and scaling, we will remove the relevant translations and/or scaling via relatively simple image preprocessing, and focus only on factoring out rotations using the vector diffusion maps algorithm. In the case that *all* relevant symmetries can be removed with straightforward preprocessing, our algorithms can extract the main direction of variability within the imaging data set. We assume that this main direction of remaining variability in these images is parameterized by the developmental time of each embryo. As a consequence, uncovering this direction should reveal the underlying dynamics.

Angular synchronization uses pairwise alignment information to register a set of images in a globally consistent way. A schematic illustration of angular synchronization is shown in Fig. 2A, where each image is represented as a vector, and the goal is to align the entire set of vectors given pairwise alignment measurements. We first compute the angles needed to align pairs of vectors (or images), which in general requires no notion of a template function (Ahuja et al., 2007; Sonday et al., 2013). In this work, we aligned pairs of images with respect to rotations by exhaustively searching over a discretized space of rotation angles to minimize the Euclidean distance between the pixels. However, pairwise alignments can also be computed by aligning appropriate image landmarks or features (Dryden and Mardia, 1998). When the data are noisy, these pairwise measurements may by inaccurate, and so we utilize *all* pairwise measurements to align the set of images robustly. Using the alignment angles between all pairs of vectors, angular synchronization finds the set of rotation angles (one angle for each vector) that is most consistent with *all* pairwise measurements (see supplementary material); this is illustrated in Fig. 2B. In this schematic, registration via angular synchronization is trivial, as the pairwise measurements contain no noise. However, the algorithm can register data sets even when many of the pairwise measurements are inaccurate (Singer, 2011).

After removing variability due to rotations, the developmental dynamics may be revealed by ordering the data along the one-dimensional curve that parameterizes most of the remaining variability in the data. Such a curve can be discovered using diffusion maps (Coifman et al., 2005), a nonlinear dimensionality reduction technique that reveals a parametrization of data that lies on a low-dimensional manifold in high-dimensional

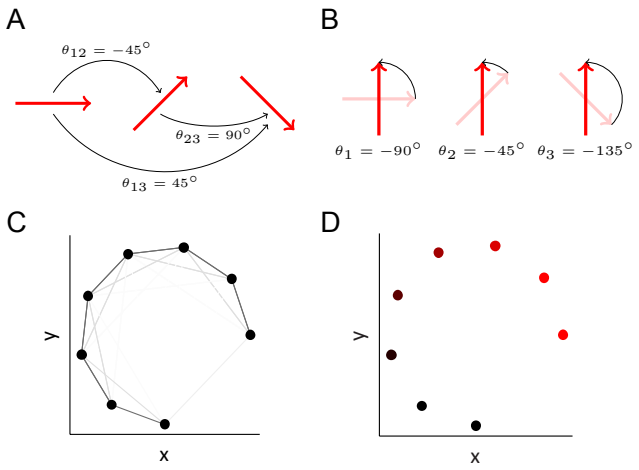


Fig. 2: Schematic illustrating angular synchronization and diffusion maps. (A) Set of vectors, each in a different orientation. The pairwise alignment angles are indicated. (B) The vectors from A, each rotated about their midpoint so that the set is globally aligned. Note that the chosen rotation angles are consistent with the pairwise alignments in A: the difference between a pair of angles in B is the same as the pairwise angle in A. (C) Data points (in black) which lie on a one-dimensional nonlinear curve in two dimensions. Each pair of points is connected by an edge, and the edge weight is related to the Euclidean distance between the points through a Gaussian kernel (see supplementary material), so that pairs of data point which are close are connected by darker ("stronger") edges. (D) The data in C, colored by the first (non-trivial) eigenvector from the diffusion map computational procedure. The color intensity is monotonic with the perceived curve arclength, thus parametrizing the curve.

space. The idea is illustrated in Fig. 2C, where the data are two-dimensional points which lie on a one-dimensional (nonlinear) curve. We use *local* information about the data to find a parametrization which respects the underlying manifold geometry, so that points which are close in high-dimensional space (e.g., images which look similar) are close in our parametrization. This idea of locality is denoted by the color of the edges in Fig. 2C: data points which are close are connected by dark edges, and clearly, the dark edges are more "informative" about the low-dimensional structure of the data. The color in Fig. 2D depicts the one-dimensional parametrization or ordering of the data that we can detect visually. A detailed example of using vector diffusion maps to register and order synthetic data is given in Fig. S2, and a step-by-step tutorial of the diffusion maps implementation is included in the supplementary material. In our working examples, each data point will be of much higher dimension (e.g., a pixelated image or three-dimensional voxel data), and so we cannot extract this low-dimensional structure visually. Instead, we will use diffusion maps which automatically uncovers a parametrization of our high-dimensional data from the eigenvectors of the appropriate matrix (see supplementary material). Furthermore, the corresponding eigenvalues will allow us to test our assumption that our data approximately lie on a one-dimensional manifold (see Fig. S3–S6).

Method validation using live imaging

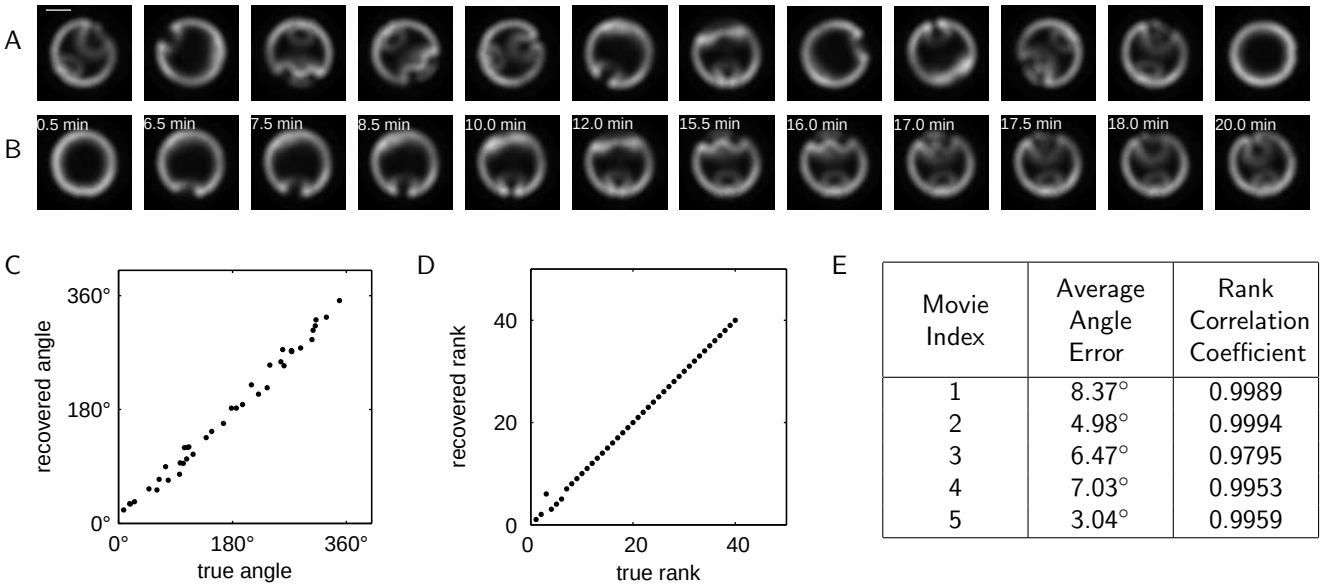
Drosophila gastrulation

To validate the proposed approach, we first applied our algorithm to a data set where the true temporal order and rotational orientation of the images were known *a priori*. This data set was obtained through live imaging near the posterior pole of a vertically oriented *Drosophila* embryo during the twenty minutes spanning the late stages of cellularization through early gastrulation. During this time window, the ventral furrow is formed, where the ventral side buckles towards the center of the embryo, internalizing the future muscle cells and forming a characteristic "omega" shape. Germband extension then causes cells from the ventral side to move towards the posterior pole of the embryo, and then wrap around to the dorsal side (Leptin, 2005). At the end of this process, cells which were originally on the ventral and posterior side of the embryo find themselves on the dorsal side, causing a similar "omega" to appear on the dorsal side.

Fig. 3A shows selected images from this live imaging data set, which contains 40 consecutive frames taken at 30 second time intervals at a fixed position within a single embryo. Each image shows an optical cross-section near the posterior pole of a vertically oriented developing embryo, with the nuclei labeled by Histone-RFP. Each frame was arbitrarily rotated, and the order of the frames was scrambled. The task is now to register these images and order them in time to reconstruct the developmental trajectory.

We used vector diffusion maps to register and order the images. Fig. 3B shows the images from Fig. 3A, now registered and ordered; the real time for each frame is also indicated. With a small number of exceptions, the recovered ordering is consistent with the real time dynamics. Fig. 3C and Fig. 3D show the correlations between the recovered and true angles and rank orders, respectively, for the entire data set. Both the angles and the ranks are recovered with a high degree of accuracy. We note that determining which end of the trajectory corresponds to early in the developmental progression is a post-processing task that requires some *a priori* information.

To assess the robustness of the proposed methodology, we repeated this procedure with four additional data sets extracted from independent live imaging studies spanning the same developmental time period. The results are shown in Fig. 3E. The errors in the recovered angles are



E

Movie Index	Average Angle Error	Rank Correlation Coefficient
1	8.37°	0.9989
2	4.98°	0.9994
3	6.47°	0.9795
4	7.03°	0.9953
5	3.04°	0.9959

Fig. 3: Method validation using live imaging of *Drosophila* embryos. (A) Selected images from a live imaging study of a *Drosophila* embryo during gastrulation. Scale bar indicates 50μm. Each frame is in an arbitrary rotational orientation, and the order of the frames has been shuffled. (B) Images from A registered and ordered by vector diffusion maps. The dorsal side of each embryo now appears at the top of each image, and the ventral side appears at the bottom. (C) The correlation between the recovered rotation angle (using vector diffusion maps) and the true rotation angle. The average absolute error in the recovered angles is 8.37°. (D) The correlation between the recovered rank (using vector diffusion maps) and the true rank. The rank correlation coefficient is 0.9989. (E) The average error in the recovered angle and the rank correlation coefficient for 5 independent live imaging studies.

all less than 10°, and the rank correlation coefficients are consistently greater than 95%, indicating that our methodology can reproducibly order data of this type.

Zebrafish epiboly

As another validation for the proposed methodology, we applied our algorithm to a time-lapse movie of zebrafish embryogenesis. We used a publicly available live imaging data of zebrafish embryogenesis (`https://zfin.org/zf_info/movies/Zebrafish.mov`, Karlstrom and Kane (1996)). Taken with a differential interference contrast (DIC) microscope, the movie records the first 17 hours of zebrafish development, from a single cell stage to a 16-somite stage. We selected 120 consecutive frames from this movie which capture 5.5 hours of epiboly (3.5–9 hours after fertilization). In this experiment, embryos were immobilized for imaging so that the position and orientation remained fixed (Kane et al., 1996). At the start of the time window, cells have divided 10–11 times and are accumulated in a cell mass above the yolk. The cell mass is then compressed and the animal-vegetal axis of the embryo (vertical axis in Fig. 4) shortens to form a spherical embryo shape by the end of the fourth hour of development. Then, the yolk syncytial layer, which forms the boundary between the yolk and the cell mass, moves upward, forming a dome-shaped structure. During this stage, the cells rearrange to form a uniform layer about four cells thick. With time, this cell layer then spreads over across the yolk and expands toward the vegetal pole. At the end of epiboly, the blastoderm completely engulfs the yolk.

As in the example of *Drosophila* embryo live imaging, the two-dimensional frames were randomly rotated and shuffled (Fig. 4A). We then used vector diffusion maps to register and order the frames. The results are shown in Fig. 4B. The recovered rotations and order are consistent with the expected developmental dynamics, as shown in the correlations between the recovered and true ranks (Fig. 4C). Quantitatively, the rank correlation coefficient for this data set is 0.9954, and the average error in the recovered angle is 4.14°. Some errors in ordering images of the early embryo result from slow cell movement during the early developmental stage where cells divide and accumulate above the yolk. During epiboly, cell movement is more dynamic and the recovered ordering is more consistent with the real dynamics.

In summary, we have shown that our approach to temporal ordering performs very well on imaging data of two different developmental processes (*Drosophila* gastrulation and zebrafish epiboly), taken with two different imaging methods (fluorescent microscopy and DIC) where the true temporal order is known *a priori*. Provided there exist significant dynamics within the data set and that the developmental trajectory is well-sampled, the developmental dynamics can be recovered.

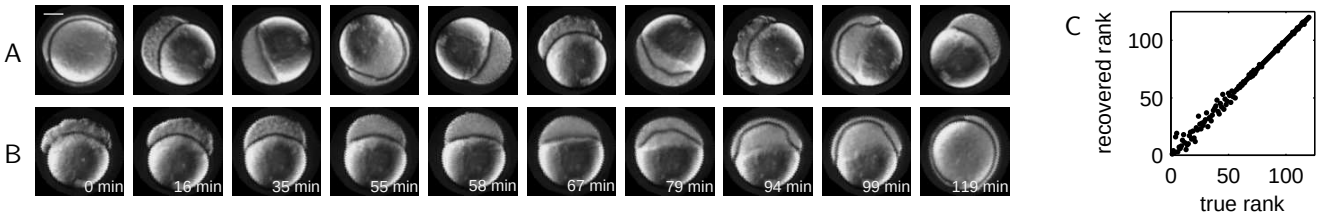


Fig. 4: Method validation using live imaging of a zebrafish embryo. (A) Selected images from a movie of zebrafish epiboly. Scale bar indicates 200μm. Each frame is in an arbitrary rotational orientation, and the order of the frames has been shuffled. (B) Images from A after registration and ordering using vector diffusion maps. The real time of each frame is also indicated. (C) Correlation between the rank recovered using vector diffusion maps and the true rank. The rank correlation coefficient is 0.9954. The larger errors in the recovered ranks towards the beginning of the trajectory are due to the slow cell movement within that time window.

Data sets with intersample variability

Fixed images of *Drosophila* gastrulation

We have analyzed how our algorithm performs on two model data sets where all images come from a single embryo. In practice, we are interested in cases where each image comes from a different embryo, and the largest source of noise in the considered data set arises from embryo-to-embryo variability. To demonstrate that our methods are robust to such variations, we constructed a synthetic time course data set by selecting a random image from one of five *Drosophila* live imaging data sets (those data sets used in Fig. 3) at each time point. The resulting data set is spatially unregistered, scrambled in time, and reflects embryo-to-embryo variability. The median rank correlation coefficient when ordering such a synthetic time course using our methodology was 0.77, indicating that the algorithm can recover the temporal order even under noisy conditions.

We then applied our approach to a data set where the true rotational orientation and temporal order was not known *a priori*. Fig. 5A shows selected images from a set of 120 images of developing *Drosophila* embryos which cover a thirty minute time interval spanning late cellularization through gastrulation. This data set is more complex than the live imaging data sets in that it contains significantly more images, each of which provides information about tissue morphology and the spatial distribution of two regulatory proteins. Each image shows an optical cross-section of the posterior view of a *different* embryo at a different rotational orientation and fixed at a different (and unknown) developmental time. The nuclei (gray) were labeled with DAPI, a DNA stain. Embryos were stained with the antibody that recognizes Twist (Twi, shown in green), a transcription factor which specifies the cells of the future muscle tissue. Another signal is provided by the phosphorylated form of the extracellular signal regulated kinase (dpERK, shown in red), an enzyme that, in this context, specifies a subset of neuronal cells (Lim et al., 2013).

Fig. 5B shows the selected images in Fig. 5A, now registered and ordered using vector diffusion maps. Registered and ordered images of individual embryos can then be used to construct a representative average trajectory. Each snapshot in the average trajectory is the (weighted) average of a group of successive images from the registered and ordered data set (see supplementary material). Averaging successive images removes some of the interembryo variability, so that sequential snapshots of this averaged trajectory, shown in Fig. 5C, serve as a summary of the stereotypic developmental dynamics.

From this average trajectory, we can now easily see the developmental progression consistent with the known dynamics: dpERK first appears as two lateral peaks at the ventrolateral side of the embryo, and a third dpERK peak then appears at the dorsal side of the embryo. During mesoderm invagination, the two ventrolateral dpERK peaks merge together, eventually forming, together with Twi, the "omega" shape. The dorsal dpERK peak then disappears during germband extension as cells from the ventral side wrap around to the dorsal side. At the end of this process, similar "omegas" formed by Twi and dpERK appear on the dorsal side of the embryo; these patterns are most readily seen in the last image of Fig. 5C. Thus, vector diffusion maps can accomplish the tasks presented in the caricature in Fig. 1, even in the absence of information about image landmarks and without *a priori* knowledge of developmental features.

To evaluate the quality of our registration and ordering, we can use prior knowledge about the developmental system. The Twi signal is known to form a single peak at the ventralmost point of the embryo. We found that the standard deviation in the location of this peak in the set of registered images was $\sim 8^\circ$, indicating that the algorithm successfully aligns the ventralmost points of the images. Because the developmental time of each embryo cannot be easily estimated, we have few options for evaluating the quality of our temporal ordering. We compared the ordering obtained from vector diffusion maps to the ordering provided by a trained embryologist who is knowledgeable about the developmental progression

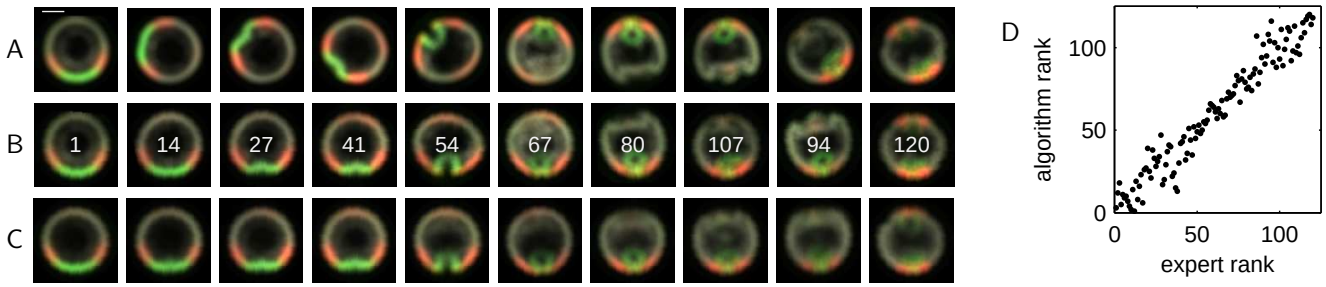


Fig. 5: Analysis of images of fixed *Drosophila* embryos. (A) Images of *Drosophila* embryos, stained for nuclei (gray), Twi (green), and dpERK (red). Scale bar indicates approximately 50μm (images have been rescaled to remove slight interembryo size variations). Each image is of a different embryo arrested at a different developmental time and in a different rotational orientation. (B) Data from A, registered and ordered using vector diffusion maps. The expert rank for each image is indicated. (C) A representative "developmental trajectory" obtained from local averaging of the entire set of registered and ordered images (see supplementary material). (D) Correlation between the image ranks calculated from the vector diffusion maps algorithm and the ranks obtained from ordering by an expert. The rank correlation coefficient is 0.9716.

and the important image features. The ranks from the ordering provided by the embryologist, which we will refer to as the "expert rank", are indicated for the images in Fig. 5B , and the rank correlation (see Fig. 5D) shows that our ordering is consistent with the expert ordering.

Fixed z-stacks of *Drosophila* wing discs

In this section we show that the approach can readily be applied to three-dimensional data. We restrict ourselves to the case where an obvious fixed axis exists, so that only rotations of the three-dimensional data around this axis need be taken into account. This does not constitute an inherent limitation for vector diffusion maps. While for simplicity here we will not discuss the general case, incorporating general 3D symmetries is possible (Arie-Nachimson et al., 2012; Wang and Singer, 2013; Cucuringu et al., 2012).

To demonstrate this approach, we used an existing three-dimensional data set of fixed *Drosophila* wing imaginal discs (Hamaratoglu et al., 2011). Imaginal discs are groups of progenitor cells in fly larva that will transform into specific organs during metamorphosis. The wing disc is an imaginal disc that turns into a wing. The data set is composed of 46 fixed wing discs whose developmental times range from 72 to 112 hours after fertilization. Each disc contains 21 z-slices taken at 1 μm intervals. The discs were dissected from larvae expressing the Dad-GFP reporter construct (green) and stained with antibodies that recognize Spalt (red), Wingless (gray), and Patched (gray), the factors that play important roles in disc patterning and growth (Fig. 6A).

In the wing disc, the anterior-posterior and dorsal-ventral axes are significantly larger than the third principal axis (see Fig. 6A). Therefore, we need not consider registration in all three dimensions, and can instead focus on registering the wing discs with respect to rotations only in the x-y plane. To register the data, we first aligned the maximum intensity projections using angular synchronization. We then used these rotations to register the full three-dimensional data in the x-y plane. Because the maximum intensity projections are two-dimensional images, this step is no more computationally intensive than the previous examples. Such an approach is possible when there are distinct major and minor axes within a three-dimensional sample, which reduces the rotational degrees of freedom.

We then used diffusion maps to order the registered three-dimensional data. Fig. 6B shows selected images from the data set ordered by diffusion maps. In the original data set, each disc was assigned to one of six time classes (72–73 hr, 76.5–77.5 hr, 79–80 hr, 89–90 hr, 100–101 hr, and 110.5–111.5 hr after fertilization) by an expert; these times are indicated in Fig. 6B. In the ordered set, the size of wing disc grows, and the intensity of the Dad-GFP signal increases as a function of time. The rank correlation coefficient based on the time class is 0.9436. The registration errors are primarily due to some wing discs having extra tissue attached to them (such as the image in Fig. 6A and the fourth image in Fig. 6B). Even with such obstructions, we can accurately order the images and extract a stereotypical developmental trajectory, shown in Fig. 6C, by averaging (see supplementary material). We can now clearly see the growth of the wing disc, even though averaging somewhat blurs some finer scale structures.

Computational requirements

The computational costs for our methodology are outlined in Fig. 7. The computational time is a function of the number of images in the data set, the number of pixels in each point, and the angular resolution to compute the pairwise rotations (see supplementary

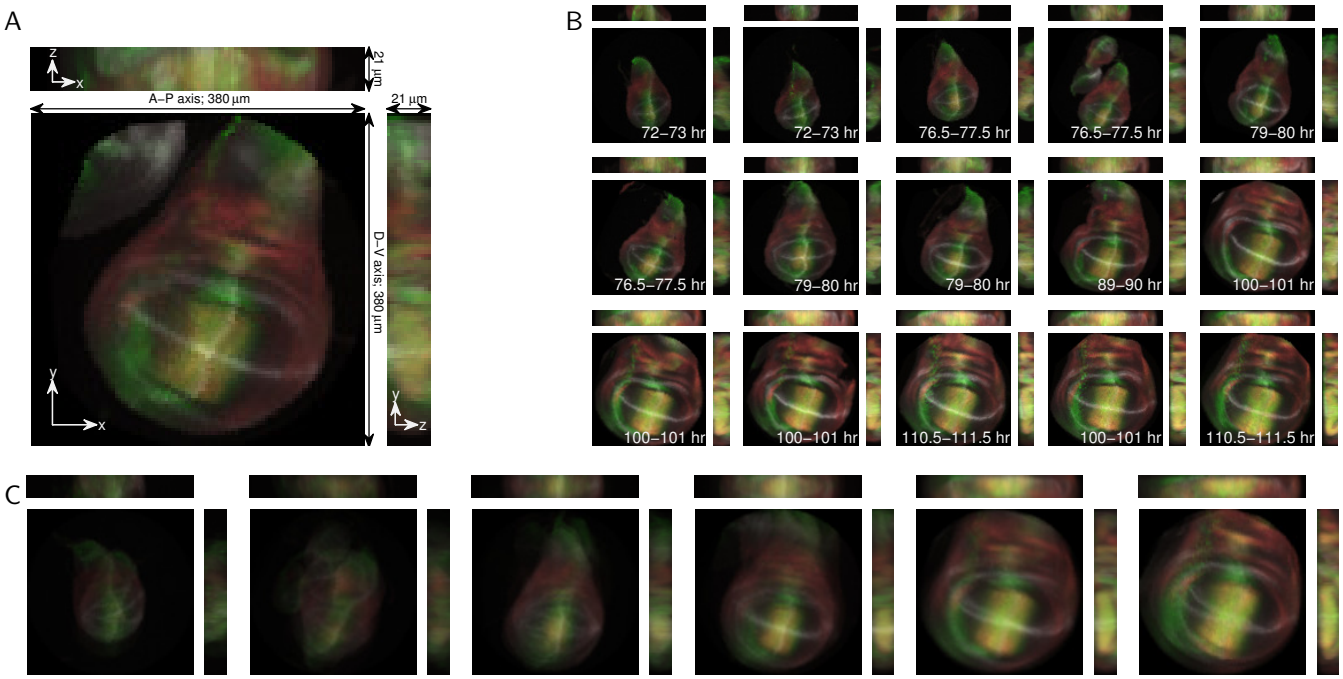


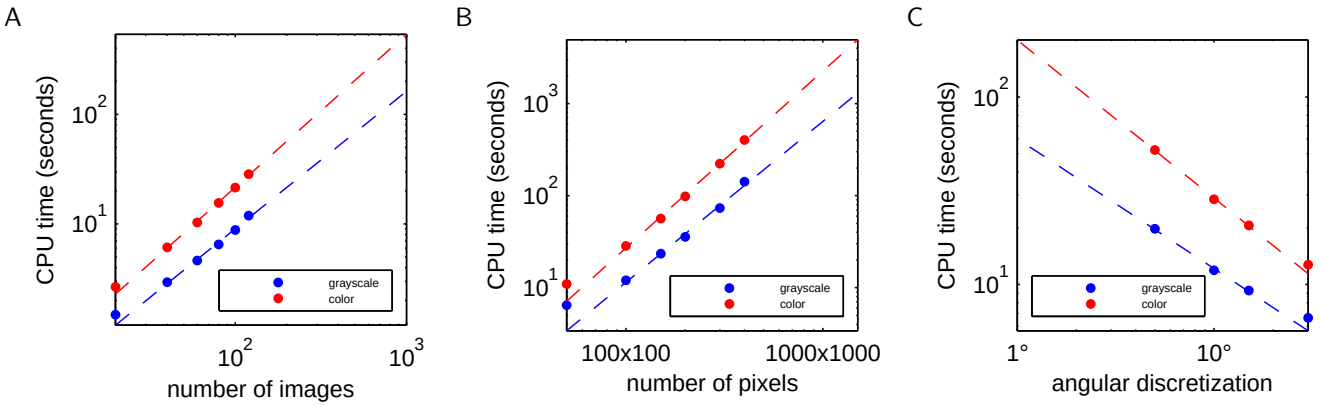
Fig. 6: Analysis of three-dimensional *Drosophila* wing disc z-stacks. (A) Maximum projections of an example three-dimensional *Drosophila* wing disc z-stack. The anterior-posterior (A-P) and dorsal-ventral (D-V) axes are indicated. Discs express the Dad-GFP reporter construct (green), and are stained for Spalt (red), Wingless (gray), and Patched (gray). Projections along the x-, y-, and z-axes are shown. (B) Example three-dimensional images, ordered using diffusion maps. The time cohort, as assessed by an expert, is indicated for each image, and the rank correlation coefficient between the diffusion maps ordering and the expert timing is 0.9427. (C) The average developmental trajectory for the registered and ordered images.

material). Furthermore, the computation of the pairwise rotational alignments, which accounts for the majority of the computational time, is trivially parallelizable, and only a subsample of the pairwise alignments need to be computed for larger data sets for accurate recovery of the underlying rotations (Singer, 2011). Because the computational cost increases with the image resolution, we chose to subsample all of our data sets to 100×100 pixels. This resolution allowed us to rapidly analyze our data sets while still retaining all of the relevant developmental features. However, as can be seen from the computational costs in Fig. 7, it is feasible to use our algorithms to analyze higher-resolution images.

The requisite user intervention and parameter tuning required for our method is relatively minor. As a first step, images must be preprocessed so that the Euclidean distance between the pixels is informative. Our software provides several preprocessing options (such as blurring, rescaling, and mean-centering), as well as some guidance for what options to select depending on the system of interest. Two algorithmic parameters, the angular discretization to compute the pairwise alignments and the diffusion maps kernel scale which determines which data points are "close" (see Fig. 2 and supplementary material), must also be defined. We also provide some guidance on selecting these parameters, and found that the results are robust to both of these parameters. Overall, the tasks of image preprocessing and parameter selection are relatively simple compared to manual registration and ordering of images, and so this methodology is promising for much larger imaging data sets which are impractical to evaluate manually.

Discussion

Temporal ordering of large-scale data was done in the context of molecular profiling studies, in which data points are vectors describing the expression levels of different mRNA (Anavy et al., 2014; Trapnell et al., 2014; Gupta and Bar-Joseph, 2008). At the same time, temporal ordering of imaging data sets was done with a significant amount of human supervision and using registered images as a starting point (Yuan et al., 2014; Surkova et al., 2008; Fowlkes et al., 2008), or using some *a priori* knowledge of the relevant developmental processes (Dubuis et al., 2013). In contrast to most of the existing registration approaches which rely on the knowledge of appropriate landmarks in the images (Dryden and Mardia, 1998) (such as the eyes in face recognition applications (Zhao et al., 2003)), algorithms based on angular synchronization can register images



D

Data Set	Data Type	Number of Channels	Number of Images	Number of Pixels	Angular Discretization	CPU time
Drosophila gastrulation (live)	2D	1	40	100 × 100	10°	3.2 sec
Zebrafish epiboly	2D	1	120	100 × 100	10°	13 sec
Drosophila gastrulation (fixed)	2D	3	120	100 × 100	10°	29 sec
Drosophila wing discs	3D	3	46	100 × 100 × 21	10°	12 sec

Fig. 7: Computational requirements for the presented methodology. (A) CPU time as a function of the number of images in the data set (for 100 × 100 pixel images, and 10° angular discretization). Empirically, the CPU time is $\sim \mathcal{O}(n^{1.33})$ in number of images. (B) CPU time as a function of the number of pixels in the images (for 120 images, and 10° angular discretization). Empirically, the CPU time is $\sim \mathcal{O}(n^{1.83})$ in the number of pixels. (C) CPU time as a function of the number of rotations (for 120 images of 100 × 100 pixels). Empirically, the CPU time is $\sim \mathcal{O}(n^{-0.77})$ in the angular discretization. (D) The algorithm settings and computational requirements for the data sets analyzed. All times are reported for an Intel Core i7 2.93 GHz processor.

even in the absence of such information, making them relevant for a wide variety of applications.

Angular synchronization and vector diffusion maps have been used to reconstruct molecular shapes from cryo-electron microscopy images (Singer and Wu, 2012; Zhao and Singer, 2014; Singer et al., 2011). Because of high levels of instrument noise in these data, thousands of images were needed for successful shape reconstruction. Based on the presented results, we expect that much smaller data sets may be sufficient for successful reconstruction of developmental trajectories from snapshots of fixed tissues. In general, the size of the data set required for accurate registration and ordering is a function of the instrument noise, interembryo variability, and the complexity of the developmental dynamics.

The benefits of our approach to image data mining are twofold. First, the algorithm can accomplish the tasks of registration and ordering in a single step. Furthermore, because our methodology is nonlinear, it can successfully order data sets which contain complex dynamics (see Fig. S7 for a comparison of ordering using linear principal component analysis versus vector diffusion maps for the data sets presented in this paper). We expect nonlinear techniques to be necessary for larger data sets which span a wider dynamic range. The main utility of our proposed methodology lies in the analysis of data sets containing hundreds of images from systems which have not been well-studied. For such data sets, manual ordering of the images can be nontrivial, and our algorithms can clearly accelerate uncovering the underlying developmental dynamics.

We acknowledge that our methods, though general, do have limitations. The first is that we require enough data to sufficiently sample the developmental trajectory. Therefore, for very small and/or very noisy data sets, our algorithms may fail. Second, the pertinent image features need to be large compared to the noise and the image resolution. In all of our examples, the relevant expression patterns and morphological structures span several pixels and are large compared to both the instrument noise and embryo-to-embryo variability,

making the Euclidean distance between pixels a good measure of images similarity.

Vector diffusion maps allow us to automatically register images, an essential task for many applications. Simultaneously, the algorithm provides us with parameters to describe each image. In the examples presented here, we have focused on ordering the images in time using the first vector diffusion maps coordinate. In general, we can recover several coordinates which concisely and comprehensively describe the data set. This parametrization can then be used for typical data analysis tasks, such as outlier detection and model fitting. Furthermore, images taken from different viewing directions can be analyzed, as the vector diffusion maps parametrization will organize the images according to the viewing angle (Singer et al., 2011). Another direction for future work is related to the joint analysis of data sets provided by different imaging approaches, such as merging live imaging data of tissue morphogenesis with snapshots of cell signaling and gene expression from fixed embryos (Krzic et al., 2012; Ichikawa et al., 2014; Rübel et al., 2010; Dsilva et al., 2013). In the future, it would also be interesting to explore the connections between our proposed approach and recently developed methods for ordering and classification of face images (Kemelmacher-Shlizerman et al., 2011, 2014) Given the rapidly increasing volumes of imaging data from studies of multiple developmental systems, we expect that dimensionality reduction approaches discussed in this work will be increasingly useful for biologists and motivate future applications and algorithmic advances.

Materials and Methods

Drosophila embryo experiments

Oregon-R was used as wild type *Drosophila* strains. Embryos were collected and fixed at 22°C. Monoclonal rabbit anti-dpERK (1:100, Cell signaling) and rat anti-Twist (1:500, a gift from Eric Wieschaus) were used to stain proteins of interest. DAPI (1:10,000, Vector Laboratories) was used to visualize nuclei, and Alexa Fluors (1:500, Invitrogen) were used as secondary antibodies. Histone-RFP strain was used to obtain time-lapse movie of gastrulating embryos at 22°C. Live embryos were loaded to the microfluidic device with PBST to keep them oxidized, and fixed embryos were loaded with 90% glycerol.

Drosophila embryo microscopy

Nikon A1-RS scanning confocal microscope, and the Nikon 60x Plan-Apo oil objective was used to image *Drosophila* embryos. Embryos were collected, stained, and imaged together under the same microscope setting. End-on imaging was performed by using the microfluidics device described previously (Chung et al., 2011). Images were collected at the focal plane $\sim$90 μm from the posterior pole of an embryo (see Fig. S1).

Image preprocessing

Images were subsampled, normalized, blurred, and centered prior to diffusion maps analysis to remove any variations due to the experimental and imaging framework. Details about the specific preprocessing operations applied to each imging data set are given in supplementary material.

Software and imaging data

All algorithms and analysis were implemented in MATLAB® (R2013b, The MathWorks, Natick, Massachusetts). Software, including documentation and tutorials, along with the full imaging data sets used in this paper are available at `genomics.princeton.edu/stas/publications.html` under "Codes and Data".

Acknowledgements

The authors thank Fisun Hamaratoglu and Markus Affolter for providing the wing disc data. The authors thank Angela DePace, Granton Jindal, Adam Finkelstein, Thomas Funkhouser, and John Storey for helpful discussions.

Author contributions

Scientific approaches were developed by C.J.D, B.L., H.L., A.S., S.Y.S, and I.G.K. Experiments were performed by B.L. Data analysis was performed by C.J.D. and B.L. The manuscript was prepared and edited by C.J.D, B.L., H.L., A.S., S.Y.S, and I.G.K.

Funding

C.J.D. was supported by the Department of Energy Computational Science Graduate Fellowship (CSGF), grant number DE-FG02-97ER25308, and the National Science Foundation Graduate Research Fellowship, Grant No. DGE 1148900. B.L. and S.Y.S. were supported by the National Institutes of Health Grant R01GM086537. H.L. was supported by the National Science Foundation Grant Emerging Frontiers in Research and Innovation (EFRI) 1136913. A.S. was supported by the Air Force Office of Scientific Research Grant FA9550-12-1-0317. I.G.K. was supported by the National Science Foundation (CS&E program).

References

Ahuja, S., Kevrekidis, I. G. and Rowley, C. W. (2007). Template-based stabilization of relative

equilibria in systems with continuous symmetry. *Journal of Nonlinear Science* **17**, 109–143.

Anavy, L., Levin, M., Khair, S., Nakanishi, N., Fernandez-Valverde, S. L., Degnan, B. M. and Yanai, I. (2014). Blind ordering of large-scale transcriptomic developmental timecourses. *Development* pp. 1161–1166.

Arie-Nachimson, M., Kovalsky, S. Z., Kemelmacher-Shlizerman, I., Singer, A. and Basri, R. (2012). Global motion estimation from point matches. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pp. 81–88. IEEE.

Belkin, M. and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* **15**, 1373–1396.

Castro, C., Luengo-Oroz, M., Desnoulez, S., Duloquin, L., Fernandez-de Manuel, L., Montagna, S., Ledesma-Carbayo, M., Bourgine, P., Peyrieras, N. and Santos, A. (2009). An automatic quantification and registration strategy to create a gene expression atlas of zebrafish embryogenesis. In *Engineering in Medicine and Biology Society*, pp. 1469–1472.

Chung, K., Kim, Y., Kanodia, J. S., Gong, E., Shvartsman, S. Y. and Lu, H. (2011). A microfluidic array for large-scale ordering and orientation of embryos. *Nat. Methods* **8**, 171–176.

Coifman, R. R. and Lafon, S. (2006). Geometric harmonics: a novel tool for multiscale out-of-sample extension of empirical functions. *Appl. Comput. Harmon. Anal.* **21**, 31–52.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F. and Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7426–7431.

Cucuringu, M., Singer, A. and Cowburn, D. (2012). Eigenvector synchronization, graph rigidity and the molecule problem. *Information and Inference* **1**, 21–67.

Dryden, I. L. and Mardia, K. (1998). *Statistical shape analysis*. Wiley series in probability and statistics: Probability and statistics. J. Wiley.

Dsilva, C. J., Talmon, R., Rabin, N., Coifman, R. R. and Kevrekidis, I. G. (2013). Nonlinear intrinsic variables and state reconstruction in multiscale simulations. *The Journal of chemical physics* **139**, 184109.

Dubuis, J. O., Samanta, R. and Gregor, T. (2013). Accurate measurements of dynamics and reproducibility in small genetic networks. *Molecular Systems Biology* **9**.

Fernández, Á., Rabin, N., Coifman, R. R. and Eckstein, J. (2014). Diffusion methods for aligning medical datasets: Location prediction in CT scan images. *Med. Image Anal.* **18**, 425–432.

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

Greenspan, H., Belongie, S., Goodman, R. and Perona, P. (1994). Rotation invariant texture recognition using a steerable pyramid. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, volume 2, pp. 162–167. IEEE.

Gupta, A. and Bar-Joseph, Z. (2008). Extracting dynamics from static cancer expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **5**, 172–182.

Hajnal, J. V. and Hill, D. L. (2010). *Medical image registration*. CRC press.

Hamaratoglu, F., de Lachapelle, A. M., Pyrowolakis, G., Bergmann, S. and Affolter, M. (2011). Dpp signaling activity requires pentagone to scale with tissue size in the growing drosophila wing imaginal disc. *PLoS biology* **9**, e1001182.

Ichikawa, T., Nakazato, K., Keller, P. J., Kajiura-Kobayashi, H., Stelzer, E. H., Mochizuki, A. and Nonaka, S. (2014). Live imaging and quantitative analysis of gastrulation in mouse embryos using light-sheet microscopy and 3D tracking tools. *Nat. Protoc.* **9**, 575–585.

Jaeger, J. et al. (2004). Dynamic control of positional information in the early *Drosophila* embryo. *Nature* **430**, 368–371.

Kane, D. A. et al. (1996). The zebrafish epiboly mutants. *Development* **123**, 47–55.

Karlstrom, R. O. and Kane, D. A. (1996). A flipbook of zebrafish embryogenesis. *Development* **123**, 461–462.

Kemelmacher-Shlizerman, I., Shechtman, E., Garg, R. and Seitz, S. M. (2011). Exploring photobios. In *ACM Trans. Graph.*, volume 30, p. 61.

Kemelmacher-Shlizerman, I., Suwajanakorn, S. and Seitz, S. M. (2014). Illumination-aware age progression. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 3334–3341. IEEE.

Krzic, U., Gunther, S., Saunders, T. E., Streichan, S. J. and Hufnagel, L. (2012). Multiview light-sheet microscope for rapid *in toto* imaging. *Nat. Methods* **9**, 730–733.

Lafon, S., Keller, Y. and Coifman, R. R. (2006). Data fusion and multicue data matching by diffusion maps. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1784–1797.

Leptin, M. (2005). Gastrulation movements: the logic and the nuts and bolts. *Dev. Cell* **8**, 305–320.

Lim, B., Samper, N., Lu, H., Rushlow, C., Jiménez, G. and Shvartsman, S. Y. (2013). Kinetics of gene derepression by ERK signaling. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 10330–10335.

Ng, L. L., Sunkin, S. M., Feng, D., Lau, C., Dang, C. and Hawrylycz, M. J. (2012). Large-scale neuroinformatics for *in situ* hybridization data in the mouse brain. *Int. Rev. Neurobiol.* **104**, 159–182.

Peter, I. S. and Davidson, E. H. (2011). A gene regulatory network controlling the embryonic specification of endoderm. *Nature* **474**, 635–639.

Richardson, L., Stevenson, P., Venkataraman, S., Yang, Y., Burton, N., Rao, J., Christiansen, J. H., Baldock, R. A. and Davidson, D. R. (2014). Emage: Electronic mouse atlas of gene expression. In *Mouse Molecular Embryology*, pp. 61–79. Springer.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326.

Rowley, H. A., Baluja, S. and Kanade, T. (1998). Rotation invariant neural network-based face detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 38–44.

Rübel, O. et al. (2010). Coupling visualization and data analysis for knowledge discovery from multi-dimensional scientific data. *Procedia computer science* **1**, 1757–1764.

Singer, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.* **30**, 20–36.

Singer, A. and Wu, H.-T. (2012). Vector diffusion maps and the connection Laplacian. *Commun. Pure Appl. Math.* **65**, 1067–1144.

Singer, A., Zhao, Z., Shkolnisky, Y. and Hadani, R. (2011). Viewing angle classification of cryo-electron microscopy images using eigenvectors. *SIAM J. Imaging Sci.* **4**, 723–759.

Sonday, B., Singer, A. and Kevrekidis, I. G. (2013). Noisy dynamic simulations in the presence of symmetry: Data alignment and model reduction. *Computers & Mathematics with Applications* **65**, 1535–1557.

Surkova, S., Kosman, D., Kozlov, K., Myasnikova, E., Samsonova, A. A., Spirov, A., Vanario-Alonso, C. E., Samsonova, M., Reinitz, J. et al. (2008). Characterization of the *Drosophila* segment determination morphome. *Dev. Biol.* **313**, 844–862.

Tenenbaum, J. B., De Silva, V. and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386.

Wang, L. and Singer, A. (2013). Exact and stable recovery of rotations for robust synchronization. *Information and Inference* p. iat005.

Yuan, L., Pan, C., Ji, S., McCutchan, M., Zhou, Z.-H., Newfeld, S. J., Kumar, S. and Ye, J. (2014). Automated annotation of developmental stages of *Drosophila* embryos in images containing spatial patterns of expression. *Bioinformatics* **30**, 266–273.

Zhao, W., Chellappa, R., Phillips, P. J. and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys (CSUR)* **35**, 399–458.

Zhao, Z. and Singer, A. (2014). Rotationally invariant image representation for viewing direction classification in cryo-EM. *J. Struct. Biol.* **186**, 153 – 166.

Zitova, B. and Flusser, J. (2003). Image registration methods: a survey. *Image Vis. Comput.* **21**, 977–1000.