RESEARCH ARTICLE

Open Access

CrossMark

Inactivity periods and postural change speed can explain atypical postural change patterns of *Caenorhabditis elegans* mutants

Tsukasa Fukunaga^{1,2,3*} • and Wataru Iwasaki^{1,4,5*}

Abstract

Background: With rapid advances in genome sequencing and editing technologies, systematic and quantitative analysis of animal behavior is expected to be another key to facilitating data-driven behavioral genetics. The nematode *Caenorhabditis elegans* is a model organism in this field. Several video-tracking systems are available for automatically recording behavioral data for the nematode, but computational methods for analyzing these data are still under development.

Results: In this study, we applied the Gaussian mixture model-based binning method to time-series postural data for 322 *C. elegans* strains. We revealed that the occurrence patterns of the postural states and the transition patterns among these states have a relationship as expected, and such a relationship must be taken into account to identify strains with atypical behaviors that are different from those of wild type. Based on this observation, we identified several strains that exhibit atypical transition patterns that cannot be fully explained by their occurrence patterns of postural states. Surprisingly, we found that two simple factors—overall acceleration of postural movement and elimination of inactivity periods—explained the behavioral characteristics of strains with very atypical transition patterns; therefore, computational analysis of animal behavior must be accompanied by evaluation of the effects of these simple factors. Finally, we found that the *npr-1* and *npr-3* mutants have similar behavioral patterns that were not predictable by sequence homology, proving that our data-driven approach can reveal the functions of genes that have not yet been characterized.

Conclusion: We propose that elimination of inactivity periods and overall acceleration of postural change speed can explain behavioral phenotypes of strains with very atypical postural transition patterns. Our methods and results constitute guidelines for effectively finding strains that show "truly" interesting behaviors and systematically uncovering novel gene functions by bioimage-informatic approaches.

Keywords: Computational ethology, Eigenworms, Worm posture analysis, Transition patterns of postures

Background

While recent advances in DNA sequencing technology have greatly facilitated genomic analysis, quantitative and reproducible analysis of animal behavior is expected to further promote data-driven behavioral genetics [1–4]. *Caenorhabditis elegans* is a model organism for which various research resources are available, including a

high-quality genome sequence, highly curated and integrated databases, and a complete neuronal wiring diagram [5–7]. Several systems that automatically track and videorecord individual worms are already available for ethological studies [8–13]. Some of these systems record not only movement trajectories but also time-series postural images of individual worms. Although these trajectories and postures are not independent from each other [14], perturbations at the molecular and cellular levels influence the latter more directly than the former.

Computational methods for analyzing *C. elegans* timeseries postural data are still under development. A classic

^{*}Correspondence: t.fukunaga@kurenai.waseda.jp; iwasaki@bs.s.u-tokyo.ac.jp
¹Department of Computational Biology and Medical Science, Graduate School of Frontier Sciences, The University of Tokyo, 277-8568 Chiba, Japan
Full list of author information is available at the end of the article



approach is to search given datasets for predefined postural patterns or behavioral parameters; however, such an approach suffers from a lack of objectivity or the ability to identify novel characteristics [15]. A more systematic approach is to use unsupervised machine learning to find frequently appearing stretches or "behavioral motif" de novo within time-series postures. Using this approach, Brown et al. analyzed 7708 movies of 307 mutant strains and detected 2223 C. elegans behavioral motifs [16]. A feature vector for each individual worm was then calculated based on the detected behavioral motifs, and clustering of the strains using these feature vectors successfully grouped mutant strains in which the responsible genes have related biological functions [16]. Szigeti et al. developed another method for finding behavioral motifs based on spline mixture models and identified motifs corresponding to turning or passive behaviors [17].

An alternative systematic approach for analyzing timeseries postural data is to quantify transition frequencies between "postural states". In this approach, worm postures are clustered based on similarities between postures and the obtained clusters are defined as postural states. Whereas the behavioral motif approach detects atypical behaviors as continuous stretches, the postural state approach detects those that rather reflect worms' prompt reaction, which might reflect their decision-making criteria, for instance. As a pioneering work, Schwarz et al. used *K*-means clustering to bin worm postures, and observed condition-specific state transition patterns [18]. However, the factors underlying atypical worm postural movement patterns were not sufficiently dissected, particularly because postures and transition patterns between them should not be independent from each other.

Here, we applied the Gaussian Mixture Model (GMM)-based binning method [19] to time-series postural data for 322 *C. elegans* strains to quantify their transition frequencies between postural states, and revealed that the occurrence patterns of the postural states and the transition patterns among these states have a relationship as expected. In addition, we discovered several strains that exhibit atypical transition patterns that cannot be fully explained by their occurrence patterns of postural states. We also propose that elimination of inactivity periods where the postural change speed is nearly equal to zero, and overall acceleration of postural change speed can explain the behavioral phenotypes of strains with very atypical transition patterns.

Methods

Dataset preparation

The original dataset was obtained from the *C. elegans* behavioral database [20] and consisted of data from 9975 hermaphroditic individual worms of 338 strains freely crawling on agar plate surfaces with food. The 338 strains

comprised 21 wild-type (including N2) and 317 N2derived mutant strains. To concisely represent their postures, we adopted four-dimensional eigenworm vectors [20, 21] that were pre-calculated in the original dataset. In brief, an eigenworm vector was calculated from each image frame of a video-recorded individual worm as follows. First, the midline of the worm body was obtained by image processing, and 48 angles were measured at regular intervals along the midline (Fig. 1). Second, the 48 angles were normalized to obtain a mean value of zero to ignore the general orientation of the body. Third, the normalized 48 values were projected onto four dimensions that were defined by four eigenvectors explaining 92% of the overall variability of worm postures [16, 21]. Such an eigenvector representation of animal shapes is widely accepted for analyzing animal behavior [17, 22].

To ensure that the data were consistent, we excluded data from any individual worm whose video length was not between 890 and 910 seconds or whose eigenworm vectors were missing in more than 40% of the entire frames (typically because of video-tracking failure). Missing eigenworm vectors in the remaining dataset were linearly interpolated using values from the two immediately flanking frames (see Additional file 1: Figure S1A for the proportions of such "gap" frames). Because various frame rates were used in the original dataset (Additional file 1: Figure S1B), we downsampled all data to five frames per second. Finally, by excluding data for any strain for which less than five different individuals were available, we obtained time-series eigenworm vector data for 8769 individual worms from 322 strains (20 wild-type and 302 mutant strains).

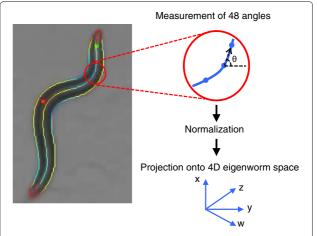


Fig. 1 Measurement and eigenworm representation of worm postures. The *left panel* shows a picture of a wild-type N2 worm; its contour and midline are highlighted. This picture was taken from the *C. elegans* behavioral database [20]. In total, 48 angles were measured at regular intervals along the midline and projected onto the four-dimensional eigenworm space after normalization

Probabilistic binning of *C. elegans* postures into postural states

To represent any eigenworm vector by discrete postural states in a probabilistic manner, we used a GMM-based binning method [19]. This method represents each four-dimensional eigenworm vector by a probabilistic mixture of multiple Gaussian distributions. First, because the total number of image frames in the entire dataset was too large, we randomly sampled 1% of the frames (i.e., 385,790 frames) for parameter estimation. Second, we plotted the eigenworm vectors of all frames and fit the four-dimensional GMM to the pooled distribution consisting of 385,790 data points. The GMM parameters were estimated by the factorized asymptotic Bayes (FAB) algorithm [23]. The FAB algorithm is similar to the conventional expectation-maximization algorithm for fitting GMM [19] but allows automatic estimation of the numbers of mixture components based on factorized information criterion (FIC) [23]. Unlike conventional information criteria such as Akaike information criterion and Bayesian information criterion, FIC can be applied to the inference of mixture models with theoretical justification. The FAB algorithm eliminates components if their mixture ratios are smaller than a given threshold ϵ after the E-step that is modified from that of the conventional expectation-maximization algorithm. The source codes of FAB-GMM algorithm are available at https://github.com/fukunagatsu/FAB-GMM. We set $\epsilon = 0.01, 0.005$, or 0.001 using initial parameter sets estimated by the K-means++ algorithm [24] with K = 100,200, or 1000, respectively (These K values are the maximum numbers of states for each ϵ). Third, after the convergence of the FAB algorithm, each Gaussian distribution component was regarded as a postural state. We obtained 44, 95, and 459 postural states when ϵ was 0.01, 0.005, and 0.001, respectively. Finally, for each frame (including the remaining 99%), the responsibility of each Gaussian distribution component for explaining its eigenworm vector was calculated using the estimated parameters. As a result, a posture of any individual i at any frame f was represented by a K-dimensional nonnegative vector $\mathbf{r}_{i,f} = (r_{i,f,1}, r_{i,f,2}, \dots, r_{i,f,K})^{\mathrm{T}}$, where K is the number of postural states, each element represents the responsibility of each state, and $\sum_{k=1}^{K} r_{i,f,k} = 1.$

For comparison, we also adopted the K-means clustering, which not probabilistically but deterministically bins any eigenworm vector to a single state. First, the same set of 1% frames were selected and their eigenworm vectors were plotted to the four-dimensional space in the same manner. Second, K-means clustering was applied to the pooled distribution. The model parameters were estimated by the Lloyd algorithm [25] with initial parameters estimated by K-means++ algorithm [24], where K

was set to 90 (a parameter used in a previous study [18]) or 44, 95, or 459 (parameters estimated by the GMM-based method earlier). Third, after the convergence of the Lloyd algorithm, a centroid of each cluster was regarded as a postural state. Finally, for each frame (including the remaining 99%), its eigenworm vector was binned to the closest postural state. Note that, any worm posture was represented by a vector $\mathbf{r}_{i,f}$ as in the case of the GMM-based method, but it was an integer vector (i.e., only one of its elements was 1 and the others were 0).

Evaluation of binning methods

In this work, we assumed that the postures of individual worms belonging to the same strain should be statistically more similar than the postures of worms belonging to different strains. Thus, if worm postures are represented more properly by the postural states, the relative state occurrence frequencies of an individual i, $\mathbf{r}_i = \frac{1}{F} \sum_{f=1}^{F} \mathbf{r}_{i,f}$, where F is the number of frames, are expected to be more similar between individuals of the same strain than those of different strains. The GMM-based method and K-means clustering with three and four different parameters, respectively, were compared based on this rationale.

Let S_i be a set of individuals that belong to the same strain as i except i, and \overline{S}_i be a set of randomly selected individuals such that $S_i \cap \overline{S}_i = \emptyset$, $i \notin \overline{S}_i$, and $|S_i| = |\overline{S}_i|$, where |S| represents the number of individuals belonging to S. For every individual i, we calculated the mean divergences of the relative state occurrence frequencies against S_i and \overline{S}_i as follows:

$$\Delta_{\text{intra}}\mathbf{r}_i = \frac{1}{|S_i|} \sum_{i \in S_i} d(\mathbf{r}_i, \mathbf{r}_j)$$

$$\Delta_{\text{inter}}\mathbf{r}_i = \frac{1}{|\overline{S}_i|} \sum_{j \in \overline{S}_i} d(\mathbf{r}_i, \mathbf{r}_j)$$

where d is the Jensen-Shannon divergence, which is a measure of divergence between two probability distributions [26]. Note that \mathbf{r}_i is a normalized vector and can be regarded as a probability distribution. Then, for each strain, we tested the hypothesis that $\Delta_{\text{intra}}\mathbf{r}_i$ of all individuals that belong to that strain are statistically smaller than $\Delta_{\text{inter}}\mathbf{r}_i$ of them by a one-sided Wilcoxon-Mann-Whitney test. The test was repeated against all strains, and the Benjamini-Hochberg approach was used to control the false-discovery rates of multiple testing (q < 0.05) [27].

Discovery of strains showing wild-type N2-like postures but atypical transition patterns

For each individual i, the relative transition frequencies between postural states were represented by a $K \times K$ matrix T_i whose element representing the transition from a state k to l is

$$T_{i,k,l} = \frac{1}{F-1} \sum_{f=1}^{F-1} r_{i,f,k} r_{i,f+1,l}$$

For each strain S, the relative state occurrence frequency \mathbf{r}_i and relative state transition frequency T_i were averaged for its individuals to obtain \mathbf{r}_S and T_S , respectively. With the exception of the wild-type N2 strain, we calculated the divergences of each strain from wild-type N2 as follows:

$$\Delta_{N2}\mathbf{r}_S = d(\mathbf{r}_S, \mathbf{r}_{N2})$$

$$\Delta_{N2}T_S = d(T_S, T_{N2})$$

where d is the Jensen-Shannon divergence. Note that T_i can also be regarded as a probability distribution. For example, a large $\Delta_{\rm N2}T_S$ indicates that strain S has a state transition pattern that is very different from that of wild-type N2.

Next, we conducted linear regression to investigate relationships between a dependent variable $\Delta_{\rm N2}T_{\rm S}$ and an explanatory variable $\Delta_{\rm N2}{\bf r}_{\rm S}$. Then, we detected strains showing atypical transition patterns using standardized residuals (Z-values) from the estimated linear model. To control the false-discovery rates of multiple testing, we used the Benjamini-Hochberg approach (q < 0.05).

Analysis of factors underlying atypical state transition patterns

To reveal factors underlying the atypical transition patterns, we created artificial N2 strains in silico by modifying the eigenworm data for the wild-type N2 strain and determined if these artificial N2 strains reproduced the atypical state transition frequencies of strains that showed atypical transition patterns. Specifically, we focused on the effects of eliminating inactivity periods and accelerating the average postural change speed. To remove inactivity periods from wild-type N2, we excluded any frame for which the Euclidean distance of the eigenworm vectors between that and the previous frame was smaller than a threshold α . To change the postural change speed as a whole, we removed frames at regular intervals to simulate movement of β -times accelerated wild-type N2. When $\beta = 1.5$, for every three consecutive frames, the eigenworm vectors of the second and third frames were replaced with the averaged vector. When $\beta = 2$, every second frame was removed. Because all strains that showed atypical transition patterns were faster than wild-type N2 on average, only acceleration was considered here (i.e., deceleration was not considered here). For each strain S, the parameters α and β were selected to minimize

$$D_{\text{eigenworm speed}} = \int_{x} |F_{aN2}(x) - F_{S}(x)| dx$$

where F_{aN2} and F_S are the cumulative distributions of the instantaneous postural change speed (Euclidean distance

between eigenworm vectors of adjacent frames) of the artificial N2 strain and *S*, respectively.

Then, we calculated \mathbf{r}_{aN2} and T_{aN2} , which are the relative state occurrence frequency and relative state transition frequency, respectively, of the artificial N2 strain. To determine whether the artificial N2 strain reproduced the behavioral characteristics of the strain S, $\Delta_{aN2}\mathbf{r}_S$ and $\Delta_{aN2}T_S$. In addition, we calculated the standardized residuals of $\Delta_{aN2}\mathbf{r}_S$ and $\Delta_{aN2}T_S$ based on the previously predicted linear model. We called this standardized residual Z_a .

Results

Evaluation of binning methods and parameters

The postures of each of the 8769 individual worms belonging to 322 strains were represented by time-series fourdimensional eigenworm vectors. Every eigenworm vector was binned to postural states by the GMM-based method and K-means clustering with three and four parameters, respectively. Then, we determined if the relative state occurrence frequencies were more similar between worms of the same strain than between worms of different strains. The number of strains for which the null hypothesis of no difference with Benjamini-Hochberg's q < 0.05was rejected is shown in Table 1. Overall, the GMM-based method detected greater numbers of such strains than Kmeans clustering, indicating that less postural information was lost during binning in the former than in the latter. Although the parameter selection did not have a strong impact on the results, K = 95 and $\epsilon = 0.005$ were the best parameters for the K-means clustering and GMM-based methods, respectively. Among the 213 strains that exhibited significance by K-means-clustering with K = 95, only ten were missed by the GMM-based method with $\epsilon = 0.005$ (Additional file 1: Figure S2). Therefore, the GMM-based method with $\epsilon = 0.005$ was adopted for postural state binning in the following analyses.

Table 1 Evaluation of binning methods and parameters

Binning method	Parameter	Number of strains
K-means	K = 44	207
	K = 90	203
	K = 95	213
	K = 459	205
GMM	$\epsilon = 0.01$	238
	$\epsilon = 0.005$	242
	$\epsilon = 0.001$	229

Shown are the numbers of strains whose relative state occurrence frequencies were significantly more similar between those of the same strain than those of different strains (one-sided Wilcoxon-Mann-Whitney test with the Benjamini-Hochberg procedure (q < 0.05)). The bold values are the highest scores

Strong relationships between postural state occurrences and transitions

After the binning of eigenworm vectors to the postural states, the relative state occurrence frequency \mathbf{r}_S and relative state transition frequency T_S of each strain S were calculated. Figure 2 shows their divergences from those of the wild-type N2 strain, where large $\Delta_{\rm N2}\mathbf{r}_S$ and $\Delta_{\rm N2}T_S$ indicate that strain S displays postures and transition patterns that are very different from those of wild-type N2. We clearly observed a general trend of a positive linear correlation between the two divergence values (The adjusted R-squared value was 0.96). This likely reflected the fact that the use of different postures naturally leads to the use of different transition patterns. Note that $\Delta_{\rm N2}T_S - \Delta_{\rm N2}\mathbf{r}_S \geq 0$ (the proof is provided in the Additional file 1).

For example, three mutant strains, unc-103, unc-1(e1598), and unc-77(e625), exhibited the largest divergences of both values from wild-type N2 (Fig. 2). The *unc-103* gene encodes an ether-a-go-go-related K^+ channel homolog, and the strain in which this gene has a gain-of-function mutation has been reported to show extremely lethargic behavior [28]. The unc-1(e1598) strain is a mutant of a stomatin-like-protein gene and has also been reported to show very slow behavior [29]. The large deviations of the postural state occurrence and transition frequencies of these two mutant strains likely reflect their exceptionally inactive phenotypes. The *unc-77(e625)* strain features a gain-of-function mutation of a subunit gene of a voltage-insensitive cation leak channel and exhibits coiled postures [29]. To reveal what postures are specific in this strain, we calculated the fold change between $\mathbf{r}_{unc-77(e625),k}$ and $\mathbf{r}_{N2,k}$ for each postural state

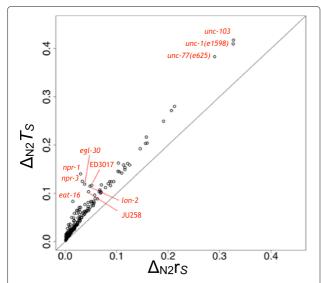


Fig. 2 Divergences of postural state occurrence and transition frequencies of 321 non wild-type N2 strains. The x- and y-axis represent $\Delta_{\rm N2} {\bf r}_{\rm S}$ and $\Delta_{\rm N2} T_{\rm S}$, respectively

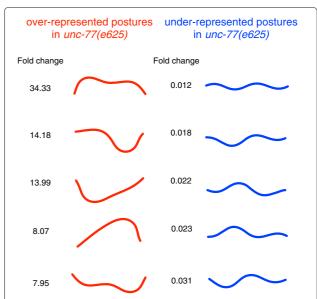


Fig. 3 Top five over-represented and under-represented postures in *unc-77(e625)* Each posture was reconstructed from the mean value of the corresponding postural states. *Red* and *blue* colors represent over-represented and under-represented postures in *unc-77(e625)*, respectively

k and detected over-represented and under-represented postures in unc-77(e625) (Fig. 3). These results showed that the unc-77(e625) strain tends to take more "C-shaped" but less "S-shaped" postures compared to the wild-type N2 strain.

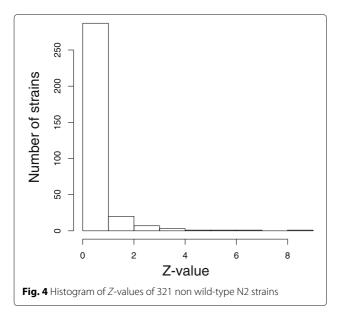
Identification of strains exhibiting atypical transition patterns

As shown in Fig. 2, although most strains strongly followed the positive linear correlation trend, several strains did not. We identified seven strains exhibiting atypical transition patterns that were significantly deviated from expectation (q < 0.05, left side of Fig. 2). Only these seven strains showed Z-values larger than 3.0 (Table 2, Fig. 4).

The two strains with the largest Z-values, *npr-1* and *npr-3*, are mutants of neuropeptide receptor (*npr*) genes. As

Table 2 Strains exhibiting wild-type N2-like postures but atypical transition patterns

transition patterns						
Strain	$\Delta_{N2}\mathbf{r}_{S}$	$\Delta_{N2} \mathcal{T}_{S}$	Z-value			
npr-1	0.0299	0.1404	8.192			
npr-3	0.0333	0.1249	6.423			
egl-30	0.0376	0.1192	5.436			
eat-16	0.0149	0.0831	4.872			
Ion-2	0.0481	0.1150	3.876			
ED3017	0.0514	0.1167	3.632			
JU258	0.0453	0.1033	3.163			



it is known that NPR-1 is expressed in ventral nerve cord motor neuron, it is reasonable that these neuropeptide receptor genes have roles in controlling postural movement [30, 31]. However, notably, other npr mutant strains did not show large Z-values, even though npr-1 and npr-3 are neither the closest paralogs in the npr gene family nor highly identical in sequence (amino-acid sequence identity% = 25.5) (Additional file 1: Figure S3). Because \mathbf{r}_S and T_S of these two strains were most similar to each other among all strains (i.e., the differences of \mathbf{r}_S and T_S between the two strains were the smallest among every strain pair that contains either of the npr-1 and npr-3 strains), the npr-1 and npr-3 genes were suggested to have closely related functions at the behavioral level regardless of their different evolutionary origins at the sequence level.

The egl-30 and eat-16 genes encode components of heterotrimeric G-protein signaling pathways. Loss of EGL-30 function depresses the behavioral activity of C. elegans, whereas EAT-16 negatively regulates EGL-30 [32, 33]. Because the egl-30 and eat-16 mutant strains in this study have gain- (ep271) and loss-of-function alleles (sa609), respectively [33, 34], their similar, active behavioral phenotypes are consistent with previous reports. Indeed, \mathbf{r}_S and T_S from the egl-30 and eat-16 strains were most similar to each other.

lon-2 encodes a glypican-family protein of heparan sulfate proteoglycans, and its mutant was reported to have a longer body than that of wild-type N2 [35]. A previous study reported that lon-2 was one of the worst-fit mutants in the eigenworm representation [16]. Although it is not clear why $\Delta_{\rm N2}{\bf r}_S$ of lon-2 is not very large, the poor fitting of the eigenworm representation may have resulted in atypical transition patterns of this strain.

ED3017 and JU258 are non-N2 wild-type strains. *C. elegans* population genomics studies revealed that N2 strains acquired gain-of-function mutations in *npr-1* during laboratory domestication [36, 37], and ED3017 and JU258 have a lower activity allele in *npr-1*. The large Z-values of these two strains may be caused by this low *npr-1* activity.

Analysis of factors underlying atypical state transition patterns

Figure 5a presents the distributions of the instantaneous postural change speed of wild-type N2 and the six strains that exhibited atypical transition patterns. Note that *lon-2* was excluded here because the earlier stage of eigenworm representation could be problematic for this strain. Overall, all six strains exhibited faster postural change speeds than those of wild-type N2. Only wild-type N2 had a mode of the postural change speed at approximately 0.1 (Fig. 5b), at which four of the other strains also had small

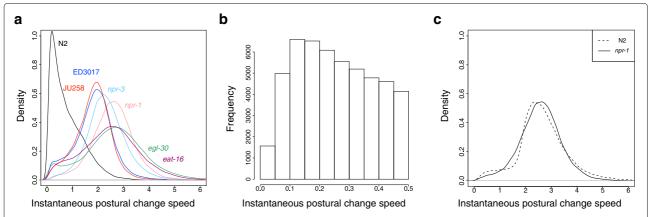


Fig. 5 Distributions of instantaneous speed of postural change. **a** The distributions of wild-type N2, *npr-1*, *npr-3*, *egl-30*, *eat-16*, ED3017, and JU258. The *y*-axis represents density. **b** A histogram that magnifies around the mode of the wild-type N2 distribution. **c** The distributions of *npr-1* and the artificial N2 strain whose postural change speed resembles that of *npr-1*

"shoulders" (Fig. 5a). Such a small speed value indicates that the individuals are under inactivity periods, which may correspond to *quiescence* worm behavior [38]. We also observed several strains that retain this mode of postural change speed at approximately 0.1 but have different distribution shapes from that of wild-type N2 (e.g., *unc-43* and *C52B9.11*; Additional file 1: Figure S4).

On the basis of these observations, we investigated whether artificial elimination of the inactivity periods and overall acceleration of postural change speed from wild-type N2 could reproduce the state transition patterns of these six strains without significantly altering the state occurrence frequencies. In time-series sequence representations of postural states, inactivity periods are represented by stretches of identical or similar state(s). Because inactivity periods likely do not occur only at specific postural states (the Jensen-Shannon divergence of relative state occurrence frequencies between frames whose posture change speed was less than and greater than 0.3 was 0.0161 for wild-type N2), the elimination of inactivity periods would change the state transition frequencies while modestly preserving the state occurrence frequencies. For the acceleration of postural change speed, for example, two-fold acceleration of a state sequence AABBCCAA... into ABCA... as a whole will also change the state transition frequencies by preserving the state occurrence frequencies.

Artificial N2 strain data were created in silico by removing inactivity periods with threshold α and accelerating β -fold as a whole from the time-series eigenworm data for wild-type N2. For each of the six strains, we chose the best parameters from $\alpha = 0.3, 0.4, \dots, 1.0$ and $\beta = 1.5, 2.0$ by examining whether the postural change speed distributions of the artificial N2 strains fit those of each of the six strains (Fig. 5c, Additional file 1: Figure S5, and Additional file 1: Table S1). Finally, we investigated whether the artificial N2 strains reproduced not only the distributions of the instantaneous postural change speed but also the postural state occurrence and transition frequencies of the six strains. All Z_a values became substantially smaller than the original Z-values, and these values decreased to the level that was not significantly different from expectation (Table 3). In other words, the atypical state transition frequencies of these four mutant strains can be explained almost entirely by the lack of inactivity periods and overall acceleration of the postural change speed.

Discussion

In this study, we used the GMM-based method for probabilistically binning worm postures into a finite number of postural states and revealed an apparent relationship between the postures and transition patterns of *C. elegans* strains. The superior binning performance of the GMM-based method reflects the fact that time-series postures

Table 3 Reproduction of atypical state transition frequencies by artificially modified N2

,				
Strain	$\Delta_{aN2}\mathbf{r}_{S}$	$\Delta_{aN2} T_{S}$	Za	Original Z-value
npr-1	0.0135	0.0314	0.484	8.192
npr-3	0.0180	0.0426	0.953	6.423
egl-30	0.0235	0.0494	0.921	5.436
eat-16	0.0147	0.0307	0.285	4.872
ED3017	0.0351	0.0695	1.360	3.632
JU258	0.0290	0.0577	1.021	3.163

of any individual are distributed along a single trajectory in the four-dimensional eigenworm space because the postures of continuous frames should be similar to each other. Thus, a worm must adopt intermediate postures while changing its posture from one postural state to another. Deterministic binning of such intermediate postures inevitably loses information or introduces noise to the representation of worm behavior. The case of the lon-2 strain in this study also indicates the importance of preserving information during the computational analysis of animal behavior, although the difficulty in this case occurred during the eigenworm representation. The strong relationship between the postural state occurrence and transition frequencies offers two important suggestions for worm postural movement analysis: a significant part of postural movement variations can be evaluated solely by examining postures without temporal information, and the effects of using different postures must be taken into account in postural movement analysis.

Several strains that exhibited atypical transition patterns among postural states were identified. Surprisingly, for the six strains that showed the most atypical postural movement, merely eliminating the inactivity periods and accelerating the postural change speed as a whole nearly reproduced their atypical transition patterns. While quantification of the transition frequencies between postural states is a powerful approach for computationally analyzing animal behavior, our results demonstrate that even very atypical state transition patterns can result from simple factors. Analyses of inactivity periods and postural change speeds both require consideration of time duration; the compression of state time duration abolishes their effects in the analysis [18]. To effectively detect strains that show truly interesting behavior, e.g., strains whose neural circuits encode special decision-making criteria, computational analyses of animal behavior must be accompanied by evaluation of the effects of more "trivial" factors such as overall change in speed (of course, we note that these trivial factors themselves would also provide many insights into worm behavior). The C. elegans

behavioral database also contains various behavioral data such as dorsal/ventral orientations, velocities, and trajectories during worm movement. Using these additional datasets, we may dissect factors that underlie interesting phenotypes more deeply, for example, effects of dorsal/ventral biases in postural change patterns and/or relationships between postural change patterns and movement trajectories.

Our analysis also revealed that the *npr-1* and *npr-3* genes have closely related functions that were unpredictable by sequence homology, the most basic principle in this genomic era. Many studies have conducted functional analyses of *npr-1* [31, 39–41], but few studies have focused on *npr-3* [42]. Therefore, we envision that existing knowledge about *npr-1* will substantially accelerate future functional analyses of *npr-3* based on the present result.

In this study, divergence of the state occurrence and transition frequencies from the wild-type N2 strain was examined. Although this would make sense for the analysis of N2-derived mutant strains, comparison among different wild-type strains can also be done by selecting another strain as a reference. We expect that the linear correlation trend between the state occurrence and transition frequencies will be recovered regardless of the reference strain choice; however, for example, it would also be of interest to select strains that have specific evolutionary context or strains that show characteristic behavior (such as ED3017 or JU258).

Conclusions

Although more than a decade has passed since the genomes of many model organisms were sequenced, significant numbers of genes remain functionally uncharacterized. Systematically deciphering their functions beyond straightforward sequence homology analysis is one of the most important goals in computational biology today, where an advantage of bioimage informatics for functional analysis is the ability of this method to directly evaluate phenotypes [43-45]. Finally, it should be noted that genome-editing technologies are enabling rapid construction of genetically engineered animal strains [46]. Bioimage-informatic analysis of their behaviors will, for example, contribute to the identification of novel genes responsible for neurological disease. We emphasize that further development of computational methods and accumulation of technical knowledge will be critical to promote this emerging field.

Additional file

Additional file 1: Supplementary materials. This file includes additional texts, figures and tables not shown in the manuscript. (PDF 477 kb)

Abbreviations

FAB: Factorized asymptotic Bayes; FIC: Factorized information criteria; GMM: Gaussian mixture model

Acknowledgments

We thank Haruka Ozaki and Hirotaka Matsumoto for critically reading the manuscript.

Funding

This study was supported by the Japan Society for the Promotion of Science [grant numbers 15J07635, 16J00129 and 16H06154], the CREST Program from the Japan Science and Technology Agency, and the Canon Foundation.

Availability of data and material

Worm posture dataset are available in the *C. elegans* behavioral database [20]. In addition, the source codes of FAB-GMM algorithm can be downloaded from https://qithub.com/fukunaqatsu/FAB-GMM.

Authors' contributions

TF and WI designed the project. TF performed the analyses. TF and WI wrote the paper. Both the authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Ethics approval and consent to participate

Not applicable.

Author details

Department of Computational Biology and Medical Science, Graduate School of Frontier Sciences, The University of Tokyo, 277-8568 Chiba, Japan.
 Faculty of Science and Engineering, Waseda University, 169-0072 Tokyo, Japan.
 Research Fellow of Japan Society for the Promotion of Science, Tokyo, Japan.
 Atmosphere and Ocean Research Institute, The University of Tokyo, 277-8564 Chiba, Japan.
 Department of Biological Sciences, Graduate School of Science, The University of Tokyo, 113-0032 Tokyo, Japan.

Received: 22 August 2016 Accepted: 6 December 2016 Published online: 19 January 2017

References

- Dankert H, Wang L, Hoopfer ED, Anderson DJ, Perona P. Automated monitoring and analysis of social behavior in *Drosophila*. Nat Methods. 2009;6:297–303.
- Anderson DJ, Perona P. Toward a science of computational ethology. Neuron. 2014;84:18–31.
- Pérez-Escudero A, Vicente-Page J, Hinz RC, Arganda S, de Polavieja GG. idTracker: tracking individuals in a group by automatic identification of unmarked animals. Nat Methods. 2014:11:743–8.
- Fukunaga T, Kubota S, Oda S, Iwasaki W. GroupTracker: video tracking system for multiple animals under severe occlusion. Comput Biol Chem. 2015;57:39–45.
- C. elegans Sequencing Consortium T. Genome sequence of the nematode C. elegans: A platform for investigating biology. Science. 1998;282:2012–8.
- Harris TW, Baran J, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, Done J, Grove C, Howe K, et al. Wormbase 2014: new views of curated biology. Nucleic Acids Res. 2014;42:789–93.
- White J, Southgate E, Thomson J, Brenner S. The structure of the nervous system of the nematode *Caenorhabditis elegans*: the mind of a worm. Philos Trans R Soc B: Biol Sci. 1986;314:1–340.
- Baek JH, Cosman P, Feng Z, Silver J, Schafer WR. Using machine vision to analyze and classify *Caenorhabditis elegans* behavioral phenotypes quantitatively. J Neurosci Methods. 2002;118(1):9–21.
- Ramot D, Johnson BE, Berry Jr TL, Carnell L, Goodman MB. The Parallel Worm Tracker: a platform for measuring average speed and drug-induced paralysis in nematodes. PLOS ONE. 2008;3(5):2208.

- Swierczek NA, Giles AC, Rankin CH, Kerr RA. High-throughput behavioral analysis in C. elegans. Nat Methods. 2011;8(7):592–8.
- Cronin CJ, Mendel JE, Mukhtar S, Kim YM, Stirbl RC, Bruck J, Sternberg PW. An automated system for measuring parameters of nematode sinusoidal movement. BMC Genet. 2005;6(1):5.
- Nagy S, Wright C, Tramm N, Labello N, Burov S, Biron D. A longitudinal study of *Caenorhabditis elegans* larvae reveals a novel locomotion switch, regulated by gαs signaling. eLife. 2013;2:00782.
- Nagy S, Goessling M, Amit Y, Biron D. A generative statistical algorithm for automatic detection of complex postures. PLOS Comput Biol. 2015;11(10):1004517.
- Stephens GJ, Johnson-Kerner B, Bialek W, Ryu WS. From modes to movement in the behavior of *Caenorhabditis elegans*. PLOS ONE. 2010;5(11):13914.
- Dell Al, Bender JA, Branson K, Couzin ID, de Polavieja GG, Noldus LP, Pérez-Escudero A, Perona P, Straw AD, Wikelski M, et al. Automated image-based tracking and its application in ecology. Trends Ecol Evol. 2014;29:417–28.
- Brown AE, Yemini El, Grundy LJ, Jucikas T, Schafer WR. A dictionary of behavioral motifs reveals clusters of genes affecting *Caenorhabditis* elegans locomotion. Proc Natl Acad Sci. 2013;110(2):791–6.
- Szigeti B, Deogade A, Webb B. Searching for motifs in the behaviour of larval *Drosophila melanogaster* and *Caenorhabditis elegans* reveals continuity between behavioural states. J R Soc Interface. 2015;12(113): 20150899
- Schwarz RF, Branicky R, Grundy LJ, Schafer WR, Brown AE. Changes in postural syntax characterize sensory modulation and natural variation of C. elegans locomotion. PLOS Comput Biol. 2015;11(8):1004322.
- Bishop CM, Nasrabadi NM, Vol. 1. Pattern Recognition and Machine Learning. New York: Springer; 2006.
- Yemini E, Jucikas T, Grundy LJ, Brown AE, Schafer WR. A database of Caenorhabditis elegans behavioral phenotypes. Nat Methods. 2013;10(9): 877–9
- Stephens GJ, Johnson-Kerner B, Bialek W, Ryu WS. Dimensionality and dynamics in the behavior of *C. elegans*. PLOS Comput Biol. 2008;4(4): 1000028.
- 22. Girdhar K, Gruebele M, Chemla YR. The behavioral space of zebrafish locomotion and its neural network analog. PLOS ONE. 2015;10(7):0128668.
- Fujimaki R, Morinaga S. Factorized asymptotic bayesian inference for mixture modeling. International Conference on Artificial Intelligence and Statistics. 2012400–408.
- Arthur D, Vassilvitskii S. K-means++: the advantages of careful seeding. ACM-SIAM symposium on Discrete algorithms. 20071027–1035.
- Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory. 1982;28(2):129–37.
- Lin J. Divergence measures based on the shannon entropy. IEEE Trans Inf Theory. 1991;37(1):145–51.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.
- 28. Park EC, Horvitz HR. Mutations with dominant effects on the behavior and morphology of the nematode *Caenorhabditis elegans*. Genetics. 1986;113(4):821–52.
- Huang KM, Cosman P, Schafer WR. Machine vision based detection of omega bends and reversals in *C. elegans*. J Neurosci Methods. 2006;158(2):323–36.
- Wang Q, Wadsworth WG. The C domain of netrin UNC-6 silences calcium/calmodulin-dependent protein kinase-and diacylglycerol-dependent axon branching in *Caenorhabditis elegans*. J Neurosci. 2002;22(6):2274–82.
- 31. Coates JC, de Bono M. Antagonistic pathways in neurons exposed to body fluid regulate social feeding in *Caenorhabditis elegans*. Nature. 2002;419(6910):925–9.
- Brundage L, Avery L, Katz A, Kim UJ, Mendel JE, Sternberg PW, Simon MI. Mutations in a C. elegans Gqα gene disrupt movement, egg laying, and viability. Neuron. 1996;16(5):999–1009.
- Hajdu-Cronin YM, Chen WJ, Patikoglou G, Koelle MR, Sternberg PW. Antagonism between Goα and Gqα in Caenorhabditis elegans: the RGS protein EAT-16 is necessary for Goα signaling and regulates Gqα activity. Gene Dev. 1999;13(14):1780–93.

- Fitzgerald K, Tertyshnikova S, Moore L, Bjerke L, Burley B, Cao J, Carroll P, Choy R, Doberstein S, Dubaquie Y, et al. Chemical genetics reveals an RGS/G-protein role in the action of a compound. PLOS Genet. 2006;2(4):57.
- 35. Gumienny TL, MacNeil LT, Wang H, de Bono M, Wrana JL, Padgett RW. Glypican LON-2 is a conserved negative regulator of bmp-like signaling in *Caenorhabditis elegans*. Curr Biol. 2007;17(2):159–64.
- Rockman MV, Kruglyak L. Recombinational landscape and population genomics of *Caenorhabditis elegans*. PLOS Genet. 2009;5(3):1000419.
- Weber KP, De S, Kozarewa I, Turner DJ, Babu MM, de Bono M. Whole genome sequencing highlights genetic changes associated with laboratory domestication of *C. elegans*. PLOS one. 2010;5(11):13922.
- 38. Gallagher T, Bjorness T, Greene R, You YJ, Avery L. The geometry of locomotive behavioral states in *C. elegans*. PLOS ONE. 2013;8(3):59865.
- De Bono M, Bargmann CI. Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in C. elegans. Cell. 1998:94(5):679–89.
- Choi S, Chatzigeorgiou M, Taylor KP, Schafer WR, Kaplan JM. Analysis of NPR-1 reveals a circuit mechanism for behavioral quiescence in *C. elegans*. Neuron. 2013;78(5):869–80.
- Cheung BH, Cohen M, Rogers C, Albayram O, de Bono M. Experience-dependent modulation of *C. elegans* behavior by ambient oxygen. Curr Biol. 2005;15(10):905–17.
- Kubiak TM, Larsen MJ, Zantello MR, Bowman JW, Nulf SC, Lowery DE. Functional annotation of the putative orphan *Caenorhabditis elegans* G-protein-coupled receptor C10C6.2 as a FLP15 peptide receptor. J Biol Chem. 2003;278(43):42115–20.
- 43. Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, Saito TL, Saka A, Fukuda T, Ishihara S, Oka S, et al. High-dimensional and large-scale phenotyping of yeast mutants. Proc Natl Acad Sci. 2005;102(52):19015–20.
- Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nat Rev Genet. 2010;11(12):855–66.
- Yu H, Aleman-Meza B, Gharib S, Labocha MK, Cronin CJ, Sternberg PW, Zhong W. Systematic profiling of *Caenorhabditis elegans* locomotive behaviors reveals additional components in G-protein Gαq signaling. Proc Natl Acad Sci. 2013;110(29):11940–5.
- Friedland AE, Tzur YB, Esvelt KM, Colaiácovo MP, Church GM, Calarco JA. Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. Nat Methods. 2013:10:741–3.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at www.biomedcentral.com/submit

