INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Szigeti B, Deogade A, Webb B. 2015 Searching for motifs in the behaviour of larval *Drosophila melanogaster* and *Caenorhabditis elegans* reveals continuity between behavioural states. *J. R. Soc. Interface* **12**: 20150899.
http://dx.doi.org/10.1098/rsif.2015.0899

Received: 12 October 2015
Accepted: 4 November 2015

Subject Areas:
computational biology, systems biology

Keywords:
Drosophila melanogaster larva, *Caenorhabditis elegans*, eigenworms, eigenmaggots, behavioural states, ethogram

Author for correspondence:
Balázs Szigeti
e-mail: b.szigeti@sms.ed.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2015.0899 or via http://rsif.royalsocietypublishing.org.

THE ROYAL SOCIETY
PUBLISHING

Searching for motifs in the behaviour of larval *Drosophila melanogaster* and *Caenorhabditis elegans* reveals continuity between behavioural states

Balázs Szigeti[1], Ajinkya Deogade[2] and Barbara Webb[3]

[1]Neuroinformatics Doctoral Training Centre, University of Edinburgh, Edinburgh, UK
[2]EMBL-CRG Systems Biology Program, Barcelona, Spain
[3]School of Informatics, University of Edinburgh, Edinburgh, UK

AD, 0000-0003-3369-0370

We present a novel method for the unsupervised discovery of behavioural motifs in larval *Drosophila melanogaster* and *Caenorhabditis elegans*. A motif is defined as a particular sequence of postures that recurs frequently. The animal's changing posture is represented by an eigenshape time series, and we look for motifs in this time series. To find motifs, the eigenshape time series is segmented, and the segments clustered using spline regression. Unlike previous approaches, our method can classify sequences of unequal duration as the same motif. The behavioural motifs are used as the basis of a probabilistic behavioural annotator, the eigenshape annotator (ESA). Probabilistic annotation avoids rigid threshold values and allows classification uncertainty to be quantified. We apply eigenshape annotation to both larval *Drosophila* and *C. elegans* and produce a good match to hand annotation of behavioural states. However, we find many behavioural events cannot be unambiguously classified. By comparing the results with ESA of an artificial agent's behaviour, we argue that the ambiguity is due to greater continuity between behavioural states than is generally assumed for these organisms.

1. Introduction

Automated analysis of behaviour is of increasing importance to biology and neuroscience. Behavioural control is the ultimate function of neural processing [1]. The recent expansion of tools for manipulating neural activity, such as optogenetics, has made it crucial to be able to screen rapidly and automatically for the behavioural consequences of these manipulations. Standardization of quantitative behavioural assays and reproducibility of analyses are thus key to progress in understanding neural circuits.

Traditional manual annotation of behavioural data is not feasible for large datasets. As a consequence, automated high-throughput behavioural annotators have been developed. An example is the Janelia Automatic Animal Behaviour Annotator (JAABA) [2]. JAABA first requires hand annotation of a subset of the data and then the software uses machine learning algorithms to find the same patterns in the unannotated data. Other researchers have developed classifiers that extract specific parameters from behavioural data and then register a state if a certain parameter (or parameter set) exceeds a user-defined threshold [3–6]. Note that for these classifiers both the set of possible behaviours and the description of those behaviours are encoded by the user. In contrast, our goal is to discover patterns in behaviour without reference to any user-defined thresholds or examples.

Posture is the main observable component of behaviour, and the behavioural annotators mentioned above mainly use postural information as input to classify behavioural states. In this context, Stephens *et al*. introduced eigenworms [7],

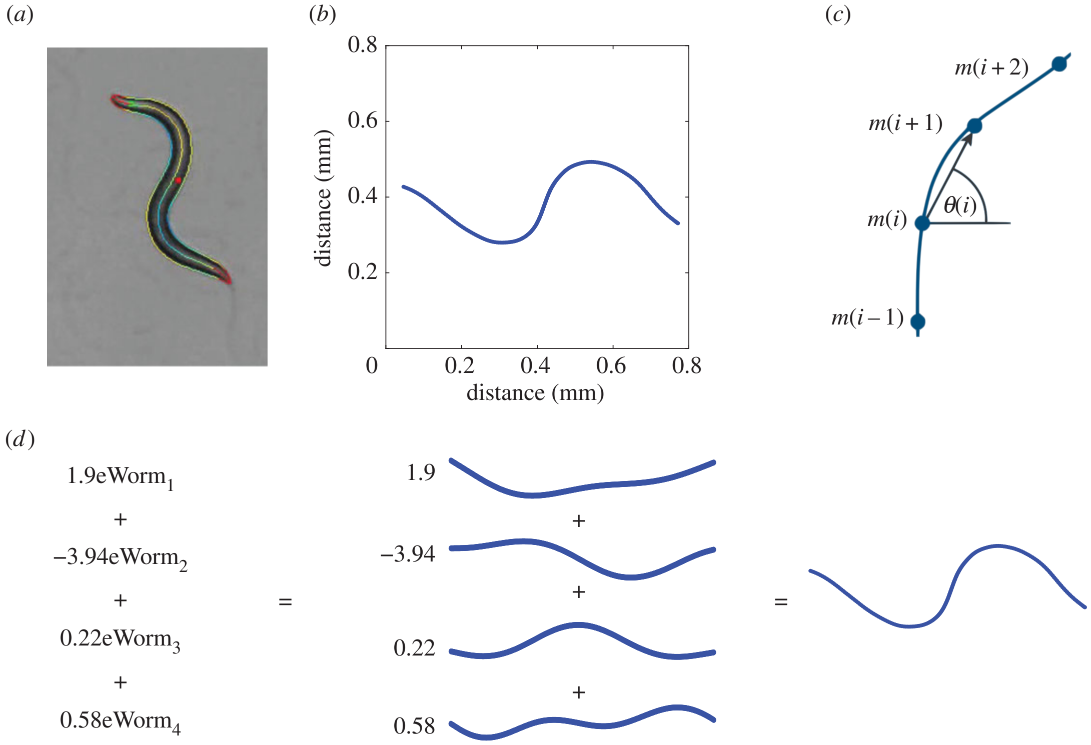

Figure 1. Constructing eigenworms. In each video frame, thresholding is used to separate the animal from the background, then the resulting binary images are skeletonized. This skeleton, or midline, is used as a proxy for the animal's posture. Panel (*a*) shows a frame from the CBD with the worm's contour and midline highlighted, panel (*b*) shows the corresponding midline. The skeleton has been rotated to remove the worm's overall rotation relative to the plate. Panels (*c*) zooms in on the midline, showing how a set of θ_i angles provide a piecewise linear approximation to the midline curvature. This angular data forms a vector for each frame, or a matrix for a movie. The matrix's principal components are the eigenworms. Panel (*d*) shows an example of a posture reconstruction. The blue shapes in the middle column are the eigenworms, which can be added together with different weights to reconstruct any actual worm posture.

using principal component analysis to produce a low-dimensional representation of *C. elegans* midline shapes. For the unrestricted free behaviour of *C. elegans*, four eigenworms account for 92% of the animal's posture variance. This means that four numbers can describe any actual worm posture with high precision. Mathematically, postures are described by a superposition of eigenworms, i.e.

$$\text{posture}(t) = \sum_{i=1}^{n} \alpha_i(t)\text{eigenworm}_i, \qquad (1.1)$$

where $\alpha_i(t)$ is the coefficient associated with the ith eigenworm at time t. Figure 1 shows the eigenworms and an example of posture reconstruction. Eigenshapes provide a compact representation of posture and hence clearly have potential use in behavioural annotation. Specifically, behaviour (change in posture over time) is represented by the time evolution of eigenshape coefficients, i.e. the time series of $\alpha_i(t)$s. This time series will be referred to as the eigenshape coefficient time series (ECTS) and forms the basis of our method.

The technical aim of this paper is the unsupervised discovery of frequently repeated ECTS subsequences. In the data mining literature, frequently repeated subsequences are also known as motifs [8]. ECTS motifs correspond to frequently repeated sequences of posture that can be viewed as behavioural states or actions [9,10]. Previous attempts to extract ECTS motifs using a simple 'sliding window' motif discovery approach [11] suffer from two major problems. First, the window for any pass is of fixed length, hence this method considers only exactly equal duration sequences as potential matches. Second, the sliding window method defines a motif as a *pair* of closest neighbour sequences. However, motifs are understood intuitively not as a single pair of subsequences, but as a *frequently repeated* subsequence. Our motif finding methodology was designed to overcome these two problems.

First, we derive the equivalent of eigenworms for larval *Drosophila*, termed eigenmaggots. The ECTS of both larval *Drosophila* and *C. elegans* are then analysed using our novel motif finding method. The ECTS motifs are used as the basis of a probabilistic behavioural annotator, the eigenshape annotator (ESA).[1] We show that the resulting annotation corresponds well to hand annotation, although a number of behaviours cannot be unambiguously classified. The ESA analysis is also applied to the behaviour of a state-based simulated maggot to show that the ambiguity is not inherent in the method, but reflects a greater continuity between behavioural states in these organisms than is generally assumed. In summary, our new method both confirms the results of previous behavioural annotation and reveals some of its limitations.

2. Methods

2.1. Overview

Our aim is to go from video of a behaving animal to annotation of its behavioural states, where those states are determined using bottom-up discovery of motifs in the sequence of postures. We start by recording freely foraging *Drosophila* larva, extract their midline as a set of angles, and apply principal component analysis to obtain a low-dimensional description of postures, the ECTS. Equivalent information for the worm is available from the *C. elegans* behavioural database (CBD). Discovering motifs in the multidimensional ECTS is a non-trivial problem, and there are no existing adequate tools. We developed a two-step process to

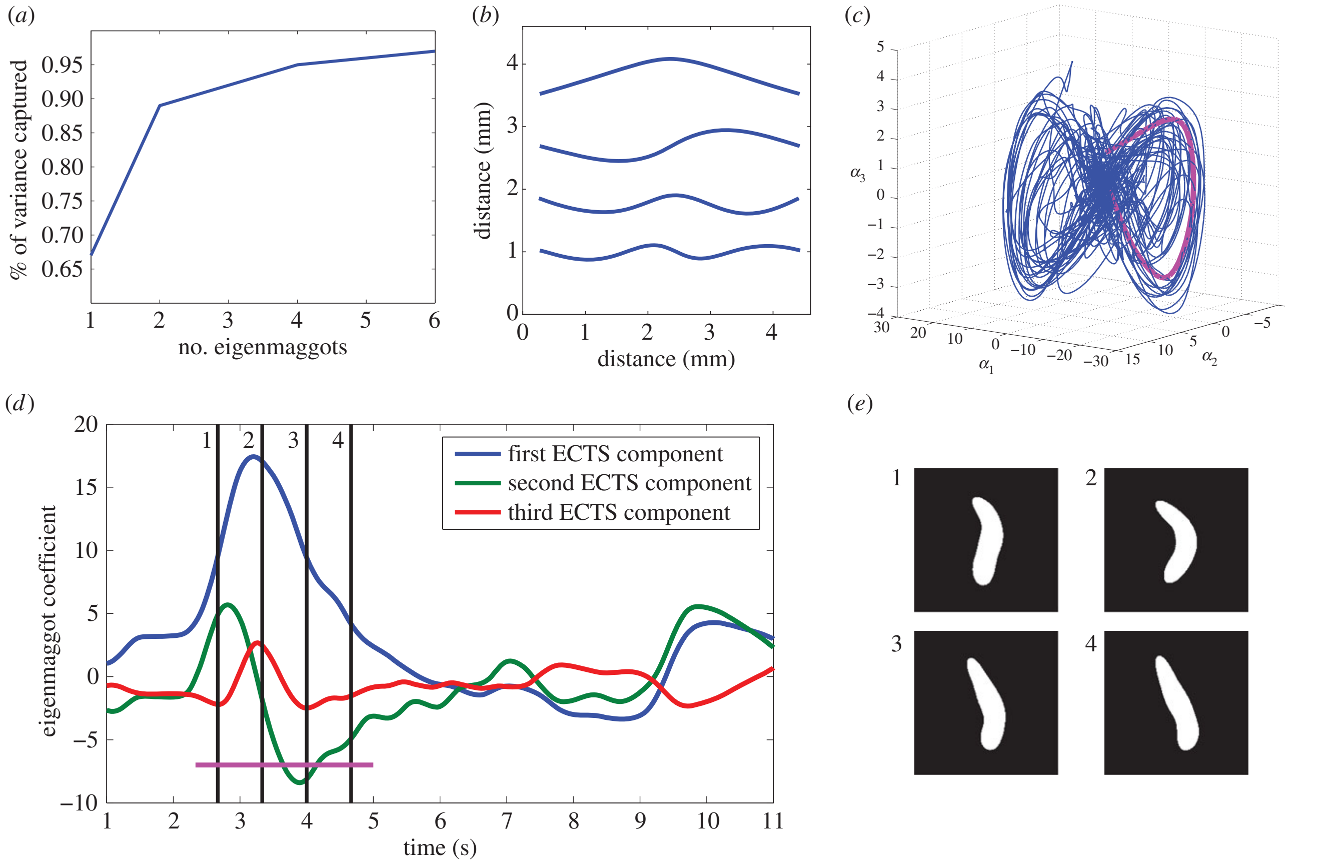

Figure 2. Results of eigenmaggot analysis. Panel (*a*) shows the percentage of the original data's variance recovered given the dimensionality of the representation. Panel (*b*) shows the eigenmaggots with the most significant eigenmaggot on top, the second below, etc. These shapes can be added up in different proportions to reproduce the larval postures (figure 1). Panel (*c*) shows a three-dimensional behavioural trajectory in eigenmaggot space, that is the time evolution of the first three eigenmaggot coefficients. The subtrajectory highlighted is an example of what we call a *turning manoeuvre*, see §3.2. Panel (*d*) shows a part of the same trajectory as three separate one-dimensional time series; the subsequence underlined corresponds to the highlighted subtrajectory on panel (*c*). Panel (*e*) shows binary images of the maggot at the corresponding time slices from panel (*d*).

first extract subsequences and then fit a statistical model to cluster the subsequences. Briefly (details are given below), we use changes in the dynamics of the ECTS to divide the sequence into variable length subsequences, with the intent that each subsequence contains a single 'action'. The subsequences are aligned and then clustered using a spline regression model [12,13], a method for analysing curves analogous to Gaussian mixture models. The resulting clusters constitute motifs by which the animal's behaviour can be annotated. The results are compared with alternative annotation systems and with hand annotation provided by a human expert, which is treated as ground truth.

2.2. Data collection

Canton-S flies were maintained on conventional cornmeal-agar molasses medium at 22°C and kept in a 12 h dark–light cycle. For the behavioural experiments, larvae in their 3rd instar stage were placed on 3% agarose and were allowed to freely forage. Across 33 individuals, 14 h of video was recorded at 30 fps. The videos were segmented (see below) into a total of 11 613 actions. The tracking and data acquisition hardware used for this publication are described in detail in [14]. Briefly, the larva moving over a fixed stage was imaged using a camera (Basler A622f) on top. The camera was mounted on a moving stage to follow the animal. The software for image capture and stage control was written in C using the OpenCV libraries.

To analyse worm behaviour, we used data from the *CBD* [6]. The database consists of videos of worms (recorded at 30 fps) browsing in bacteria. For every video, there is a corresponding feature file, which contains many precalculated statistics of worm morphology. The feature files also contain the eigenworm coefficient time series. The worm analysis in this paper uses this precalculated ECTS. Twenty-two thousand and sixty-six actions were analysed from 100 experiments with N2 worms, corresponding to 25 h of video.

2.3. Constructing eigenmaggots

In each video frame, the larva was separated from the background by a thresholding algorithm. The resulting binary images were skeletonized using the built-in MATLAB function [15]. Midlines were rotated such that the endpoints, corresponding to the head and tail of the animal, lie along the x-axis. This operation removes the overall rotation of the animal's body relative to the plate. The midlines were normalized such that they consist of 71 points placed equidistant from each other. The length of the larva can change, but is neglected in this analysis, i.e. we treat every midline as if it is the same length. The eigenshapes in figures 1 and 2 have been reconstructed to reflect the average physical size of the midlines. The angles among consecutive points defining the midline were restricted to the interval $-\pi < \theta_i \leq \pi$. As a result of these operations, each frame is associated with a 70 dimensional vector, where the ith component is θ_i (figure 1*c*). These vectors are concatenated to form an $n*70$ data matrix, where n is the number of frames. Principal component analysis is applied to this data matrix to construct the eigenshapes and the associated ECTS.

2.4. Eigenshape coefficient time series

For both the larval and worm analysis, the coefficients of the three most significant eigenshapes were included in the ECTS, that is $\mathrm{ECTS}(t) = [\alpha_1(t), \alpha_2(t), \alpha_3(t)]$, see equation (1.1). After principal component analysis, the inspection of the eigenvalues reveals that for both organisms three coefficients account for approximately 90% of the posture variance [16], thus provide an accurate description of posture. At the same time, a three-dimensional ECTS is small enough to avoid 'the curse of dimensionality' that could lead to difficulties during the clustering step [17].

2.5. Dropped frames

Both the larval and the maggot ECTS contains dropped frames. If a gap was short (less than 0.5 s), then ECTS was linearly interpolated. After the interpolation, 1.1% of the *Drosophila* and 4.2% of the *C. elegans* frames were still missing. For both organisms on a significant portion of the dropped frames, the animal was curled up in a 'doughnut shape' from which it is difficult to extract a biologically meaningful skeleton. For *C. elegans*, more frames were dropped, because the worms were browsing in food. The layer of bacteria can obscure the worm in the image, making separation of the body of the worm from the background more challenging. Note that the inability to analyse curled-up postures introduces a bias to the pipeline, as no posture with self-intersection is included.

2.6. Segmentation

The intuition behind the segmentation algorithm is that boundaries between windows should be located where the dynamics of ECTS changes. ECTS was smoothed using a weighted running average filter with a window size of four frames and weights inversely proportional to the distance from the window's centre. Segmentation operates on a 'body score' time series that is created by calculating a weighted sum of the separate dimensions of ECTS, where the weights are set by the eigenvalues associated with the eigenshapes. The segmentation algorithm scans the body score to find local minima and maxima. An action is defined as a local maxima in body score bounded by minimas. The minimas define the start and end of the segmented subsequence. Figure 3 shows the result of segmentation for *Drosophila* and *C. elegans* with the corresponding body score time series.

The maxima/minima finding algorithm is controlled by a master parameter. The results are not strongly dependent on the precise parameter setting: adjusting it by ±25% leaves 92% of the annotation unchanged.

The behavioural videos of *C. elegans* were recorded while the worms were browsing in food. In this environment, worms often show low activity. Our segmentation was designed to identify periods where the body score rapidly changes, hence the identification of low activity periods required an extra step. Low activity periods were identified by intervals where the time derivative of body score remained under half of its average value for more than 0.5 s. These periods were added to the collection of actions prior to proceeding to the clustering step. If the two parameters (less than 50% of average body score for more than 0.5 s) are adjusted ±25%, then 97% of the action's classifications are not altered. Thus, fine tuning of the parameters is not necessary.

2.7. Curve alignment and clustering

Segmentation produces a large set of subsequences, or actions, each of which is a continuous ECTS curve. Hence, splines, locally smooth piecewise polynomials, are a natural choice to parametrize actions. Spline regression [12,13] was used to assign the actions to clusters. This method is analogous to Gaussian mixture models, but instead of Gaussian distributions, clusters are parametrized by splines.

To improve the consistency of spline fitting, the ECTS subsequences were aligned in the time domain. The frame with the highest body score was used as a reference, and actions were shifted in time such that their point of highest body score coincides, see electronic supplementary material, figure S2 for illustration. Note that if ECTS = [0, 0, 0], then the posture is a flat line (for both organisms). The higher the coefficients are, generally the more curved the postures are (although the bend caused by the coefficients can be in opposite directions and cancel each other). Therefore, the maxima of the body score correspond to the frame with the most bent posture and as such this frame is a rational choice to define a reference point in time by which subsequences of different lengths can be aligned.

Splines had three internal knot points and each polynomial had an order of 3. An expectation–maximization (EM) algorithm [13] was used to learn model parameters. EM was initiated 500 times with random boundary conditions, and the solution with the highest likelihood was kept. Bayesian information criteria (BIC) [18–20] was used to identify the optimal number of clusters. BIC is defined as

$$\mathrm{BIC} = 2\ln(\mathscr{L}_{\mathrm{model}}) - k * \ln(n), \tag{2.1}$$

where $\mathscr{L}_{\mathrm{model}}$ is the likelihood of the fitted model, k is the number of free parameters and n is the number of observations. The first term reflects goodness of fit of the model, and the second is a penalty term is for the number of free parameters.

Spline regression clustering produces a membership probability that a given action belongs to a cluster. Therefore, this method avoids rigid cluster assignments and also allows classification uncertainty to be quantified. To measure the classification uncertainty Shannon entropy [21] was used, defined as

$$H = -\sum_i p_i \log_2 p_i, \tag{2.2}$$

where p_i is the probability that a given action belongs to a cluster i. Note that the most uncertain situation is when the probability is equally distributed among the clusters, correspondingly H has a maximum when all $p_i = 1/i_{\max}$ ($i_{\max}$ is the number of clusters).

2.8. Comparison of behavioural annotations

In the following, a 'behavioural event' means an interval of consecutive frames tagged with the same behaviour. A behavioural event marked by an automated annotator (ESA, JAABA or CBD) was counted as true positive if at least 50% of it was also tagged by ground truth annotation with the same behaviour. Otherwise, the event was either counted as a false positive (automated annotator marked a behavioural event that had less than 50% overlap with an identically annotated behavioural event in the ground truth annotation) or a false negative (ground truth marked a behavioural event that had less than 50% overlap with an identically annotated behavioural event in the automated annotation).

Furthermore, we had to consider the problem that different annotations used different behavioural state spaces. The behaviours were always matched to the closest behaviour in the ground truth annotation. Specifically, for larval *Drosophila*, ESA's *turning manoeuvre* was treated as a match to both *stop cast* and *turn* in the ground truth annotation. That is, if ground truth contained either a *turn* or a *stop cast* behaviour and at least 50% of the frames were tagged as a *turning manoeuvre* by ESA, then it was counted as a true positive. *Run casts* are the same behaviour across ground truth, JAABA and ESA. For *C. elegans*, the ground truth hand annotation's *dwelling* was treated as a match to CBD's *pause* and ESA's *passive* state. The CBD's Υ and Ω turns were both treated as a match to the ground truth's *turn* behaviour.

Parts of the time series were excluded from the analysis when the video frames could not be segmented and hence midline information was not accessible. Note that JAABA, CBD and ground truth annotation is available for these periods as they do not exclusively rely on contour information.

We modified the output of JAABA to avoid the problem of 'flickering annotation'. Flickering annotation occurs when single frames within a behavioural event are not classified as

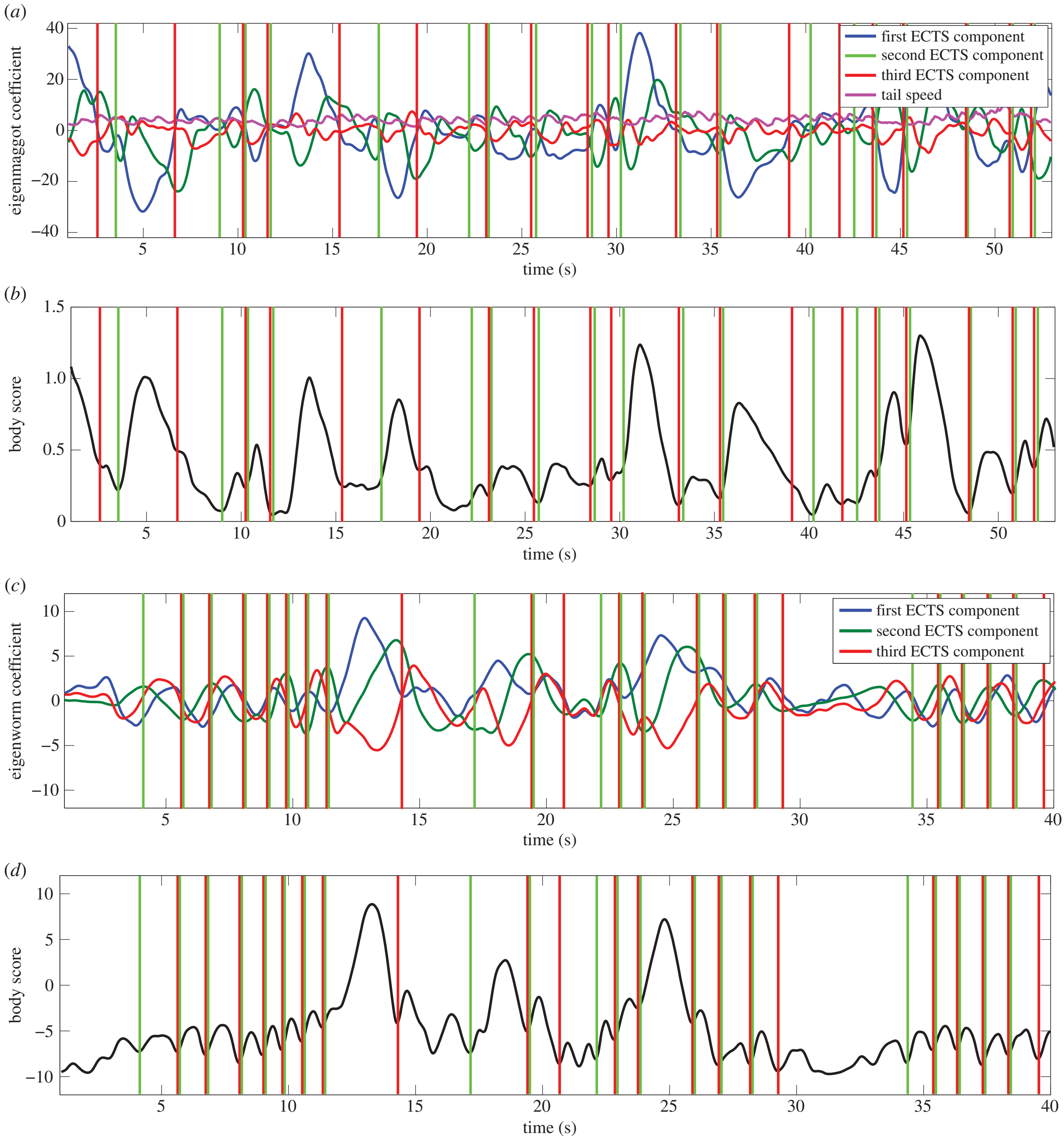

Figure 3. The segmentation algorithm. Panel (*a*) shows a screenshot of the larval ECTS and the tail speed time series. Panel (*b*) shows the corresponding body score, calculated as a weighted average of the ECTS dimensions, where the weights are set by the eigenvalue associated with each eigenshape. Local maximas and minimas in body score determine boundaries between actions, marked as green and red vertical lines for the beginning and end of actions respectively, in both panels (*a*) and (*b*). Panels (*c*,*d*) show the same information as (*a*,*b*) for *C. elegans*. The sinusoidal segments correspond to locomotion, note that segmentation resolves these into 'steps'.

part of the event, e.g. the sequence 0011011100 (where 1 means that the frame corresponds to a given behaviour, 0 means it does not). JAABA works on a frame-by-frame basis, hence these sequences are present when an event is near threshold value. To avoid the false positives caused by the small gaps, we have connected behavioural events that are less than three frames apart. Hence, the sequence above would become 0011111100.

To summarize annotation accuracy, we report the precision (positive predictive value) and sensitivity (also known as recall and true positive rate) [22] in tables 1 and 2. Sensitivity is the percentage of events recognized by the annotator, and precision is the proportion of events tagged by the annotator that are true positives. Furthermore, these two measures are combined as the *F*-score, defined as

$$F = \frac{2(\text{precision} \times \text{sensitivity})}{\text{precision} + \text{sensitivity}}, \qquad (2.3)$$

which is commonly used to quantify the goodness of classification.

2.9. Visualization, density cross sections and feature histograms

To produce figures 4*b* and 5*b* and figure S1, the standard MATLAB [15] implementation of metric multidimensional

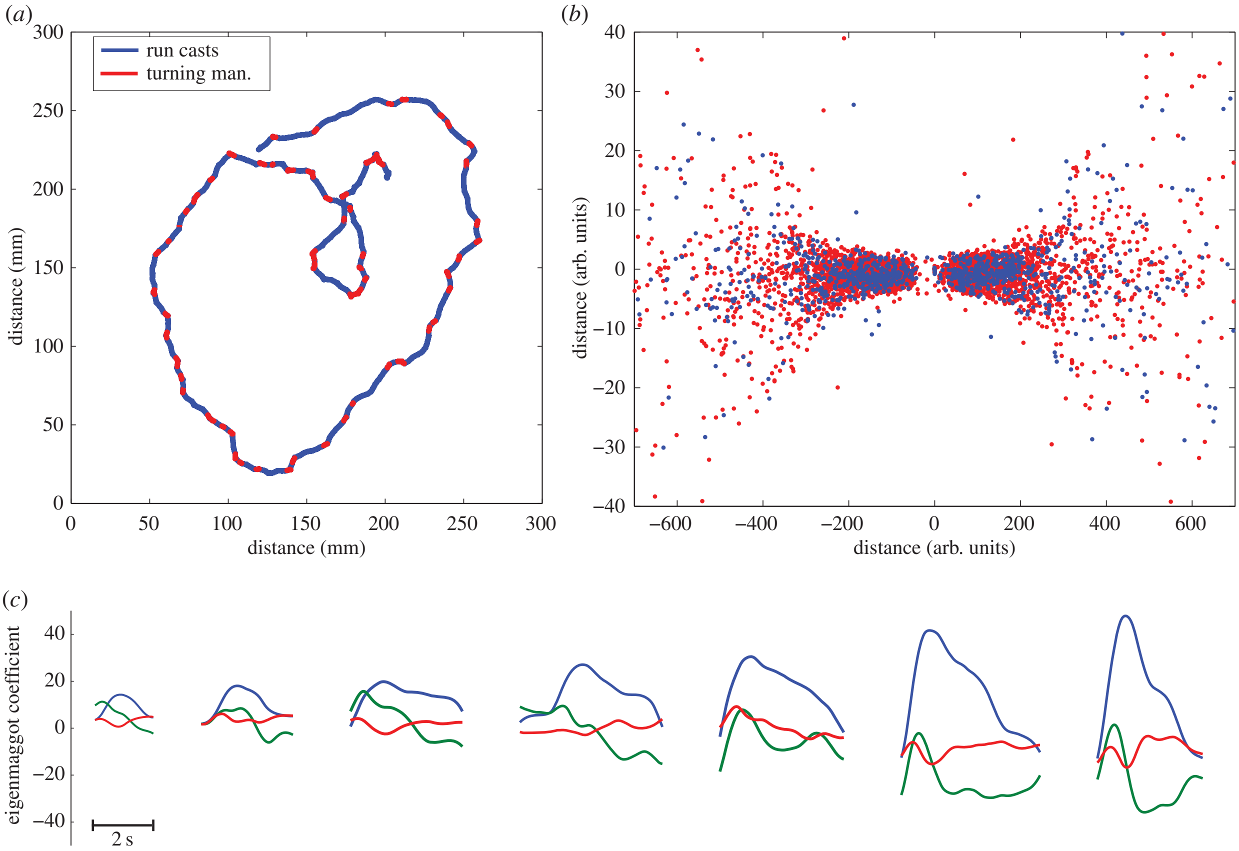

Figure 4. The structure of behavioural motifs for larval *Drosophila*. ESA identifies two motifs in the larva's behaviour; panel (*a*) shows a trajectory colour coded for the two motifs. Note that turning manoeuvres tend to happen when direction changes. Panel (*b*) shows a two-dimensional map of the distances among actions as measured by dynamic time warping ($R^2 = 0.85$), see §2.9 for details. Panel (*b*) uses the same colour scheme as panel (*a*) to distinguish behaviours. The symmetry in the figure corresponds to the left/right symmetry in the animal's behaviour. Note that the points corresponding to the two behavioural motifs are concentrated in separate regions, yet there is no clear boundary between the two set of points. Panel (*c*) illustrates that similar ECTS subsequences can be found at every scale. These actions have been selected by starting in the middle of the map in (*b*) and picking example actions at regularly spaced distances along the *x*-axis, going from left to right.

Table 1. Statistics of the annotation of larval *Drosophila* behaviour. Precision, sensitivity and *F*-score values have been derived from electronic supplementary material, table S1. See electronic supplementary material, video S3 that shows the larva's behaviour ground truth, JAABA annotation and ESA annotation next to each other.

	run cast			stop cast			turn			all behaviours		
	Pre.	Sen.	F	Pre.	Sen.	F	Pre.	Sen.	F	Pre.	Sen.	F
JAABA	0.49	0.95	0.65	0.67	0.89	0.76	0.53	0.98	0.69	0.54	0.94	0.68
ESA	0.64	0.91	0.75	0.74	0.75	0.75	0.7	0.51	0.59	0.67	0.77	0.72

Table 2. Statistics of the annotation of *C. elegans* behaviour. Precision, sensitivity and *F*-score values have been derived from electronic supplementary material, table S2. See electronic supplementary, video S4 that shows the worm's behaviour ground truth, CBD annotation and ESA annotation next to each other.

	locomotion			turn			dwelling			all behaviours		
	Pre.	Sen.	F	Pre.	Sen.	F	Pre.	Sen.	F	Pre.	Sen.	F
CBD	0.77	1	0.87	0.96	0.79	0.87	0.89	0.94	0.92	0.86	0.9	0.88
ESA	0.83	0.93	0.9	0.67	1	0.8	0.73	0.83	0.77	0.74	0.95	0.82

scaling was used. The distance matrix was constructed using weighted dynamic time warping (DTW), where the weights are set by the eigenvalue associated with each dimension of ECTS. DTW is a standard measure of similarity in time-series analysis that uses a nonlinear time warping to find the optimal match between a pair of subsequences [24]. Note that the Euclidean distance among the points (corresponding to the actions) on the map correlates with the DTW distance among the

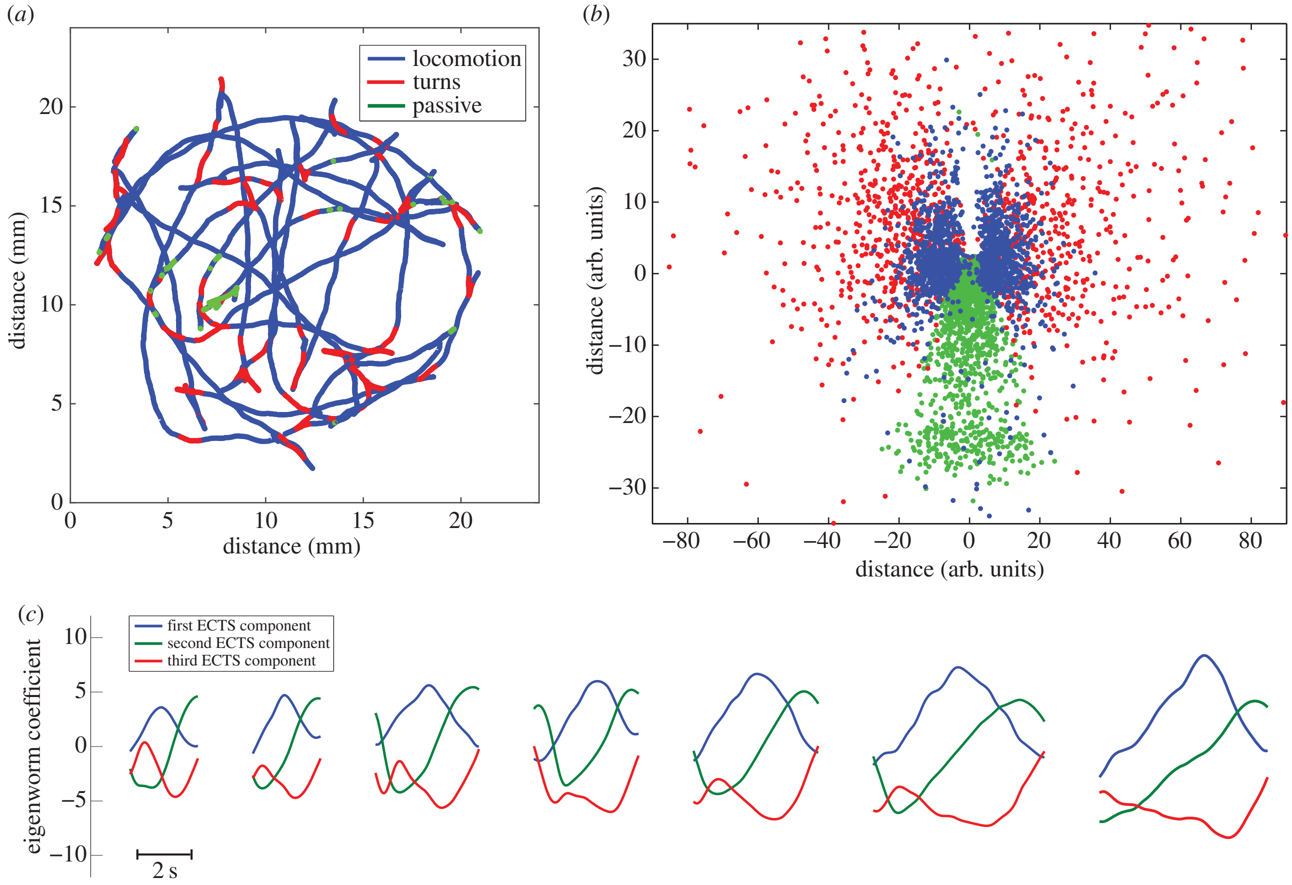

Figure 5. The structure of behavioural motifs for *C. elegans*. Panel (*a*) shows a trajectory colour coded for behaviour. Panel (*b*) shows a two-dimensional map of the distance among actions as measured by dynamic time warping ($R^2 = 0.78$), see §2.9 for details. The symmetry in the figure corresponds to the dorsal/ventral symmetry in the animal's behaviour. Note that turn events are denser on the negative side of the *x*-axis. This effect is due to the ventral bias of Ω-turns [23]. Panel (*c*) illustrates that the ECTS subsequence corresponding to turns can be found at various scales, indicating that Ω-turns are not distinct behaviour, but a part of the continuum of turning behaviours.

subsequences, but the distances on the map are in arbitrary units. To construct each map, a random sample of 5000 actions were used. The algorithm was run 500 times with random initial conditions and the solution with the highest R^2 was kept.

The density cross sections of aggregated ECTS curves were visualized to see possible density fluctuations (see §3.4). Sets of stereotypical curves would form high-density regions in the cross sections. Hence, the cross sections can be used to detect stereotypical curves corresponding to stereotypical posture sequences. Density cross sections are measured on aggregated and aligned ECTS curves at specific 'time slices' as shown in figure 6*a*. To estimate the density of curves, a kernel density estimation method was used [25]. Figure 6 only shows the cross section for one time slice, see electronic supplementary material, figure S3 for additional cross sections.

To create the histograms of *C. elegans* behavioural features, data were directly imported from the CBD feature files. These features are defined in [6]. The hardware and software that was used to obtain the behavioural features for larval *Drosophila* is described in [4].

3. Results

3.1. Eigenshapes

The eigenworm analysis pipeline extracts a vector of angles between consecutive points along the animal's midline, and applies principle component analysis to reduce the dimensionality of this description. The same method was adapted to create the analogous set of shapes for *Drosophila* larva, the eigenmaggots (figure 2). We find that eigenmaggots (figure 2*b*) are as efficient to describe larval postures as the eigenworms (figure 1*d*) are to describe worm postures. The inspection of eigenvalues reveals that three eigenmaggots account for over 90% of the postural variance [16] (figure 2*a*). Thus, eigenmaggots provide an accurate low-dimensional description of larval postures.

In contrast to eigenworms, eigenmaggots do not capture forward locomotion [7]. This difference is due to the different mode of locomotion. *C. elegans* propels itself by moving its body in a sinusoidal wave perpendicularly to the direction of motion [26]. Larval *Drosophila* crawls forward using peristaltic contraction waves [27]. The peristaltic waves can be recognized by the contraction of the abdominal sections, but this contraction does not alter the animal's midline shape from the camera's top view, and therefore is not captured by the eigenmaggot description. It is noted here that we have experimented with supplementing the larval ECTS with the tail speed time series as an extra dimension. The idea is that tail speed captures the state of peristalsis. However, the additional information did not improve the classification when evaluated against the ground truth annotation.

3.2. Motifs for *Drosophila* larva

For foraging *Drosophila* larva, the BIC for the spline regression model gave the best fit when assuming the presence of two

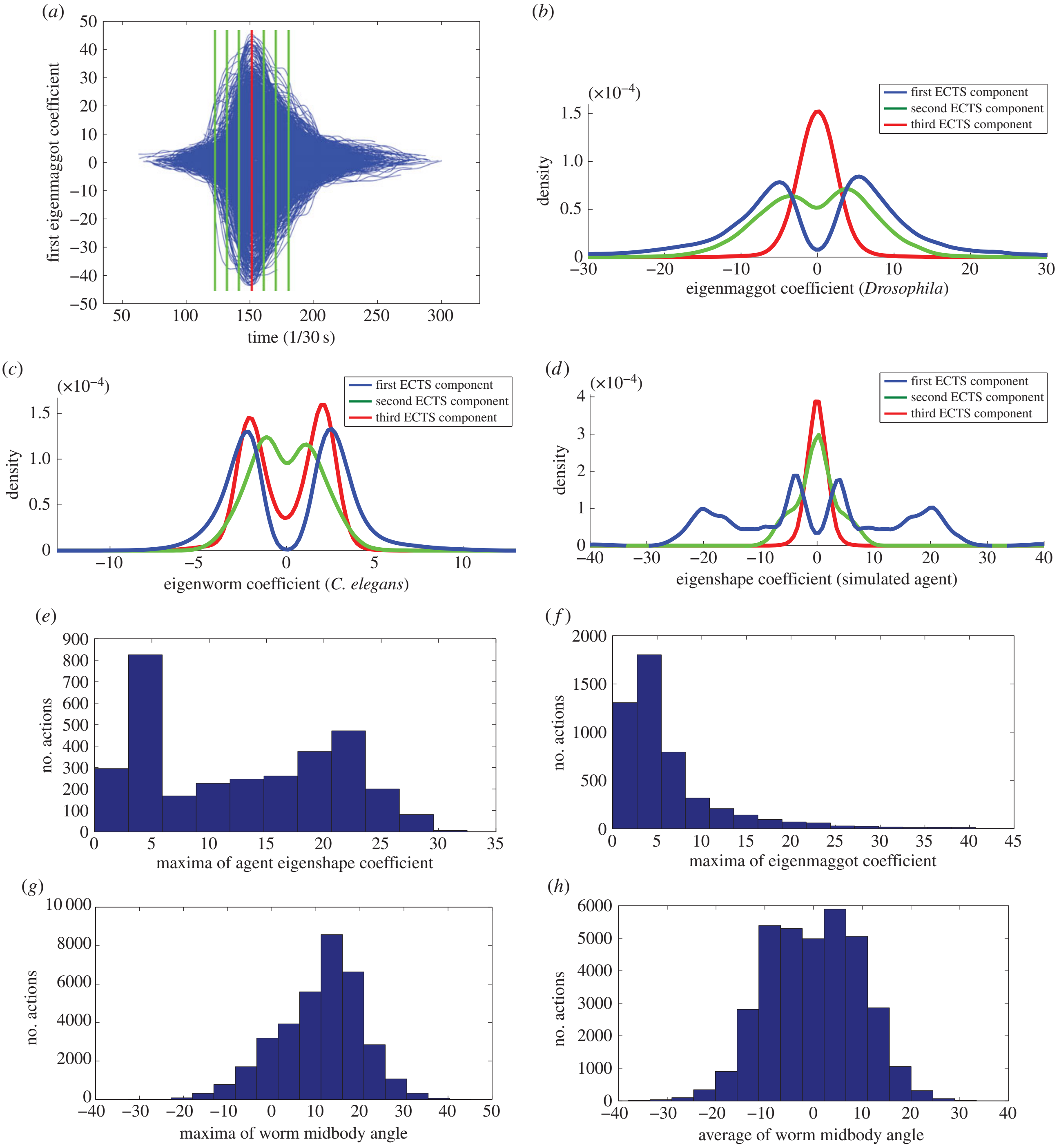

Figure 6. Continuity among behavioural states. Panel (*a*) shows the cross section taken across the aggregated ECTS curves. The cross section across the time of peak curvature (red line on *a*) is shown for *Drosophila* larva, *C. elegans* and the simulated agent on panel (*b–d*), respectively. For clarity, straight runs were removed from the agent's cross section. Panels (*e,f*) show the histogram of the maxima of first ECTS component during actions for the agent and *Drosophila* respectively. For the agent, the bimodal distribution indicates two distinct behaviours, but there is no clear cut-off amplitude for the real organism. Panel (*g,h*) shows the histogram of the maxima and the average of midbody bend for *C. elegans* actions. Again, we do not find a multimodal distributions, indicating that there is no data-defined threshold to distinguish separate behaviours.

behavioural motifs. The first motif we call a *run cast*. A run cast is a low amplitude head cast while the larva is moving approximately straight [28,29]. Successive run casts make up the larva's typical forward locomotion. The second motif corresponds to high amplitude head casts that may or may not be followed by a sharp change of direction. Some previous analyses of larval behaviour distinguish 'stop casts' (or simply 'casts'), where the larva stops locomotion and sweeps its head laterally, from 'turns', which start in a bent body shape and end as the larva resumes locomotion in a new direction [4,9]. This classification scheme is not unique; others have proposed alternatives [30]. We do not find evidence to support the distinction between 'stop casts' and 'turns' instead our analysis describes these behaviours as a single motif, the *turning manoeuvre*. See electronic supplementary material, video S1 and figure 4 for an annotated trajectory and a visualization of the relationship among the motifs.

ESA annotation was evaluated against hand annotation. Across all behaviours, ESA produced an *F*-score of 0.72 (precision = 0.67 and sensitivity = 0.77), where the dominant source of error was a large number of false positive run casts. On the same behavioural experiments, JAABA annotation produced an F-score of 0.68 also with many false positive events. See table 1 for the precision, sensitivity and F-score statistics for each behaviour for both JAABA annotation and ESA. Electronic supplementary material, video S3 shows the binary video of the larva, hand annotation, JAABA and ESA annotations next to each other, so that the reader can gain a good understanding of how the different annotations relate to the larva's behaviour.

Typically, disagreements happen between ESA and hand annotation when an action has high classification uncertainty. Classification uncertainty is quantified by the Shannon entropy [21] and it is denoted by H. Seventy-three per cent of the ESA actions have a low uncertainty, meaning $H < H_{\max}/4$, where $H_{\max} = \log_2 2$, because two states have been found. For these low uncertainty actions, hand annotation and ESA agree on 87%. When classification entropy is high, $H > H_{\max}/4$, then the agreement rate between the two annotations drops to 49%. In short, action labels typically differ where ESA is uncertain. When hand annotation and ESA are in disagreement, it is often debatable which one is correct. In §3.4, we argue that the difficulty to resolve disagreements is due to an unbroken continuity between the two behavioural motifs.

3.3. Motifs for *Caenorhabditis elegans*

ESA was developed with the analysis of larval *Drosophila* in mind, but can also be applied to *C. elegans*. The worm behavioural data were obtained from the CBD. The database contains movies of worms browsing in bacteria, an environment where worms tend to pause for long periods. These pauses required an extra step in the segmentation process, see Methods for details.

In this case, BIC for the spline regression model fit indicated the presence of three behavioural motifs, corresponding to *locomotion*, *turns* and *passive periods*. Segmentation divides locomotion into 'steps', where each step is a $\pi/2$ advancement of the locomotion wave. Multiple locomotion steps make up the characteristic undulatory motion of the worm. The turn behaviour as defined by ESA also includes classic Ω turns, lower amplitude turns and sharp pirouettes [23]. The passive periods are a mixture of pauses, dwelling and quiescence [31]. Figure 5 shows a visualization of the relationship between the motifs and an annotated trajectory, and electronic supplementary material, video S2 provides a dynamic illustration of the annotation.

To benchmark ESA, its performance was compared against hand annotation. ESA produced an F-score of 0.82 (precision = 0.74 and sensitivity = 0.95), where the dominant source of error was a large number of false positive turn events. This finding is not surprising given that the turning behaviour as defined by ESA is very permissive. Existing automated behavioural annotation of the CBD resulted in an F-score of 0.88 (precision = 0.86 and sensitivity = 0.9). See table 2 for the precision, sensitivity and F-score statistics for each behaviour for both CBD annotation and ESA. Furthermore, see electronic supplementary material, video S4, which shows the video of the worm, hand annotation, CBD and ESA annotations next to each other.

As for larval *Drosophila*, there is a significantly increased chance of a *C. elegans* action to be labelled differentially by ESA and hand annotation if the action has a high classification uncertainty ($H > H_{\max}/4$, where $H_{\max} = \log_2 3$ as three behavioural states have been detected) according to ESA. The probability that hand annotation labels these uncertain actions identically decreases to 39% from the population average 77%.

3.4. Do the larva and the worm exhibit discrete behaviours?

For both animals, the above analysis produces a substantial proportion of actions (around 25%) for which classification uncertainty is high. This suggests that the identified behaviours are not discrete, where 'discrete' means clearly distinguishable and stereotypical. Rather we see a continuous spectrum of behaviour. This is in contrast with the overwhelming majority of the literature that treats behaviour of these animals as a set of discrete states, although we are not the first to suggest a continuum among behavioural states for *C. elegans* [31].

To compare our results to what might be expected if there were discrete states, ESA was used to annotate the behaviour of an agent-based simulation of *Drosophila* larva which had been developed independently to study chemotaxis [32]. The agent's behaviour is controlled by a Markov chain model with three states: stop cast, run cast and straight run. Within each state, the precise motion (e.g. body bend) is determined by the current sensory conditions so can vary significantly. Videos were recorded of the agent in its virtual world, and the videos were put through the ESA pipeline (i.e. extracting eigenshape representation, segmentation, clustering). In this way, we test the ESA pipeline for its ability to detect underlying discrete states. We also present several alternative analyses that reveal distinct actions in the simulation but suggest a continuum of actions in the real animals.

3.4.1. Clustering results

For the simulated agent ESA produced three clusters and for 94% of the time, it produced the same behavioural classification as ground truth annotation. BIC indicated a difference between the agent and the animals. For the agent, BIC provided strong evidence to distinguish the three clusters ($\Delta\mathrm{BIC}_{\min} = 7.57$). In contrast, for both *Drosophila* larva and *C. elegans*, there was weak statistical evidence to justify the number of clusters (in both cases $\Delta\mathrm{BIC}_{\min} < 3.75$) [33]. In other words, BIC is confident that there are three distinct clusters among the agent's actions, but for the two animals, the cluster structure is statistically much less justified.

3.4.2. Structure in aggregated eigenshape coefficient time-series segments

We can directly examine this difference in cluster structure by visualizing the presence or absence of clear density bands in the aggregated ECTS subsequences (see §2.9). Sets of stereotypical curves form high-density regions in the cross sections, hence the cross sections can be used to detect stereotypical curves corresponding to a stereotypical posture sequences. Figure 6*a* shows the aggregated ECTS curves for the first ECTS component of larval *Drosophila*. Figure 6*b–d* shows the density cross sections for larval *Drosophila*, *C. elegans* and the agent, respectively. Note that the positive/negative asymmetry of ECTS values along the x-axis corresponds to the left/right asymmetry in larval behaviour and to the dorsal/ventral distinction for *C. elegans*. For both organisms, there is a single band in each half of the *x*-axis. This profile is in contrast with the two distinct bands of the agent's density cross section. The curves forming each high-density band correspond to one Markov state of the agent. Seven cross sections at various *x*-values were examined in each dimension for both the *C. elegans* and *Drosophila* (electronic supplementary material, figure S3), but they all had the same qualitative features as the cross section shown in figure 6, i.e. the animals do not have distinct bands that would support the inference of separable behavioural states.

3.4.3. Structure in behavioural features

Weathervaning, or klinotaxis, is a steering process that results in the animal's trajectory bending towards higher concentration of odour [34]. For *Drosophila* larva, low amplitude head casts are hypothesized to be responsible for weathervaning [29]. These weathervaning casts are distinguished from head casts by the amplitude of body angle [28,29], which is very closely related to the amplitude of the first ECTS component, see figure 2*b*. The agent's behaviour was coded with this distinction in mind, so head casts tend to cause a higher body angle than weathervaning casts. Figure 6*e* shows the histogram of the maxima of first ECTS component during the agent's actions. The bimodal distribution clearly indicates two distinct behaviours. Based on this observation, we examined the maxima and average of a number of features of larval *Drosophila* (head speed, head angle, body angle, body angle speed and head angle speed) and *C. elegans* (eccentricity, head, midbody and tail angles) actions, see figure 6*e*–*h* and electronic supplementary material, figures S4 and S5. We hoped to find multimodal distributions and possibly sharp cut-off values because these could be used as data-defined thresholds to distinguish actions. However, in all cases, a smooth, unimodal distribution was found.

3.4.4. Multidimensional scaling

A final way to examine this issue is to use multidimensional scaling to visualize the distance matrix of actions. DTW was used to measure distance, where the weights are set by the eigenvalue associated with each dimension of ECTS. Figures 4*b* and 5*b* show the larval *Drosophila* and *C. elegans* maps, respectively. As can be seen, there is no clear boundary in either figure to unambiguously separate behavioural motifs. This is in contrast with the agent's map, electronic supplementary material, figure S1, where clearly separated regions can be seen.

4. Discussion

This paper introduces eigenshape annotation, a bottom-up unsupervised method that searches for frequently repeated posture sequences in behavioural data. This problem is closely related to behavioural annotation, but not identical to it. Most behavioural annotators recognize behaviours through user-defined thresholds or training data [2–6]. In both cases, the set of possible behaviours and the description of those behaviours are determined by the user. In contrast, ESA is trying to discover the behavioural states directly from the data without any user input. Note that this task is considerably more challenging than behavioural annotation owing to the lack of *a priori* constraints. Thus, the novelty of this work is to create a data processing pipeline that discovers behavioural motifs in an unsupervised manner, where a behavioural motif is defined as a frequently repeated posture sequence.

The behavioural motifs discovered were generally consistent with behaviours described in the literature. However, many ESA motifs were more permissive than the definitions in other studies. For example, the ESA 'turning manoeuvre' for larva includes turns and high amplitude head casts [4], whereas the ESA 'turning behaviour' for the worm is a mixture of classic and wide Ω turns [6,23]. In both cases, there was no justification in the data for making any further subdivision of turns. Note that it can also be difficult for human observers to distinguish these behaviours consistently.

ESA was also unable to unambiguously classify many actions. The seeming continuity of the action distance maps, figures 4*b* and 5*b*, motivated us to further consider whether there are 'defining features' that could objectively distinguish behaviours. In a simulated agent that was coded with distinct behavioural states, it is straightforward to find such features, for example, the amplitude of body bend (figure 6*e*). We searched for multimodal distributions in a variety of features of the *Drosophila* and *C. elegans* data, but failed in both cases. It remains possible that some feature we did not consider might reveal multimodality, or that discrete behaviours can be distinguished by considering a combination of multiple features.

There is an extensive literature that treats the behaviour of these animals as a set of discrete states. Despite our observation of continuity among behavioural states, our results are not necessarily in contradiction with the discrete treatment of behaviour. Discrete states can be seen as coarse graining (or binning) the continuous behavioural states. For example, the CBD defines Ω turns as a bend greater then $\pi/6$ propagating through the body. If the bend is between $\pi/12$ and $\pi/6$, then the event is called an Υ turn. Thus, this classification scheme treats turning as a two state variable (Ω/Υ turn). In contrast, ESA produces a membership probability that an action is a turn, instead of discretizing non-turns, Υ and Ω turns at arbitrary thresholds. Coarse graining simplifies the underlying postural dynamics, and it can be an appropriate simplification for many studies. For example, the CBD's turn annotation is appropriate for studies looking at the worm's biased random walk. On the other hand, if an analysis requires the precise characterization of the worm's turning behaviour, then the continuous classification scheme of ESA can be advantageous.

However, adopting a coarse-grained description for convenience does not justify the widespread treatment in the research literature of behaviour as actually consisting of a set of discrete states, an assumption that needs to be independently evaluated. There is a risk that initially arbitrary distinctions between behaviours have become reified as qualitatively distinct behaviours of the animal, and treated as a set of actions between which it selects. For example, it is sometimes assumed that the underlying neural activity has a modularity that matches the behavioural states, and that this should guide investigation of neural circuits. In our results, the lack of stereotypical and distinguishable behavioural states suggests that the underlying neural activity is not stereotypical or modular. It remains possible that a highly stereotypical activity pattern of neurons implements a behavioural state, but owing to biomechanical effects, the resulting posture sequences are not so stereotypical. These alternate possibilities can only be addressed by studies of neural activity that do not exclusively depend on behavioural annotators that make *a priori* assumptions about the existence of discrete states.

A further possibility is that the lack of discrete actions observed in our study was a consequence of the particular behavioural conditions in which the animals were tested. Both environments were free of stimulus gradients: larval *Drosophila* was crawling on plain agar, whereas *C. elegans* was browsing in bacteria (although the bacterial layer could have minor inhomogeneities leading to shallow gradients). In future work, we will examine whether the behavioural space changes under different environmental conditions, for example, during directed chemotaxis in larval *Drosophila*.

ESA could be improved by advances in computer vision. Standard thresholding and skeletonizing algorithms fail when the animal intersects itself (2.5). The exclusion of self-intersecting postures introduces a bias to the pipeline, as no posture with self-intersection is included in the analysis. It is a possibility that there are discrete elements of behaviour in the self-intersecting sequences of postures.

The idea behind ESA is to find motifs in behaviour. We represented behaviour as posture, and posture as an ECTS, but the framework presented is not specific to either. ECTS can be replaced with any time series capturing behavioural features, or alternatively ECTS can be supplemented with such time series. Time series of higher-level features provide extra information for the classifier, potentially increasing its accuracy. For example, including a 'direction of locomotion' time series could lead to the detection of reversals as a separate state.

Alternative motif finding algorithms could be used on ECTS as well. For example, the subsequences yielded by segmentation can also be clustered using distance-based methods. We have experimented with several methods [35,36] in combination with standard distance measures (Euclidean and DTW), but it always led to results inferior to spline regression clustering in terms of the classification performance evaluated against hand annotation. We think that the performance difference is due to the ambiguous separation of clusters. Because of its probabilistic nature, spline regression clustering is better equipped to deal with datasets where many of the entries cannot be unambiguously classified.

Finally, we note that motif discovery is a challenging problem and it is an area of intense research in the machine learning community. Owing to the abundance of sequencing data most of the effort is focused on discrete, one-dimensional time series. To the best of our knowledge, the combination of segmentation and clustering is a novel approach to multi-dimensional motif finding. As discussed earlier, the framework is not specific to ECTS, therefore, we expect that with minor modifications the framework could also make contributions in other applications.

Funding. This work was supported by grants nos. EP/F500385/1 and BB/F529254/1 for the University of Edinburgh School of Informatics Doctoral Training Centre in Neuroinformatics and Computational neuroscience from the UK Engineering and Physical Sciences Research Council (EPRSC), UK Biotechnology and Biological Sciences Research Council (BBSRC) and the UK Medical Research Council (MRC) and the FP7 program MINIMAL.

Authors' contributions. B.S. designed the study, implemented the code and wrote the manuscript. A.D. collected and hand annotated the larval data and helped to interpret the results. B.W. supervised the work and helped to write the manuscript.

Competing interests. We declare we have no competing interests.

Acknowledgements. We thank Matthieu Louis and Andre Brown for the thoughtful comments on the manuscript.

Endnote

[1] Code and data used for this publication are available at http://tinyurl.com/p7eculs.

References

1. Marin AG, Paton JJ, Kampff AR, Costa RM, Mainen ZF. 2014 Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nat. Neurosci.* **17**, 1455–1462. (doi:10.1038/nn.3812)
2. Kabra M, Robie AA, Alba MR, Branson S, Branson K. 2013 JAABA: interactive machine learning for automatic annotation of animal behavior. *Nat. Methods* **10**, 64–67. (doi:10.1038/nmeth.2281)
3. Ohyama T, Jovanic T, Denisov G, Dang TC, Hoffmann D, Kerr RA, Zlatic M. 2013 High-throughput analysis of stimulus-evoked behaviors in *Drosophila* larva reveals multiple modality-specific escape strategies. *PLoS ONE* **8**, e71706. (doi:10.1371/journal.pone.0071706)
4. Marin AG, Stephens GJ, Louis M. 2011 Active sampling and decision making in *Drosophila* chemotaxis. *Nat. Commun.* **2**, 441. (doi:10.1038/ncomms1455)
5. Salvador LCM, Bartumeus F, Levin SA, Ryu WS. 2014 Mechanistic analysis of the search behaviour of *Caenorhabditis elegans*. *J. R. Soc. Interface* **11**, 20131092. (doi:10.1098/rsif.2013.1092)
6. Yemini E, Jucikas T, Grundy LJ, Brown A, Schafer WR. 2013 A database of *Caenorhabditis elegans* behavioral phenotypes. *Nat. Methods* **10**, 877–879. (doi:10.1038/nmeth.2560)
7. Stephens GJ, Johnson-Kerner B, Bialek W, Ryu WS. 2008 Dimensionality and dynamics in the behavior of *C. elegans*. *PLoS Comput. Biol.* **4**, e1000028. (doi:10.1371/journal.pcbi.1000028)
8. Fu T-C. 2011 A review on time series data mining. *Eng. Appl. Artif. Intell.* **24**, 164–181. (doi:10.1016/j.engappai.2010.09.007)
9. Green CH, Burnet B, Connolly KJ. 1983 Organization and patterns of inter- and intraspecific variation in the behaviour of *Drosophila* larvae. *Anim. Behav.* **31**, 282–291. (doi:10.1016/S0003-3472(83)80198-5)
10. Berman GJ, Choi DM, Bialek W, Shaevitz JW. 2014 Mapping the stereotyped behaviour of freely moving fruit flies. *J. R. Soc. Interface* **11**, 20140672. (doi:10.1098/rsif.2014.0672)
11. Brown AE, Yemini EI, Grundy LJ, Jucikas T, Schafer WR. 2013 A dictionary of behavioral motifs reveals clusters of genes affecting *Caenorhabditis elegans* locomotion. *Proc. Natl Acad. Sci. USA* **110**, 791–796. (doi:10.1073/pnas.1211447110)
12. Gaffney SJ, Smyth P. 2005 Joint probabilistic curve clustering and alignment. In *Advances in Neural Information Processing Systems 17* (*NIPS 2004*). Cambridge, MA: MIT Press.
13. Gaffney SJ. 2004 Probabilistic curve-aligned clustering and prediction with regression mixture models. PhD thesis, University of California, USA.
14. Schulze A *et al.* 2015 Dynamical feature extraction at the sensory periphery guides chemotaxis. *eLife* **4**, e06694. (doi:10.7554/eLife.06694)
15. MATLAB. 2014 *Version 8.3.0.* (*R2014a*). Natick, MA: The MathWorks Inc.
16. Jolliffe I. 2002 *Principal component analysis*. Hoboken, NJ, USA: Wiley Online Library.
17. Verleysen M, François D. 2005 The curse of dimensionality in data mining and time series prediction. In *IWANN'05 Proc., Computational Intelligence and Bioinspired Systems, Barcelona, Spain, 8–10 June* (eds J Cabestany, A Prieto, F Sandoval), Lecture Notes in Computer Science 3512, pp. 758–770. Berlin, Germany: Springer.
18. Schwarz G. 1978 Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464. (doi:10.1214/aos/1176344136)

19. Fraley C, Raftery AE. 1998 How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **41**, 578–588. (doi:10.1093/comjnl/41.8.578)

20. Konishi S, Kitagawa G. 2008 *Information criteria and statistical modeling*, pp. 211–237. Berlin, Germany: Springer.

21. Shannon C. 2001 A mathematical theory of communication. *ACM Sigmobile Mobile Comput. Commun. Rev.* **5**, 3–55. (doi:10.1145/584091.584093)

22. Powers DM. 2011 Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2**, 37–63.

23. Huang K-M, Cosman P, Schafer WR. 2006 Machine vision based detection of omega bends and reversals in *C. elegans*. *J. Neurosci. Methods* **158**, 323–336. (doi:10.1016/j.jneumeth.2006.06.007)

24. Müller M. 2007 *Information retrieval for music and motion*, pp. 69–84. Berlin, Germany: Springer.

25. Botev ZI *et al.* 2010 Kernel density estimation via diffusion. *Ann. Stat.* **38**, 2916–2957. (doi:10.1214/10-AOS799)

26. Berri S, Boyle JH, Tassieri M, Hope IA, Cohen N. 2009 Forward locomotion of the nematode *C. elegans* is achieved through modulation of a single gait. *HFSP J.* **3**, 186–193. (doi:10.2976/1.3082260)

27. Heckscher ES, Lockery SR, Doe CQ. 2012 Characterization of *Drosophila* larval crawling at the level of organism, segment, and somatic body wall musculature. *J. Neurosci.* **32**, 12 460–12 471. (doi:10.1523/JNEUROSCI.0222-12.2012)

28. Ohashi S, Morimoto T, Suzuki Y, Miyakawa H, Aonishi T. 2014 A novel behavioral strategy, continuous biased running, during chemotaxis in *Drosophila* larvae. *Neurosci. Lett.* **570**, 10–15. (doi:10.1016/j.neulet.2014.04.011)

29. Marin AG, Louis M. 2014 Multilevel control of run orientation in *Drosophila* larval chemotaxis. *Front. Behav. Neurosci.* **8**. (doi:10.3389/fnbeh.2014.00038)

30. Kane EA *et al.* 2013 Sensorimotor structure of *Drosophila* larva phototaxis. *Proc. Natl Acad. Sci. USA* **110**, E3868–E3877. (doi:10.1073/pnas.1215295110)

31. Gallagher T, Bjorness T, Greene R, You Y-J, Avery L. 2013 The geometry of locomotive behavioral states in *C. elegans*. *PLoS ONE* **8**, e59865. (doi:10.1371/journal.pone.0059865)

32. Davies A, Louis M, Webb B. In press. A model of *Drosophila* larvae chemotaxis. *PLoS Comput. Biol.*

33. Kass RE, Raftery AE. 1995 Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795. (doi:10.1080/01621459.1995.10476572)

34. Lockery SR. 2011 The computational worm: spatial orientation and its neuronal basis in *C. elegans*. *Curr. Opin. Neurobiol.* **21**, 782–790. (doi:10.1016/j.conb.2011.06.009)

35. Rodriguez A, Laio A. 2014 Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496. (doi:10.1126/science.1242072)

36. Ankerst M, Breunig MM, Kriegel H-P, Sander J. 1999 Optics: ordering points to identify the clustering structure. *ACM SIGMOD Rec.* **28**, 49–60. From the proceedings of the 1999 ACM Sigmod conference on management of data. (doi:10.1145/304181.304187)