A Posture-Based Markov Analysis of Behavioral States in *Caenorhabditis elegans*

by

Rebekah I. Clark

S.B. Electrical Engineering and Computer Science, MIT (2018)
S.B. Brain and Cognitive Science, MIT (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2019

Author . . . .
Department of Electrical Engineering and Computer Science
June 6, 2019

Certified by . . . .
Steven W. Flavell, Ph.D.
Assistant Professor
Thesis Supervisor
June 6, 2019

Accepted by . . . .
Katrina LaCurts, Ph.D.
Master of Engineering Thesis Committee
June 6, 2019

A Posture-Based Markov Analysis of Behavioral States in *Caenorhabditis elegans*

by

Rebekah I. Clark

Submitted to the Department of Electrical Engineering and Computer Science
on June 6, 2019, in partial fulfillment of the
requirements for the Degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Organisms of varying degrees of complexity coordinate their diverse behavioral outputs over time, yet the internal neural dynamics underlying such behavioral organization is not completely understood. Behavior coordination can be captured through the quantitative description and subsequent analysis of behavioral states. Here, we develop an analytical method for the characterization of behavioral states in *C. elegans*. We observe posture sequences in wild type *C. elegans* and utilize a hidden Markov model to detect the behavioral states giving rise to these posture sequences. We then demonstrate that this method is generalizable to different *C. elegans* strains by applying this posture-based Markov analysis to *C. elegans* mutants and survey how these mutants differentially exhibit key behaviors within these behavioral states. This methodology provides a framework by which behavioral states can be quantified for further study of the neural dynamics underlying behavior coordination.

Thesis Supervisor: Steven W. Flavell, Ph.D.
Title: Assistant Professor

Acknowledgments

We thank H.R. Horvitz and the Caenorhabditis Genetics Center (CGC) for strains used in this work. The author would like to acknowledge Dr. Nathan Cermak for the design and implementation of the microscope used for data acquisition, Stephanie Yu for her contributions in attaining the data referenced in this thesis, the entire Flavell lab for helpful comments and conversations, and Professor Steven Flavell for his invaluable mentorship and making the work in this thesis possible. Lastly, the author wishes to thank Daniel Clark for his continuous support, productive discussions, and providing copious amounts of coffee.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation for the study of behavioral states

A behavior is any action taken by an organism. A behavioral state can be defined as a set of coordinated behaviors that tend to co-occur over an extended period of time. The behaviors that result from a given stimulus depend on the behavioral state as well as the stimulus. That is, a behavioral output is a function of both behavior state and exogenous input. Such behavioral systems thus allow for the integration of inputs and subsequent output to depend on the history of inputs. Additionally, behavioral states are the result of both exogenous and endogenous triggers. These aspects of behavioral states result in sophisticated behavioral responses to the environment.

Behavioral states tend to last for prolonged periods of time. However, neural signaling occurs on the order of milliseconds. This dichotomy in timescales between behavioral states and neural signaling suggests that there are underlying neural circuit dynamics that control how behaviors are coordinated and how information is integrated into the system. To analyze the neural circuit dynamics responsible for such coordination effectively, discerning approaches are required.

Currently, the neural circuit dynamics, and even the neural circuits themselves, underlying such behavior coordination are not completely understood. This may be due in part to the difficulty of defining and characterizing behavioral states in a quantitative and generalizable manner. This difficulty partially stems from the sheer complexity in the coordination of individual behaviors, which are themselves already complex, for higher-order organisms. This is especially true in natural, non-controlled settings. Furthermore, it is difficult to define a behavioral state in a rigorous and consistent manner.

1.2 *C. elegans* as a model organism

Caenorhabditis elegans (*C. elegans*) is a microscopic, transparent nematode that consumes bacteria in decaying organic matter. *C. elegans* has two sexes, hermaphrodite and male, where a hermaphrodite is able to self-fertilize but cannot fertilize a different hermaphrodite.[1] The nematode becomes a young adult approximately 72 hours after fertilization and has a typical lifespan of two to three weeks at 20°C [1]. In this

[1] In contrast, a male can fertilize any hermaphrodite, but cannot fertilize another male.

thesis, we will only consider young adult, hermaphrodite *C. elegans* that consume the bacteria strain *Escherichia coli* (*E. coli*) OP50.

C. elegans has been utilized as a model organism for research in cellular biology, developmental biology, and neurobiology. *C. elegans* was the first multicellular organism to have its entire genome sequenced. Additionally, the simplicity of the organism has permitted the study of various mutations in *C. elegans* [2, 3, 4]. One notable biological discovery that utilized *C. elegans* as a model organism is the identification of genes underlying apoptosis. Furthermore, the use of *C. elegans* enabled the development of green fluorescent protein (GFP) as a tool in biological research [5, 6, 7, 8]. Owing to these and many other applications, *C. elegans* are widely considered a useful model organism in biological research.

C. elegans are a useful model organism in systems neuroscience research in particular. One reason for this is the relatively small number of neurons in these nematodes.[2] *C. elegans*, consequently, is the only species whose connectome has been mapped in its entirety [10]. These aspects of the *C. elegans* nervous system allow for the study of neural circuits in a manner not currently possible in humans.

The highly stereotyped behavior of *C. elegans* facilitates the study of behavior. For instance, the dwelling state and the roaming state are commonly observed behavioral states used to describe locomotion patterns. The dwelling state is characterized by

[2] The nervous system of a hermaphrodite *C. elegans* is comprised of 302 neurons. The human brain, in contrast, contains approximately 100 billion neurons [1, 9].

low forward velocity and high angular speed, while the roaming state is characterized by high forward velocity and low angular speed [11]. One example of locomotion dependence on environmental cues is the effect of food on locomotion. Specifically, the animal tends to dwell when in the presence of an increased concentration or quality of food. Alternatively, the animal tends to roam in the presence of a reduced concentration or quality of food [12, 13].

Another behavior of interest in *C. elegans* is egg-laying. Hermaphrodites can lay a total of up to approximately 300 eggs and tend to do so at an average rate of 4 to 10 eggs per hour [1]. Specific locomotion patterns and egg-laying behavior have been previously shown to be correlated. Specifically, there is an increase in forward velocity and a ceasing of reversal events shortly before an egg-laying event [14]. As such, there is evident coordination between locomotion and egg-laying behaviors. While the exact neural mechanisms underlying this coordination are not yet fully understood, the relative simplicity of *C. elegans* makes this nematode a useful model organism in the study of this and other aspects of behavior coordination.

Despite the simplicity of *C. elegans*, they share many key biological properties with humans. In particular, various genes in the *C. elegans* genome have functional counterparts in humans. Additionally, previous research concerning associative learning, habituation, and sensitization in *C. elegans* have demonstrated that the organism is a useful model in the study of learning [15, 16]. Furthermore, humans and *C. elegans* have several common nervous system features, which range from the general structure of neurons to the organization of neuron subtypes. The relevance of understanding the simpler *C. elegans* nervous system for understanding the more complicated human system justifies its use as a model organism in systems neuroscience research.

1.3 Posture-based behavioral state modeling

While the use of *C. elegans* as a model organism in studying behavioral states has the benefit of characterizing simple exhibited behaviors, the issue of creating a generalizable method for defining behavioral states remains. Indeed, behaviors can be variable across mutants, so a behavioral state analysis based only on a set of variable behaviors may be problematic. However, with a few notable exceptions, the types of body postures expressed across mutants are fairly uniform. Furthermore, these postures contain information about other behaviors. For instance, when an animal is in motion, the postures the animal assumes must be capable of carrying out that motion. Additionally, in general, there is an increase in the rate of posture change when the animal is in motion than when the animal is resting. By leveraging these aspects of body posture, it is possible to create a quantitative method of defining behavioral states.

Previous works have utilized postures to quantitatively extract behavioral information. An example of this includes using posture structure and dynamics to characterize locomotion [17, 18]. Additionally, analyzing posture sequences allows for the data-driven study of behaviors across mutants [19]. Furthermore, previous research has explored how posture sequences vary on the basis of environment and population [20]. These studies motivate the use of a posture-based framework to study behavior coordination.

We now describe a posture-based analytical framework to study behavioral states in *C. elegans*. Specifically, our method heavily leverages the fact that postures are informative about other behaviors to extract behavioral states in a data-driven manner. In Chapter 2, we discuss data acquisition, quantification, and the processing of posture data from freely moving *C. elegans*. In Chapter 3, we use a hidden Markov model on this posture data to find different behavior states. We then apply this model to different *C. elegans* mutants with known behavioral phenotypes to demonstrate that our method results in the same conclusions as those in the literature. Using this analysis, we conclude by demonstrating that this framework allows for the detection of behavior coordination in a manner that can direct and facilitate future research on the biological mechanisms underlying behavior coordination.

Chapter 2

Data Acquisition and Processing

In this chapter, we present the procedure for extracting posture data from raw images. We first describe the handling of *C. elegans* and how recordings were performed. We then discuss how posture information was attained from these recordings and processed to facilitate the analytical quantification of the behavioral states described in Chapter 3.

2.1 Methods overview

Wild type *C. elegans* were recorded using an open-source automated tracking microscope. The posture of the animals at every frame of the recording was then quantified using fourteen reference points along the length of the animal. Next, every frame's recorded posture was matched to a compendium of general posture categories. Then, the transition dynamics between these general posture categories was established. These transition dynamics were subsequently used to generate transition groups, and every frame was assigned a transition group. Lastly, this sequence of transition groups was binned to coarsen the timescale of the sequence. Bins were categorized into clusters that represented the average transition group composition of a given time bin. Every bin was then assigned a cluster number according to the cluster classification of the bin. This sequence of binned clusters was then utilized as the input for the hidden Markov model described in Chapter 3.[1]

2.2 *C. elegans* growth and handling conditions

Nematode maintenance was conducted through commonly practiced methods [2]. Populations were maintained on NGM agar plates supplemented with *E. coli* OP50 bacteria at approximately 20°C.

Approximately 72 hours after fertilization, the animals were randomly selected for recording and transferred from their original growth plate to a low-peptone recording NGM agar plate. The animals were then recorded after a brief habituation period. Each recording consisted of only a single animal on the recording plate.

[1] All computational analyses in Sections 2.3 - 2.4 are performed using R [21], and all subsequent analyses are performed using MATLAB [22].

2.3 Data acquisition

Recordings were conducted on freely moving hermaphrodite young adult *C. elegans*. These recordings were performed using the OpenAutoScope, an open-source, low-magnification bright-field microscope with an automated tracking stage developed by Dr. Nathan Cermak. This microscope automates the tracking of the animal's movement along the recording plate, which enables recording of the animal without active supervision. The recordings were performed for approximately six continuous hours per animal at a rate of 20 frames per second (fps). For every frame of the recording, the posture of the animal, egg-laying, and a variety of other behaviors were recorded. An example raw image of a freely moving *C. elegans* as recorded by the tracking microscope is given in Figure 2-1, and a sample raw image of an egg-laying event is given in Figure 2-2 where the egg can be found at the midbody.

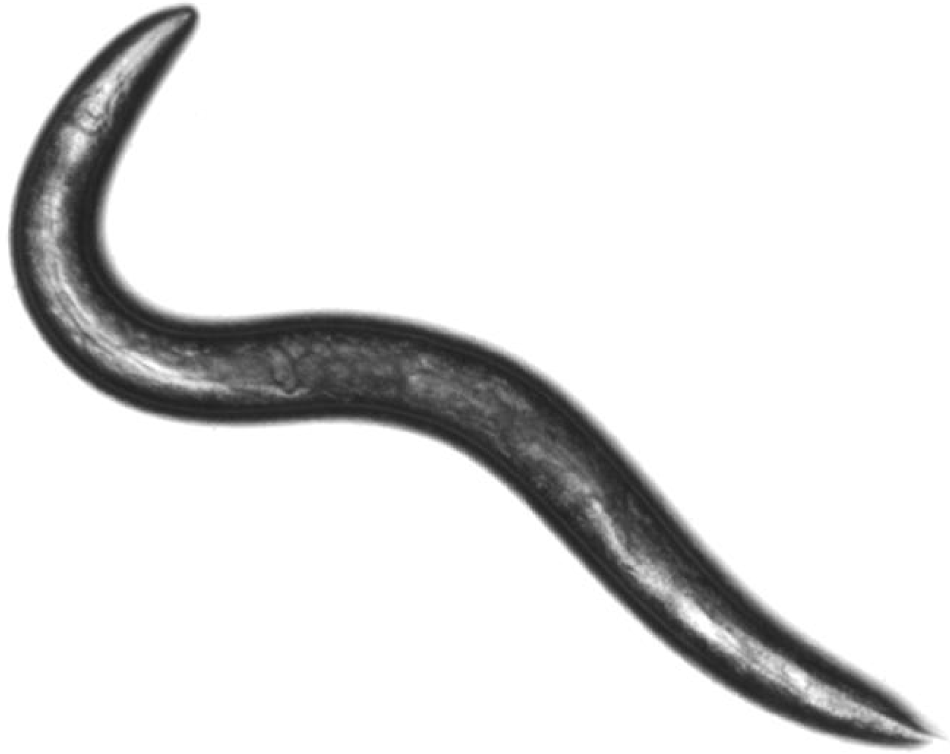

Figure 2-1. Bright-field image of a freely moving *C. elegans*: A sample raw image using the OpenAutoScope of freely moving *C. elegans*

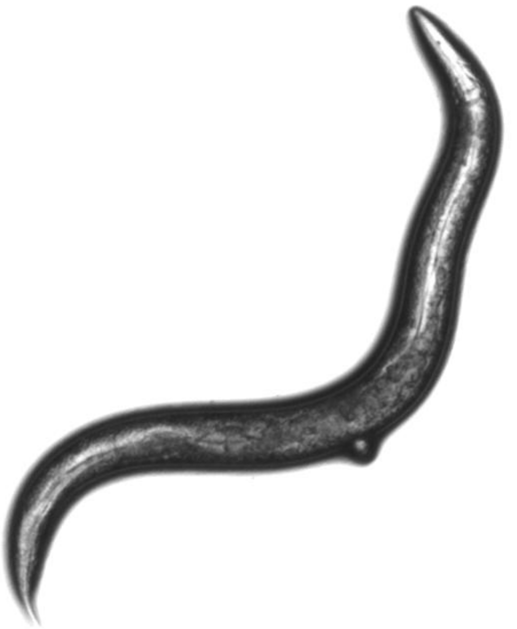

Figure 2-2. ***C.elegans* during an egg-laying event:** A sample raw image of freely moving *C. elegans* during an egg-laying event

2.4 Posture quantification

The posture was quantified in a manner consistent with existing literature [18]. For completeness, this quantification procedure is described below.

For every frame, spline interpolation was used to approximate the raw posture of the worm as a curve through the center of the animal's body. This curve was encoded with fourteen uniformly distributed segments. Each of the fourteen segments was encoded as the angle between the tangent vector of the worm's posture curve at that segment and an axis defined relative to the worm's orientation.

Formally, a posture matrix $\mathbf{P}$ is an $F \times S$ matrix, where F is the number of frames in the recording and $S = 14$ is the total number of body segments. The angle corresponding to body segment s in frame t, denoted by θ_t^s, is the row t, column s element of $\mathbf{P}$. The posture of the worm at frame t, ρ_t, is simply the t^{th} row of $\mathbf{P}$. That is,

$$\rho_t = [\theta_t^1, \theta_t^2, ..., \theta_t^s] \text{ for } t \in [1, 2, ..., F],\ s \in [1, 2, ..., S]. \tag{2.1}$$

Thus, a posture in a given frame is represented by a fourteen dimensional vector and can be visualized as 14 connected line segments, as in Figure 2-3, which is the output of the quantification procedure when the raw input is Figure 2-1.

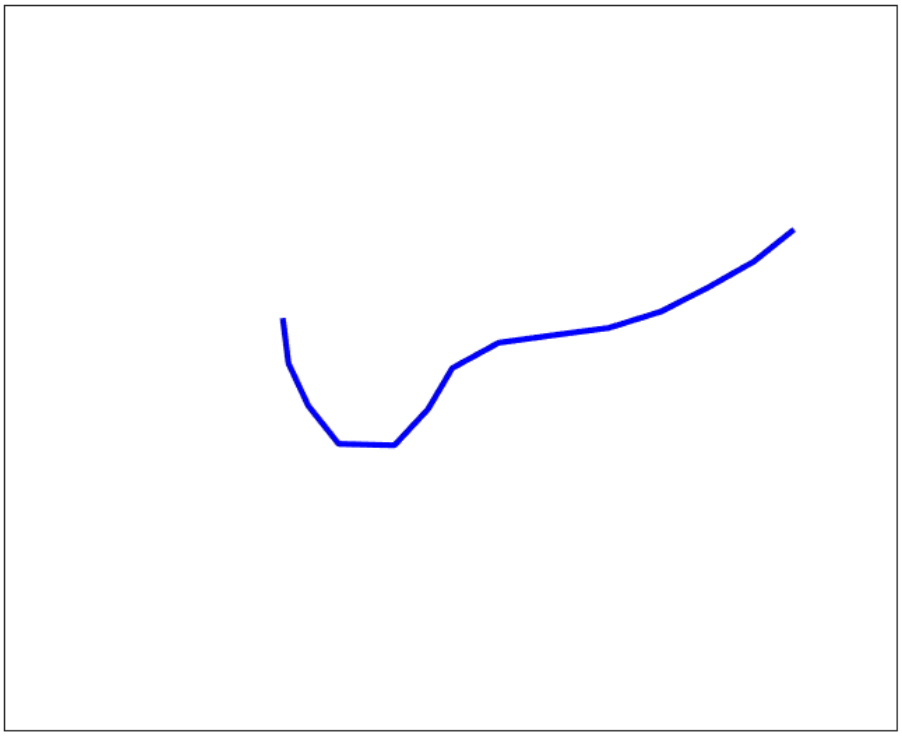

Figure 2-3. Posture quantification: A visual representation of the output of the quantification procedure when the raw input is Figure 2-1. The recorded posture has been reoriented for consistency across different frames.

2.5 Posture categorization

Following posture quantification, the posture of each frame was classified according to a compendium of posture categories [18]. This compendium was generated by performing hierarchical clustering analysis [23], where each posture category in the compendium corresponded to a cluster center. The size of the compendium was chosen to be the smallest compendium such that the explained variance percentage between the compendium and observed postures was greater than 75% (Figure 2-4) [18]. Using this criterion, a compendium of 100 postures was chosen (Figure 2-5).[2] This compendium can be described as a set of 100 elements, $\hat{\mathbf{P}}$ where each element is an S-dimensional vector. An element of $\hat{\mathbf{P}}$, thus, takes the form,

$$[\hat{\theta}^1, \hat{\theta}^2, ..., \hat{\theta}^s] \text{ for } s \in [1, 2, ...S]. \tag{2.2}$$

Every observed posture is then assigned a label $\hat{\rho} \in [1, ..., 100]$ that corresponds to the compendium posture that most resembles the observed posture. In this way, every frame whose posture was previously described as a vector ρ_t in Section 2.4 can now be described as $\hat{\rho}_t \in [1, ..., 100]$ where t represents the time of the posture observation. An example of this is given in Figure 2-6, which depicts the compendium posture that is assigned to the observed posture in both Figure 2-1 and Figure 2-3.

[2] This result was found using approximately 20% of the observed postures and verified by randomly sampling different subsets of observations.

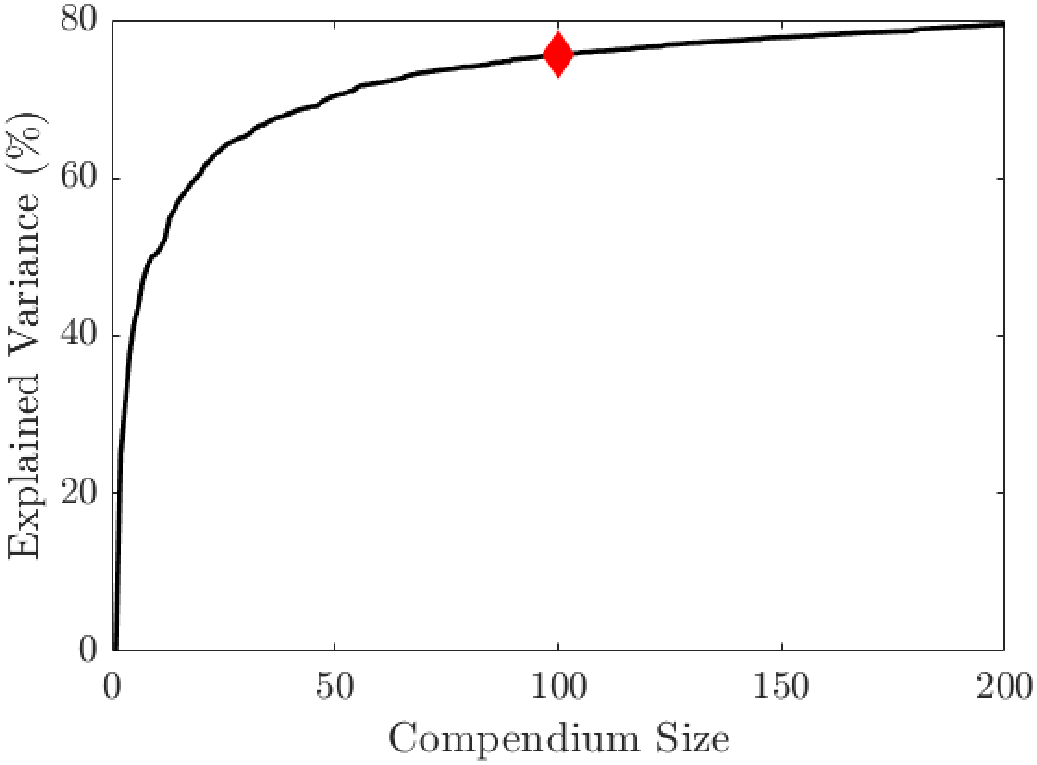

Figure 2-4. Posture compendium size selection: The percentage of explained variance between the compendium and observed postures for each compendium size is given by the black curve. The chosen size of the compendium, 100, is marked by the red diamond.

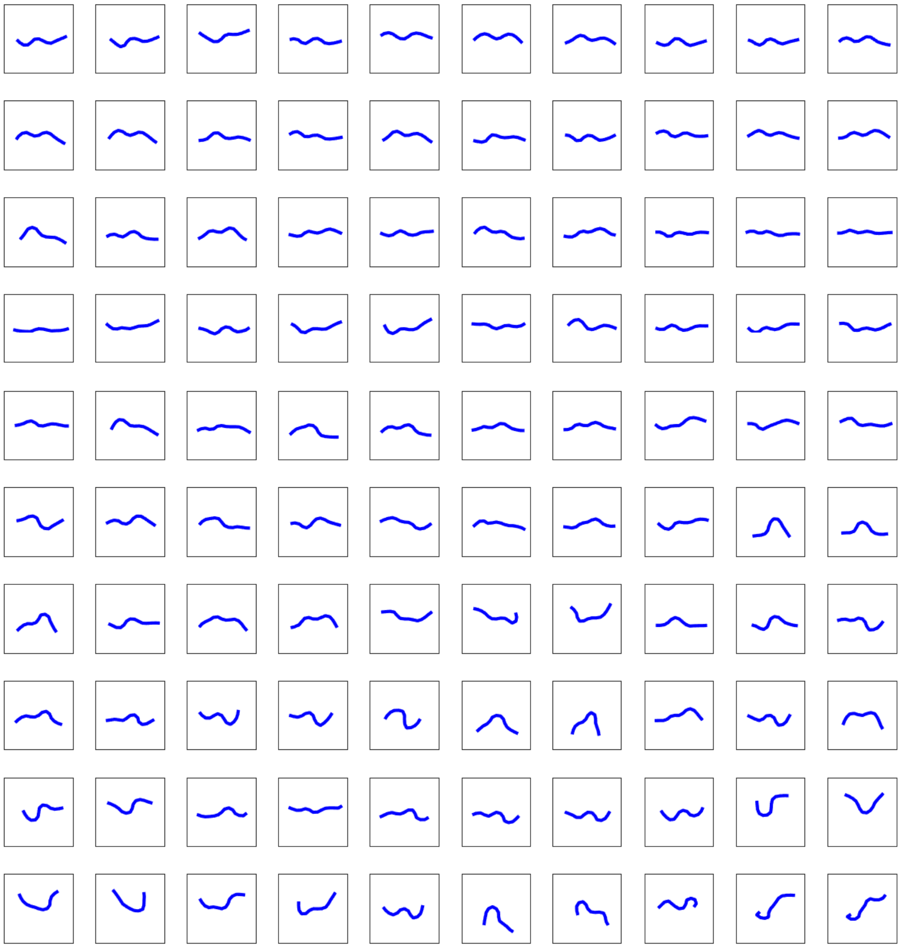

Figure 2-5. Posture compendium: Each of the subfigures depicts one of the 100 posture categories in the posture compendium. The posture in each frame is assigned the compendium posture to which it is most similar in shape.

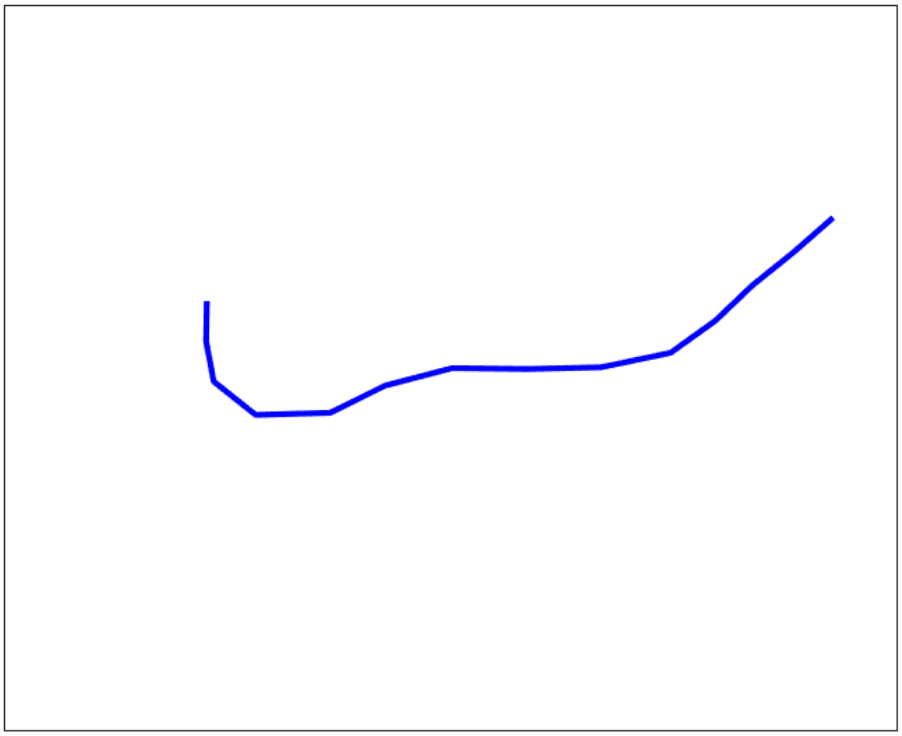

Figure 2-6. Compendium posture assignment: A visual representation of the compendium posture corresponding to both Figure 2-1 and Figure 2-3.

Using these posture classifications, we see that, on average, almost all of the postures are displayed for a similar amount of time (Figure 2-7). This suggests that no single posture or group of postures dominates the worm's behavior.

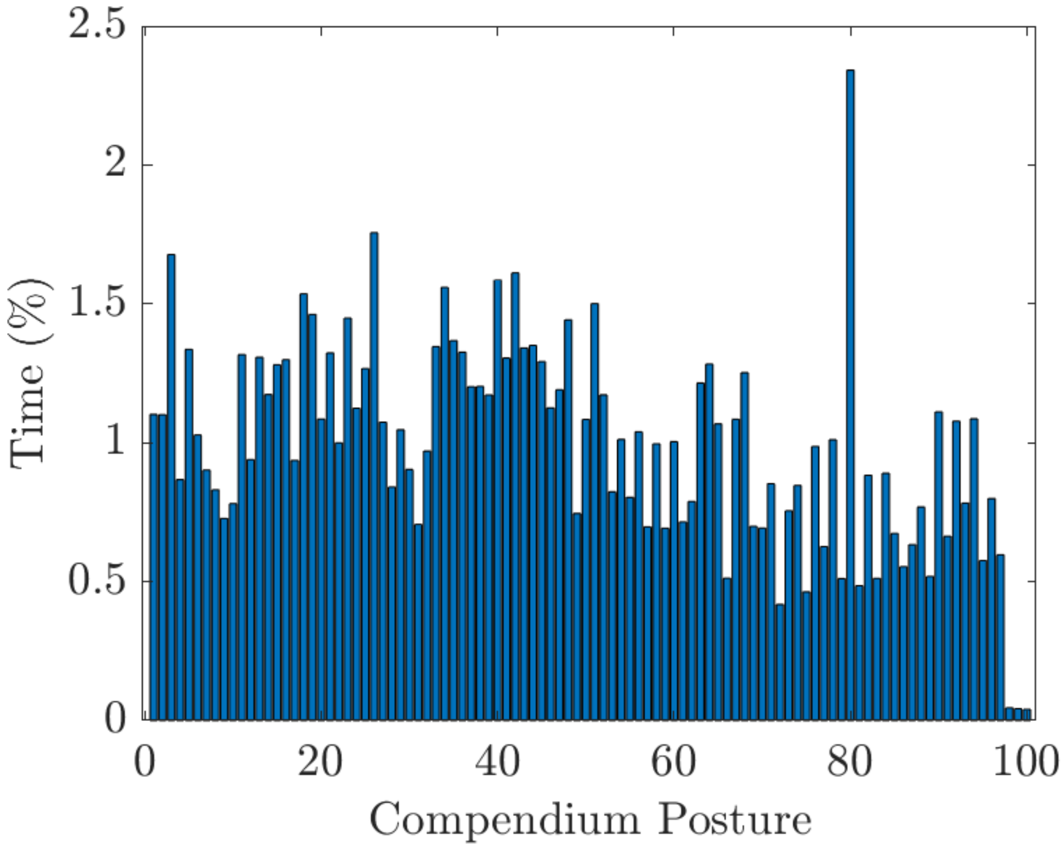

Figure 2-7. Percentage of time in posture categories: The average amount of time animals spent in each posture category (compendium posture). Note that category postures are exhibited approximately equally.

2.6 Posture transition dynamics

Once postures had been classified according to the posture compendium, we further classified the postures in each frame according to their transition group. We define a transition group as a subset of posture categories in which the postures in the group tend to transition frequently to the other postures within the group and infrequently to postures outside of the group.

This transition classification serves two main purposes. The first purpose is to correct for potential imperfections in the classification described in Section 2.5. Such an imperfection is demonstrated in the following example. Consider an observed posture that closely resembles both Posture Category A and Posture Category B. The observed posture may be trivially more similar to Posture Category A, and hence will be categorized as A. However, as the nematode continues in this general posture shape, the classification could oscillate between Posture Categories A and B as slight postural changes shift the classification. Such postural classification changes do not reflect any biological phenomena of interest, but rather are due to imperfections in the discretization and classification of postures. This imperfection, which does in fact occur in practice, is mitigated by the classification of posture types into transition groups.

The second purpose of these transition groups is to reduce the dimensionality of the posture sequences while preserving posture transition information. In the following section, we describe the method used to obtain these transition groups.

2.6.1 Posture transition network

We define a posture transition matrix that captures single time step transitions in postures (Figure 2-8). For a sequence of postures, an observed posture at frame t can be represented by $\hat{\rho}_t$. The dynamics of the sequence can be described by the transition matrix T where the element in row i, column j represents the probability that posture i at frame t transitions to posture j at frame $t+1$. Formally,

$$T = \begin{bmatrix} \Pr(\hat{\rho}_{t+1} = 1 | \hat{\rho}_t = 1) & \dots & \Pr(\hat{\rho}_{t+1} = C | \hat{\rho}_t = 1) \\ \vdots & \ddots & \vdots \\ \Pr(\hat{\rho}_{t+1} = 1 | \hat{\rho} = C) & \dots & \Pr(\hat{\rho}_{t+1} = C | \hat{\rho}_t = C) \end{bmatrix}. \tag{2.3}$$

Recall that $C = 100$.

We exclude self-transitions from consideration as self-transitions are significantly more probable than transitions to different postures, and we are interested in highlighting inter-posture dynamics. Therefore, we set every element in the diagonal of a modified transition matrix $\mathbb{T}$ to be zero. Every row of $\mathbb{T}$ is subsequently normalized to sum to 1, so the probability of a posture $\hat{\rho}_t = i$ transitioning to some posture $\hat{\rho}_{t+1} \neq i$ according to $\mathbb{T}$ is 1. More precisely, the element in row i, column j of $\mathbb{T}$ is,

$$\mathbb{T}(i,j) = \begin{cases} \frac{\Pr(\hat{\rho}_{t+1}=j|\hat{\rho}_t=i)}{\sum\limits_{j \neq i} \Pr(\hat{\rho}_{t+1}=j|\hat{\rho}_t=i)}, & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}. \tag{2.4}$$

2.6.2 Defining transition groups

Once the transition dynamics for the sequence of posture categories were found, these dynamics were captured by a directed graph, $\mathbb{G} = (V, E)$, where V is the set of nodes and E is the set of edges in the graph $\mathbb{G}$. Each element of V, v_i for $i \in [1, ..., 100]$, represents a single posture category. Each element of E, e_{ij}, has an associated weight

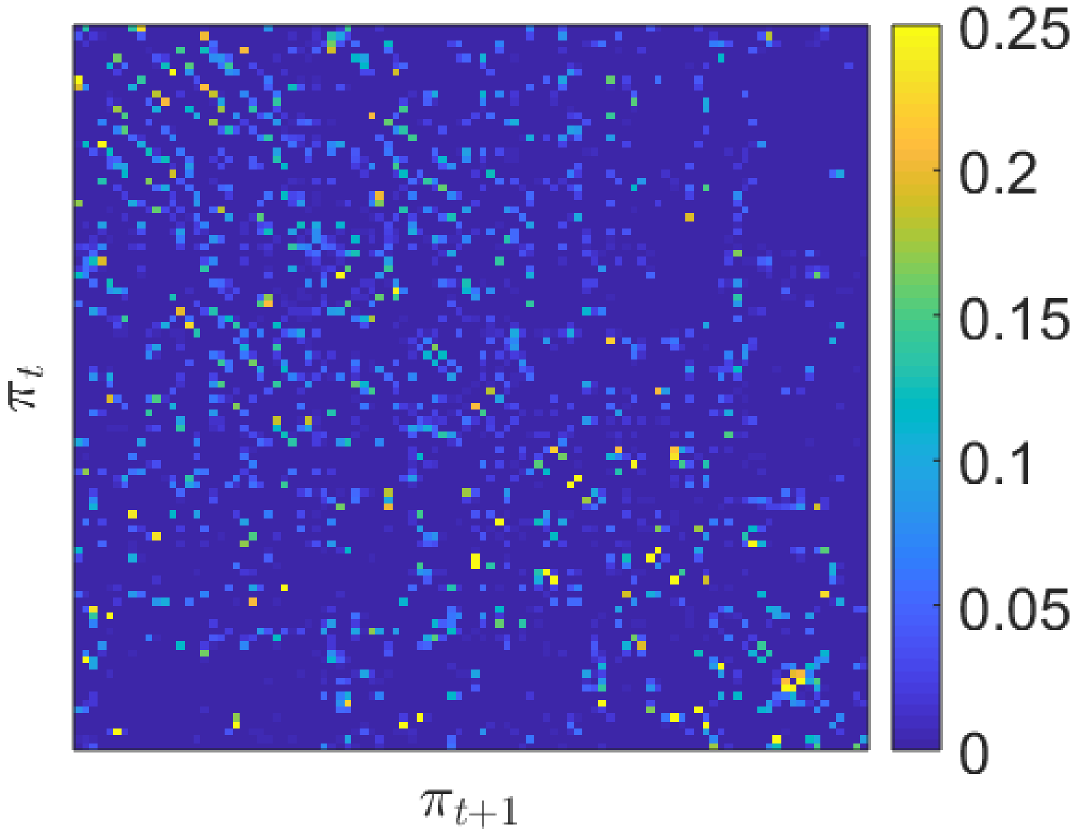

Figure 2-8. Posture transitions: A visual representation of the posture transition dynamics as a matrix where the rows represent the posture at frame t, the columns represent the posture at frame $t+1$, and the color at row i, column j represents the corresponding probability of transition, $\Pr(\hat{\rho}_{t+1} = j | \hat{\rho}_t = i)$. The color axis is rescaled such that all probabilities of 0.25 and greater are the same color.

$W(e_{ij}) = \mathbb{T}(i,j)$, that represents a transition probability from some v_i to some v_j. We can visualize the posture transition dynamics using G, where we limit the outdegree of each node to be three for ease of representation (Figure 2-9). On average, the outdegree and indegree of each node is 40. For each posture in the network, there are approximately 5 other postures to which it regularly transitions.

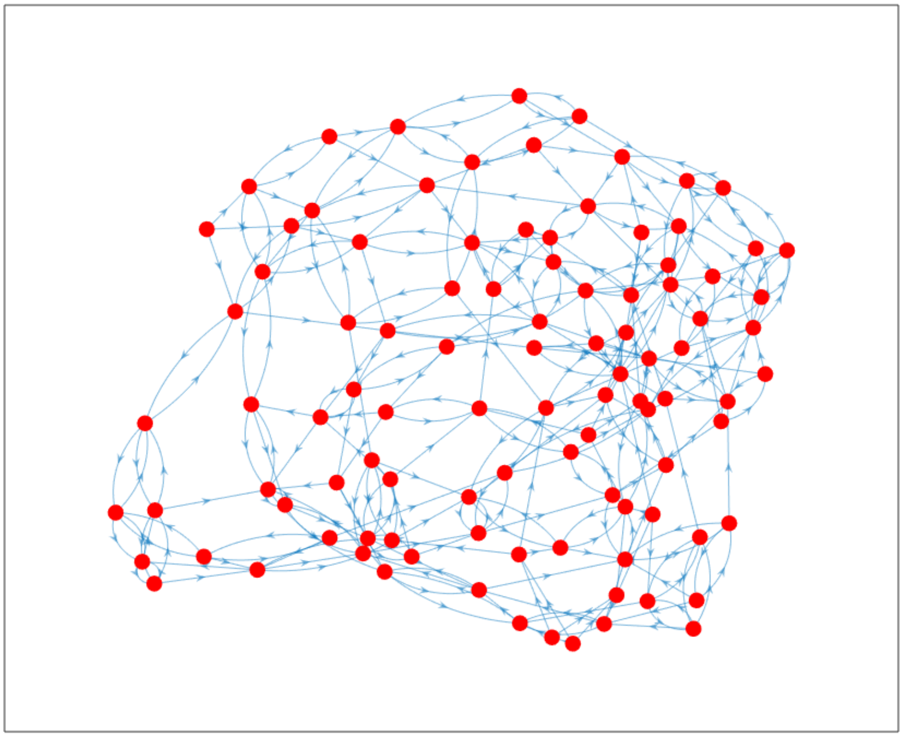

Figure 2-9. Posture transition network: A visual representation of the posture transition dynamics as a network. For ease of representation, each node has an outdegree of three, which correspond to its three most probable transitions. Additionally, the edge weights have been modified so that an edge representing a highly probable transition has a lower weight than an edge representing an unlikely transition. This was done so that nodes that are more likely to transition to one another are in closer proximity to one another in the network.

Next, we cluster the network via spectral clustering [24] to determine a set of transition groups, $\mathbf{H} = \{\eta^1, ...\eta^k\}$, where k is the number of transition groups and, for each $g \in [1, 2, ...k]$, η^g is a set of $\hat{\rho}_t$ that transition to one another with high likelihood. We select the number of transition groups by determining the clustering that maximizes the probability that a posture $\hat{\rho}_t$ would transition to a posture $\hat{\rho}_{t+1}$

within the same transition group (η^g), which equivalently minimizes the probability that a posture would transition to a posture outside of the group. However, this needs to be performed while controlling for cluster size. To illustrate the need to control for cluster size, consider a partition into two clusters: one large cluster that contains nearly all the postures and one small cluster. This partitioning, though not necessarily meaningful, would maximize the likelihood of intra-cluster posture transitions as most postures are in the same cluster.

We now describe the selection procedure in detail. For a posture $\hat{\rho}_t \in \eta^g$, we define the transition ratio, R_k, for a partitioning of transition groups. This ratio captures the quality of the clustering, where the quality is defined as the likelihood $\hat{\rho}_t$ will transition to any $\hat{\rho}_{t+1} \in \eta^g$ for each transition cluster. This ratio is defined as,

$$R_k = \frac{\frac{1}{k}\sum\limits_{g=1}^{k} \alpha_g}{\frac{1}{k}\sum\limits_{g=1}^{k} \bar{\alpha}_g}, \tag{2.5}$$

where α_g is the average probability of an intra-group transition and $\bar{\alpha}_g$ is the average probability of an inter-group transition. We define α_g as,

$$\alpha_g = \frac{1}{h}\sum_{\hat{\rho}_t=1}^{h} \Pr(\hat{\rho}_{t+1} \in \eta_g | \hat{\rho}_t \in \eta_g), \tag{2.6}$$

and we define $\bar{\alpha}_g$ as

$$\bar{\alpha}_g = \frac{1}{h}\sum_{\hat{\rho}_t=1}^{h} \Pr(\hat{\rho}_{t+1} \notin \eta_g | \hat{\rho}_t \in \eta_g), \tag{2.7}$$

where h is the size of the transition group. We then vary the number of clusters from 2 to 10 and repeat each clustering for multiple iterations. Using this criterion, we find the iteration of the clustering that maximizes this transition ratio. By comparing these transition ratios, we determine the optimal number of transition groups to be eight (Figure 2-10).[3]

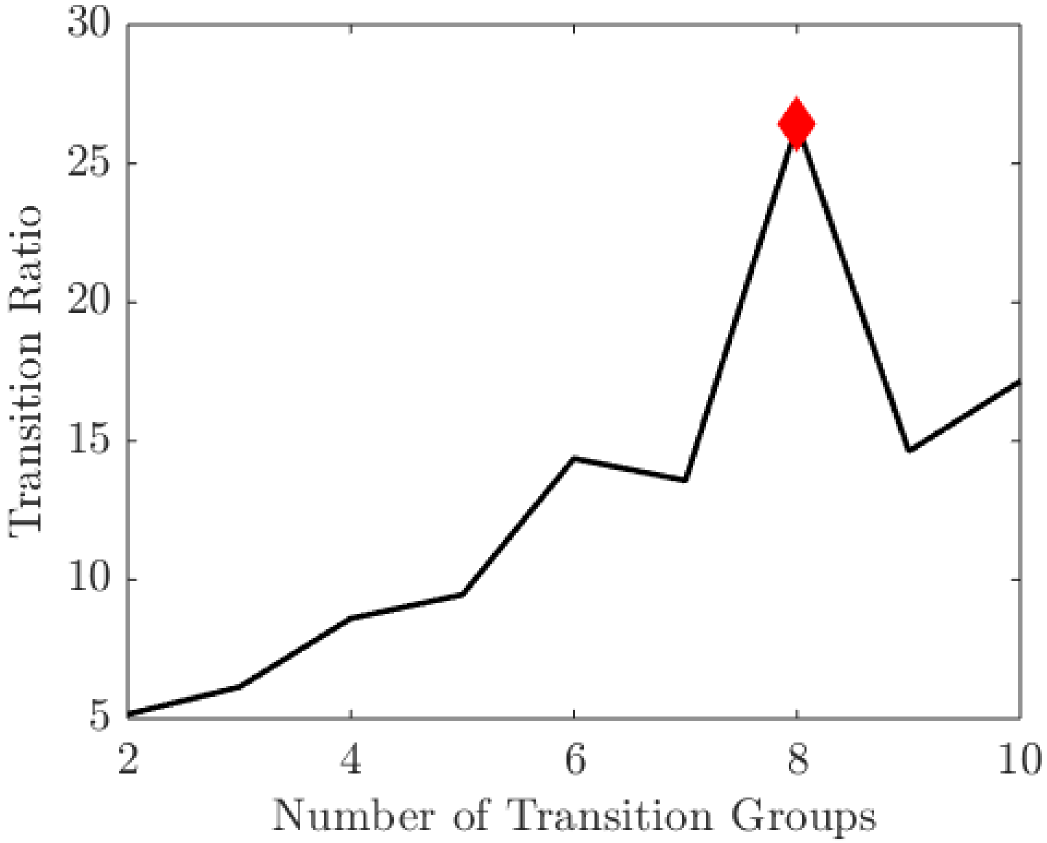

Figure 2-10. Transition group selection: The curve generated by the clustering selection criterion as described in Equation (2.5). As we maximize this value, we selected a clustering of 8 clusters, as denoted by the red diamond.

[3] It is worth noting that occasionally, the maximum transition ratio value occurred for a clustering of 7 or 9 clusters rather than 8, though 8 generally appeared to be the most likely clustering to maximize the selection criterion.

Using the optimal clustering[4] and iteration of that clustering, we can then visually represent the transition groups by reorganizing the transitions in Figure 2-8 so that the rows and columns are ordered by transition group, rather than by numerical label (Figure 2-11). The transition groups do not encode perfect transition clusters where all postures that are likely to transition to one another are classified in the same group. However, the clustering does capture the broad transition features, as can be seen by comparing Figure 2-8 and Figure 2-11. Furthermore, we expect that there are stereotyped inter-cluster transitions wherein a posture of a given transition group acts as a “gateway" posture into a different transition group. As such, the transition clustering creates an adequate transition grouping to extract posture transition dynamics, reduces dimensionality of the posture sequences, and minimizes the effect of classification noise. Using the transition groups, every frame's posture is assigned a transition group (arbitrarily labelled 1-8). That is to say, every posture previously described as $\hat{\rho}_t$ in Section 2.5 is now assigned a value $\eta_t \in [1, ..., 8]$ corresponding to the transition group, η^g, that contains the posture at frame t.

[4] This clustering was found by using 50% of the data (N=12) and verified by randomly sampling different subsets of observations.

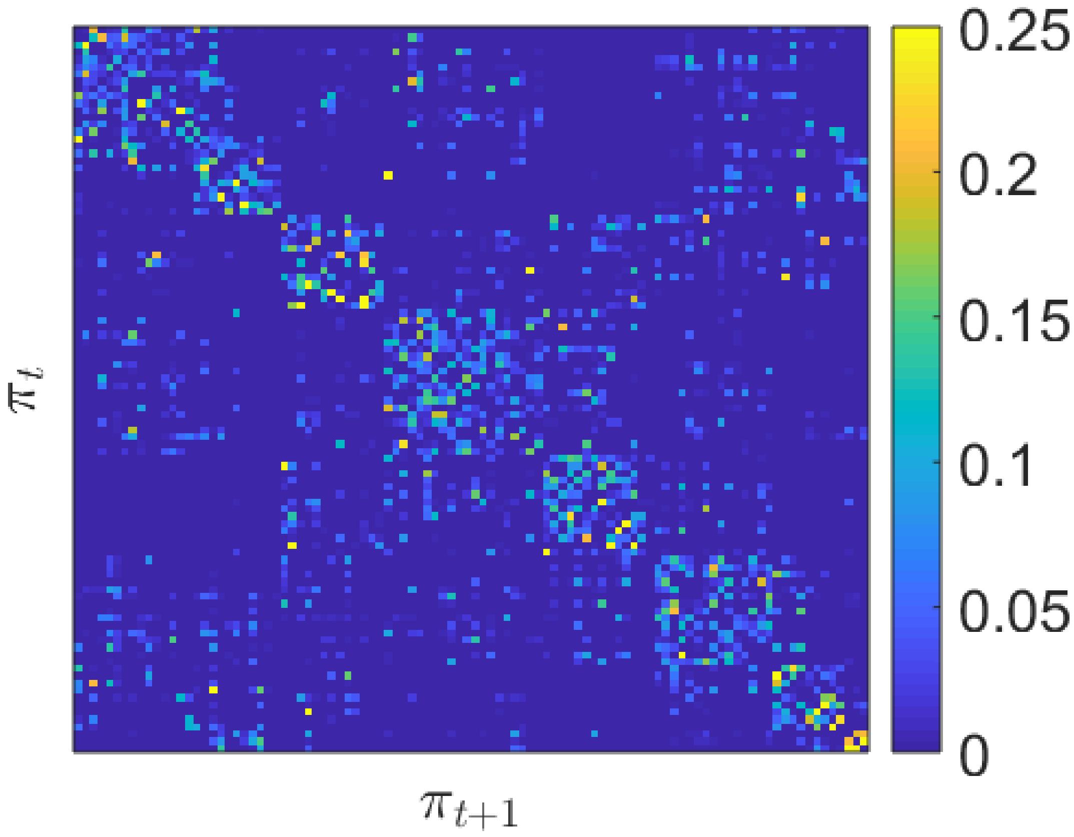

Figure 2-11. Clustered posture transitions: A visual representation of the posture transition dynamics as a matrix where the rows represent the posture at frame t, the columns represent the posture at frame $t + 1$, and the color at row i, column j represents the corresponding transition probability, $\Pr(\hat{\rho}_{t+1} = j|\hat{\rho}_t = i)$. The rows and columns are ordered by transition group rather than by numerical posture category. As such, the diagonal squares represent the transitions within a cluster. The color axis is rescaled such that all probabilities of 0.25 and greater are the same color.

2.7 Posture sequence time-scale expansion

Currently, every observation occurs at every frame (0.05 seconds) and our current data processing only analyzes sequences on these short timescales. However, we are interested in longer timescale phenomena. As such, we coarsen the timescale of observations so that a single observation becomes the binned data over a window of sixty frames (3 seconds). In this section, we describe the binning procedure.

For every sixty-frame window, we determine the percentage of time spent in each of the η_g transition groups during the window. This is done in segments rather than continuously. Consequently, a data recording of duration F would result in $F/60$ bins. This results in a set of $F/60$ vectors that represent the transition group compositions of each of these windows of time. This data is partitioned via k-means clustering method [25], and the optimal number of clusters is determined using silhouette analysis [26]. Figure 2-12 depicts the average silhouette value for each clustering, where the highest silhouette value is given by a clustering consisting of nine partitions. Once the optimal cluster number is found, the partitioning is chosen by which iteration of the algorithm minimized the sum of the distances from each point to their respective cluster centroid. Subsequently, the posture previously assigned the value η_t in Section 2.6.2 is now assigned a label $\beta_\tau \in [1, ..., o]$ where $\tau \in [1, 2, ..., \frac{F}{60}]$ and $o = 9$ is the number of bin clusters.[5]

[5] This result was found using 50% of the data (N=12) and verified by randomly sampling different subsets of the data.

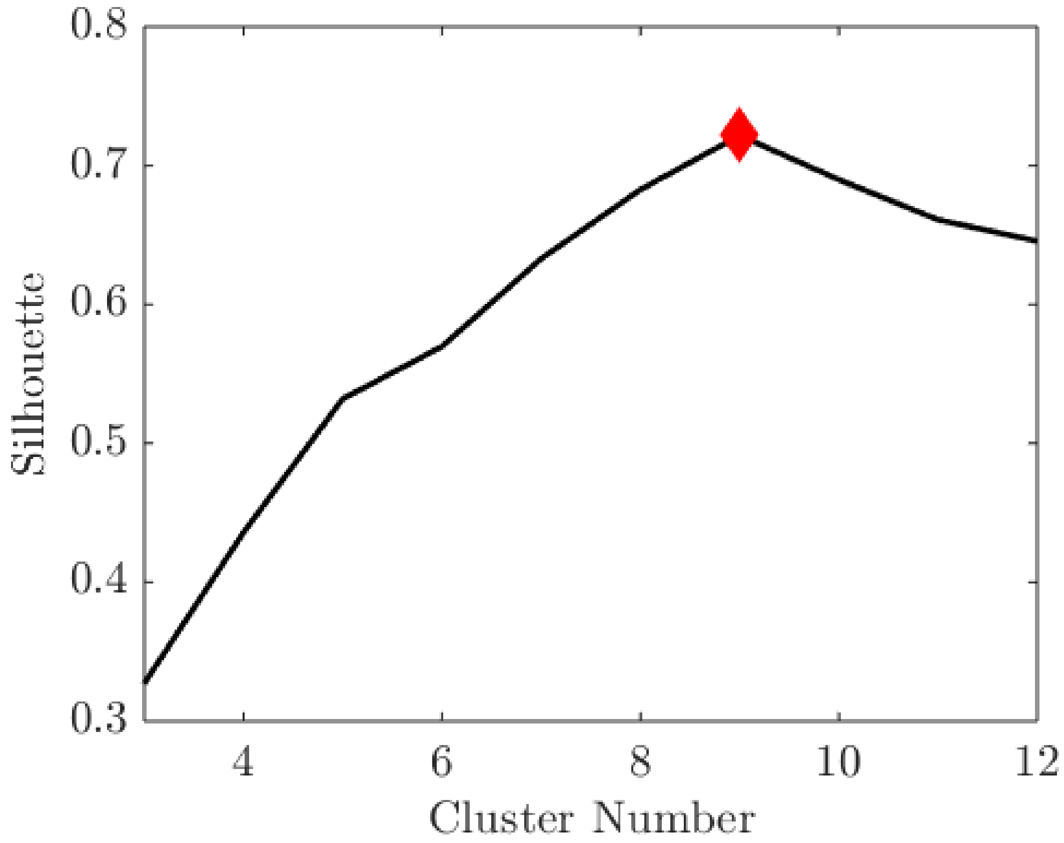

Figure 2-12. Silhouette curve for bin clustering: The average silhouette value for the clustering of data with the number of partitions ranging from 3 to 12; the optimal number of clusters, 9, is denoted by the red diamond.

In summary, Chapter 2 describes the quantitative processing of postures prior to the application of Markov analysis. For a given raw image of a freely moving N2 *C. elegans* at time t, the postures are quantified (ρ_t), categorized into a posture category ($\hat{\rho}_t$), assigned a transition group (η_t), and finally binned and assigned a bin cluster (β_τ). This sequence of bins is subsequently used as the input for the analytical method described in Chapter 3.

Chapter 3

Defining Behavioral States through Markov Analysis

In Chapter 2, we outlined the method for processing postural data. Here, we describe the method by which processed postural data is analyzed using a hidden Markov model to quantitatively estimate hidden behavioral states. We then characterize the states extracted by the model. We conclude by demonstrating the application of the posture classifications previously found to different mutant *C. elegans*, and discuss future research using this method.

3.1 Method overview

A hidden Markov model is a statistical model used to describe a system with a set of hidden states that probabilistically emit a sequence of observables in a state-dependent manner. Here, the sequence of emissions is B $= [\beta_1, ..., \beta_\tau, ..., \beta_{F^*}]$ where $F^* = \frac{F}{60}$, and the set of possible emissions is $\beta = \{1, ..., o\}$. The hidden states are given by $\mathbf{X} = \{\chi_1, ..., \chi_\sigma\}$ where σ is the number of hidden states in the system. In this analysis, we use a hidden Markov model to estimate the parameters of the system governing the emission of **B**. Specifically, we use the Baum-Welch algorithm to estimate the transition and emission probabilities of the hidden Markov model [27, 28, 29]. Once the parameters have been estimated, we use the Viterbi algorithm to find the most likely sequence of hidden states that generated a given observed sequence of emissions [30]. We then use the most likely state path as a proxy for the hidden behavioral states of the animal occurring along this sequence. Finally, we analyze these states in conjunction with other behaviors for both wild type *C. elegans* as well as mutants.

3.2 Model training

We trained the model using the Baum-Welch algorithm on the wild type dataset (N=12) while varying the number of states in a trained model.[1] Different models

[1] Separately training the model on the remaining data (N=12) confirmed the following results.

with the number of states ranging from 2 to 13 were each initialized with random emission and transition probabilities. Then, for each model size, we found the highest likelihood model. Using Bayesian information criterion (BIC) [31], we determined the optimal number of states to be 10 (Figure 3-1).

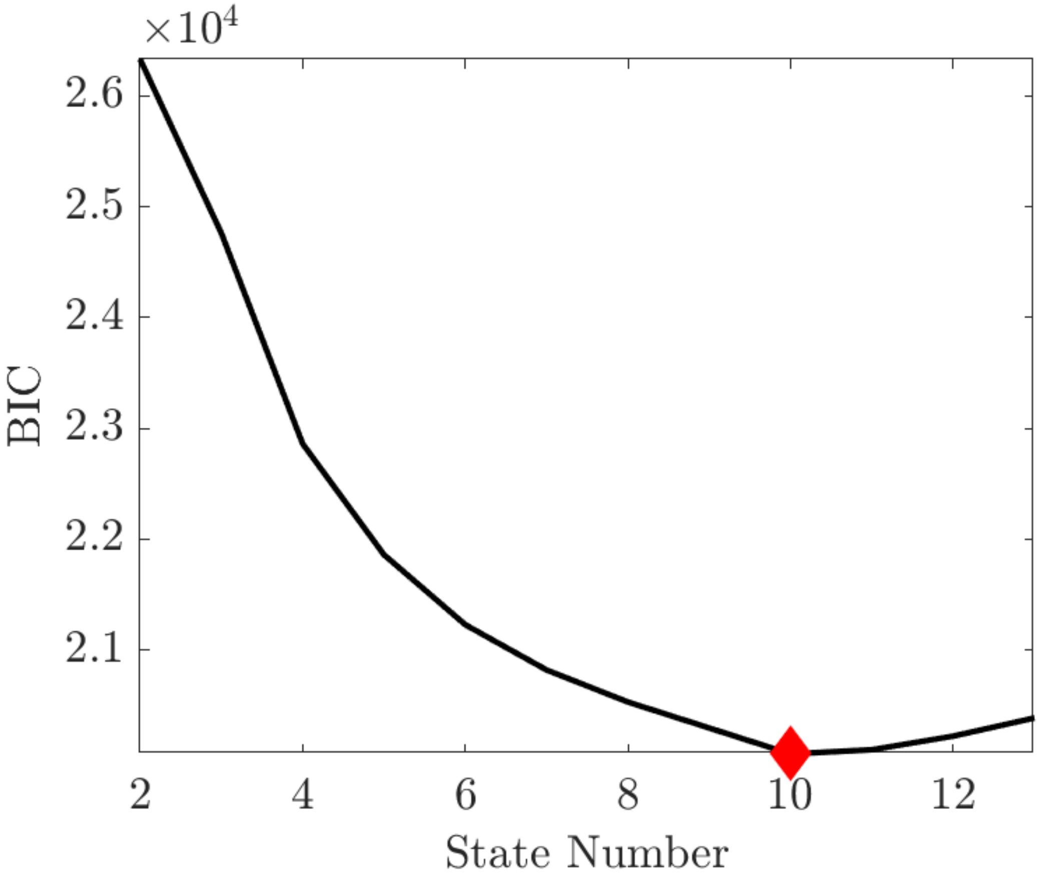

Figure 3-1. Hidden Markov model selection: Using the Bayesian information criterion (BIC), we found the optimal number of hidden states to be 10. The curve generated by finding the BIC value for each model is given in black, while the selected number of states is marked in red.

From the models trained with 10 hidden states, we selected the model with the highest likelihood. After doing so, we used the emission (Figure 3-2) and transition (Figure 3-3) probabilities of the model to estimate the hidden states of the observed posture sequences.

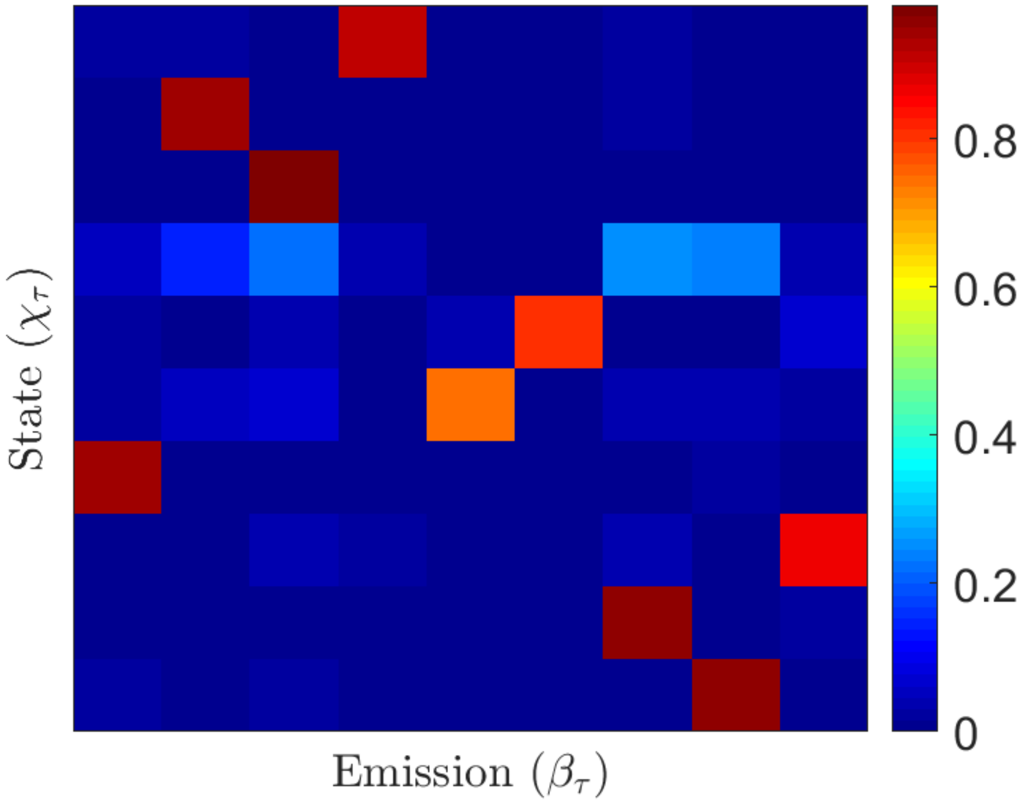

Figure 3-2. Hidden Markov model emission probabilities: The emission probabilities of the selected hidden Markov model are given in matrix form where the probability in row i, column j is the probability that when the system is in state i, observation j will be emitted. More formally, this is the probability $\Pr(\beta_\tau = j|\chi_\tau = i)$, and this value is illustrated by color.

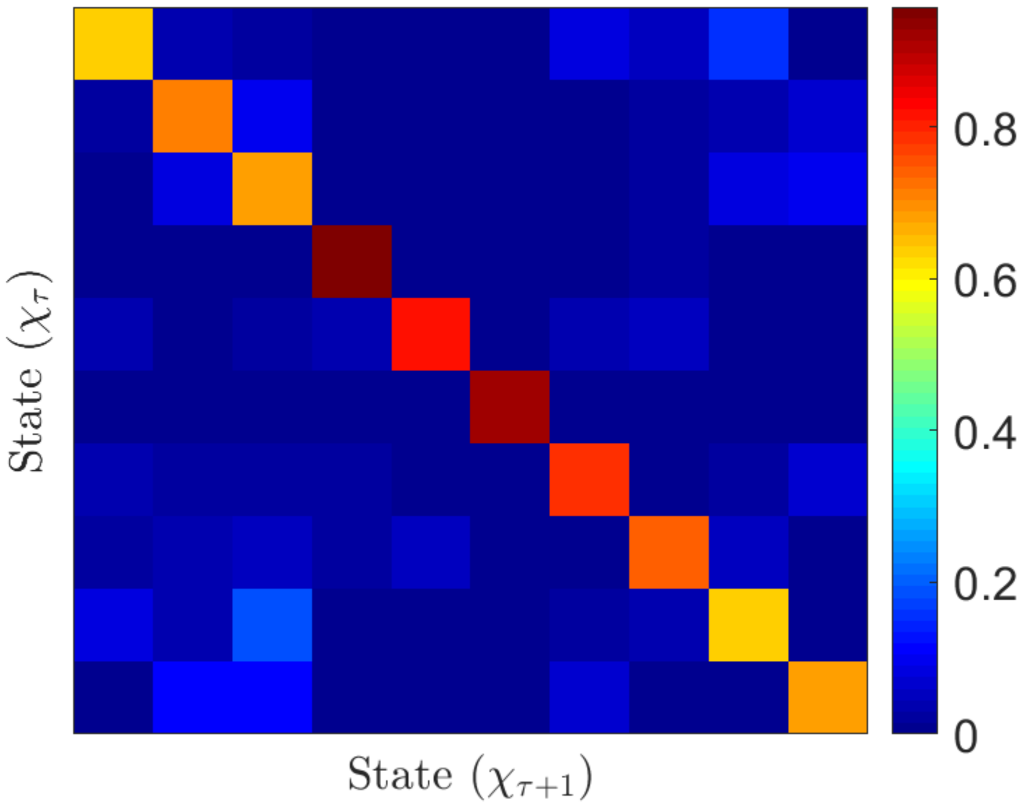

Figure 3-3. Hidden Markov model transition probabilities: The transition probabilities of the selected hidden Markov model are given in matrix form where the probability in row i, column j is the probability that when the system is in state i at time τ, the system will transition to state j at time $\tau + 1$. More formally, this is the probability $\Pr(\chi_{\tau+1} = j|\chi_\tau = i)$, and this value is illustrated by color.

3.3 Behavioral state characterization

3.3.1 State attributes

Using the estimated emission and transition probabilities of the system, we now classify each frame according to the worm's state at that time, where the states are arbitrarily labelled 1 to 10.

The probability of a worm remaining in a given state exponentially decays over time (Figure 3-4). This is consistent with the system adhering to the memoryless property of a stochastic process (Markov property). Furthermore, we can see that the average continuous durations of the states are not uniform across states, and some states have long durations (Figure 3-5). As such, there is heterogeneity between the states. Lastly, we see that the distribution of time in the states is, again, not homogenous (Figure 3-6). We will discuss the implications of this when we discuss individual behaviors linked to the various states.

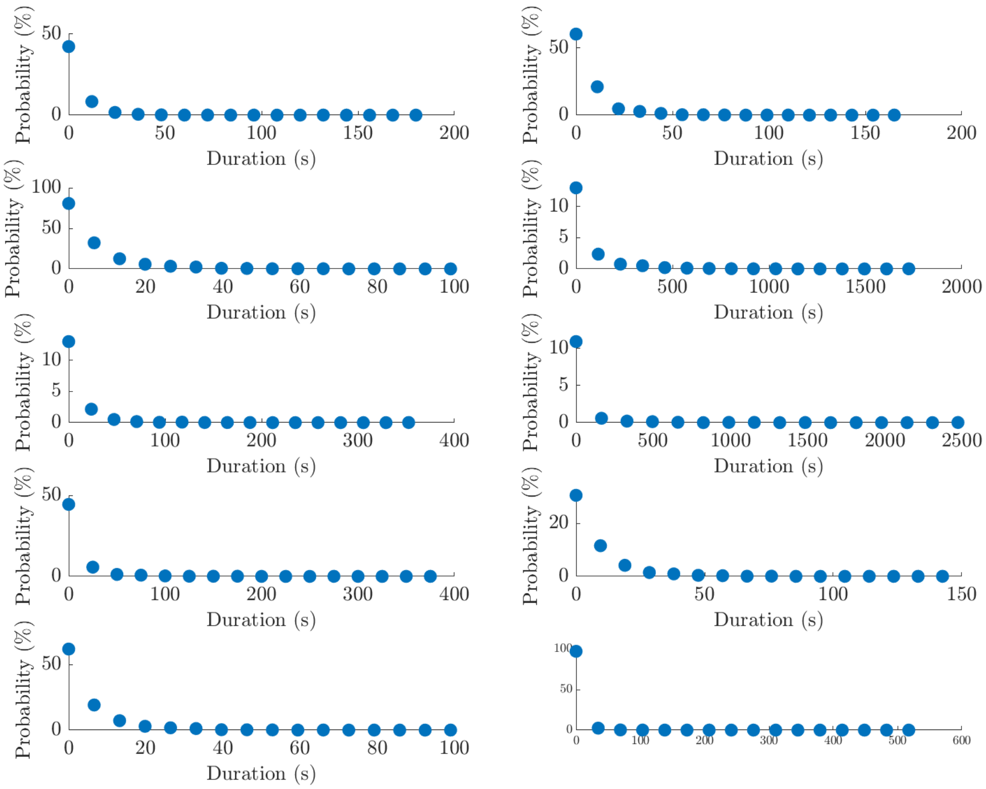

Figure 3-4. Probability of state duration: The probability of remaining in a state for a given duration of time; the probabilities were found by tracking the number of instances in which an animal remained in the state for the given duration.

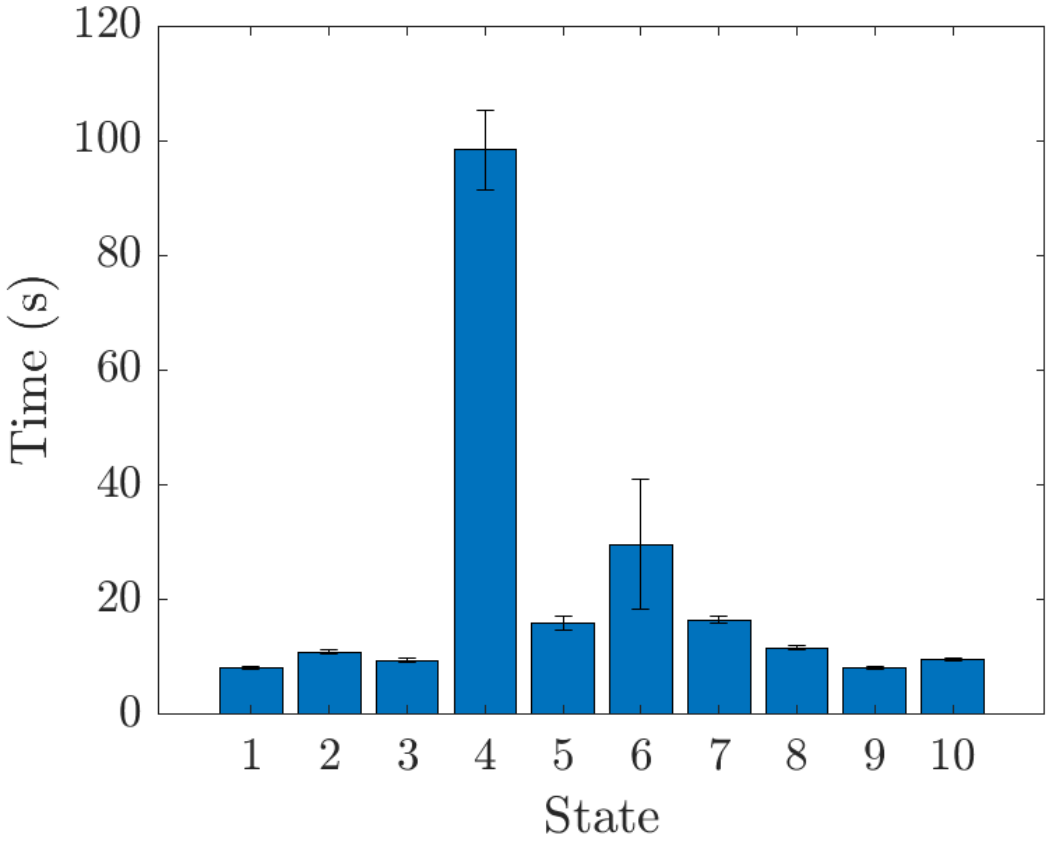

Figure 3-5. Continuous duration of states: The continuous duration of each state in seconds (mean ± SEM)

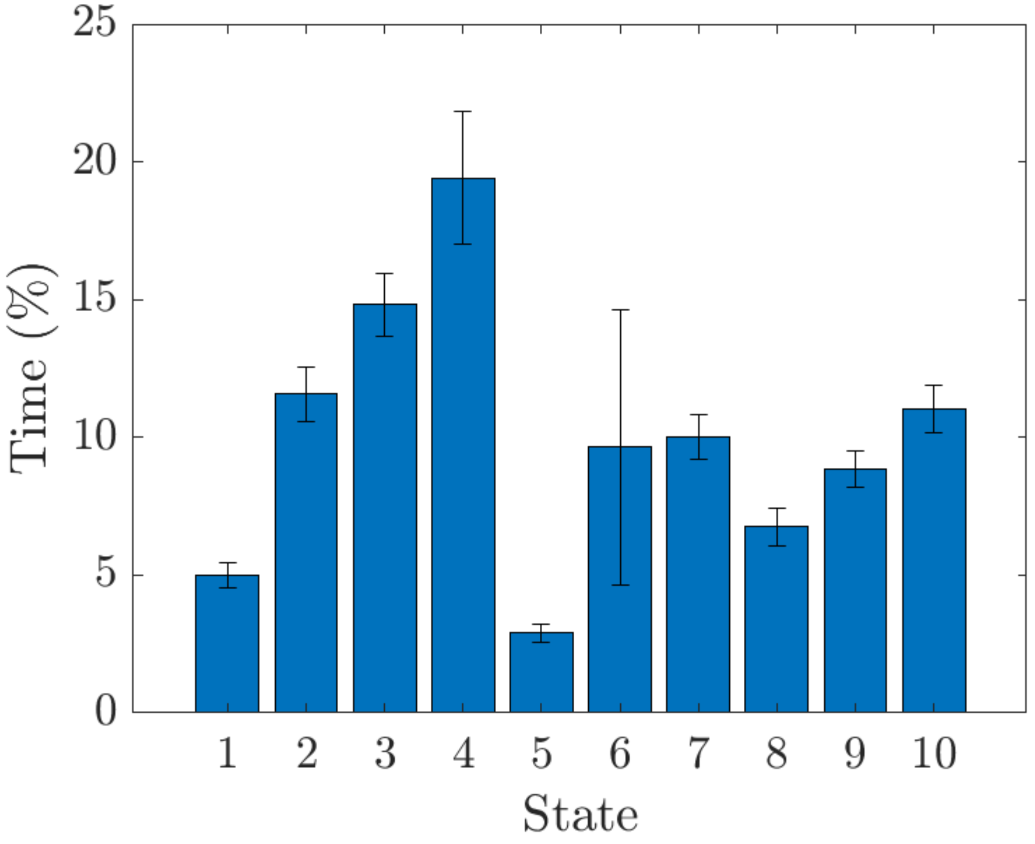

Figure 3-6. Percentage of time in states: The percentage of a recording the animals spent in each state (mean ± SEM)

3.3.2 Analysis of behaviors across states

Next, we analyze how these states correspond to different behaviors in wild type *C. elegans*. As described in Chapter 1, the roaming and dwelling state are the two most salient behavioral states in *C. elegans*. The dwelling state is characterized by reduced forward velocity and increased angular speed, while the roaming state is characterized by increased forward velocity and reduced angular speed. Therefore, for this model to be valid, it should extract roaming and dwelling states. If we consider the average velocity and angular speed of the worm in each state, we find that, indeed, the model extracts roaming and dwelling states. More specifically, we find that State 4 is an example of a roaming state as it is characterized by high forward velocity and reduced angular speed, while the other states are likely dwelling states (Figures 3-7, 3-8). Furthermore, *C. elegans* have been previously found to spend approximately 20% of their time in a roaming state [13]. As can be seen in Figure 3-6, the worms spend, on average, approximately 20% of their time in State 4.

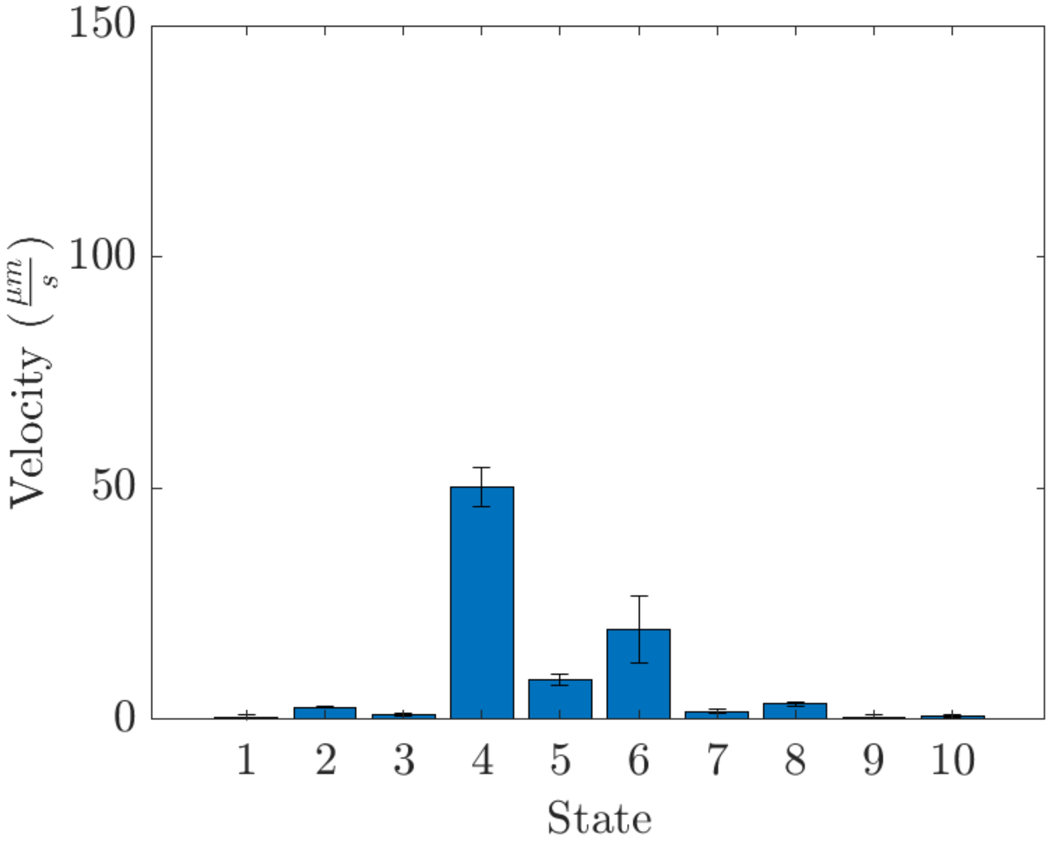

Figure 3-7. Velocity per state: The velocity animals exhibited per state, given in μm per second (mean ± SEM)

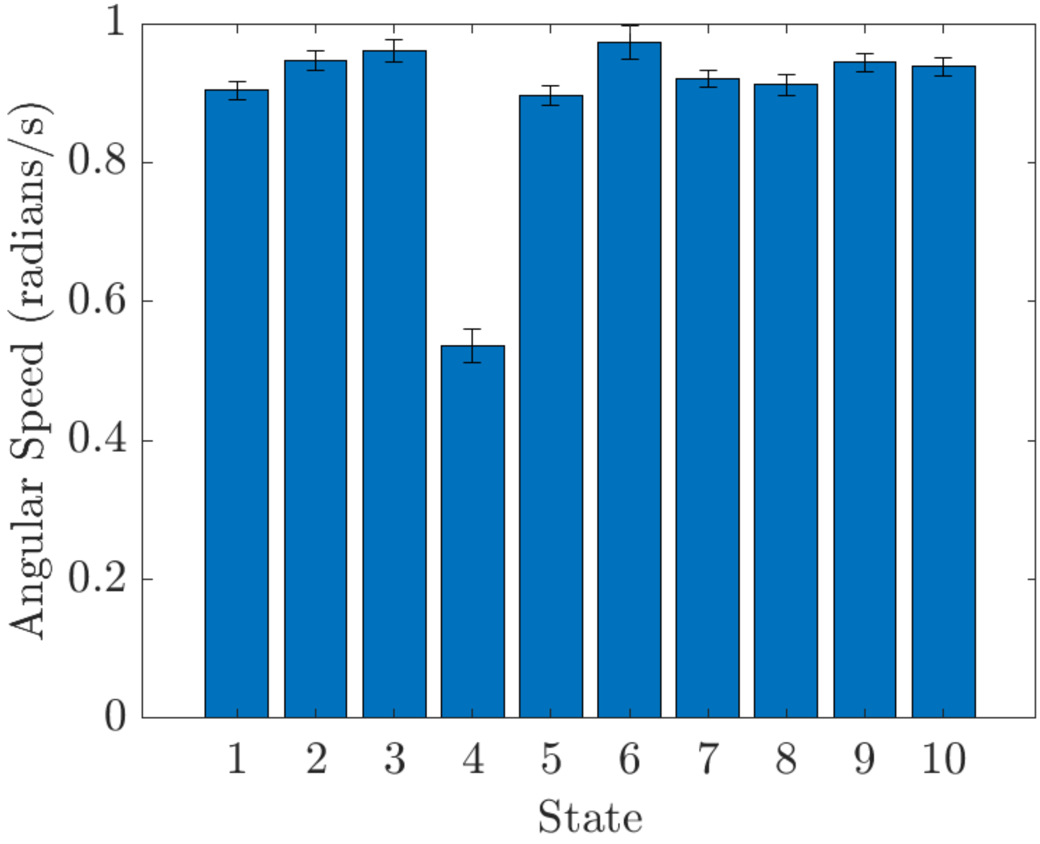

Figure 3-8. Angular speed per state: The angular speed animals exhibited in each state, given in radians per second (mean ± SEM)

Finally, we consider egg-laying behaviors in relation to these states. The egg-laying rate in each state is given in Figure 3-9. We see that egg-laying increases during the roaming state, which is consistent with previous research [14]. Furthermore, we see that the majority of eggs are laid in the roaming state (Figure 3-10).

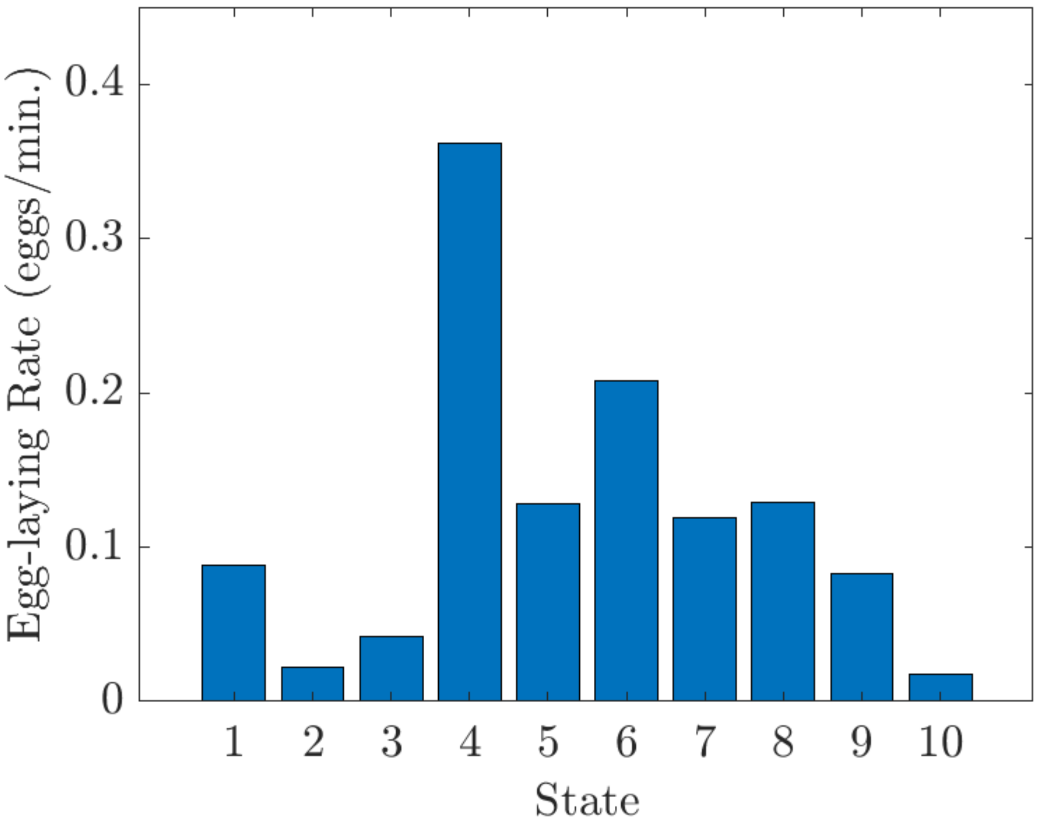

Figure 3-9. Egg-laying rate per state: The average egg-laying rate of animals in each state, given in eggs per minute

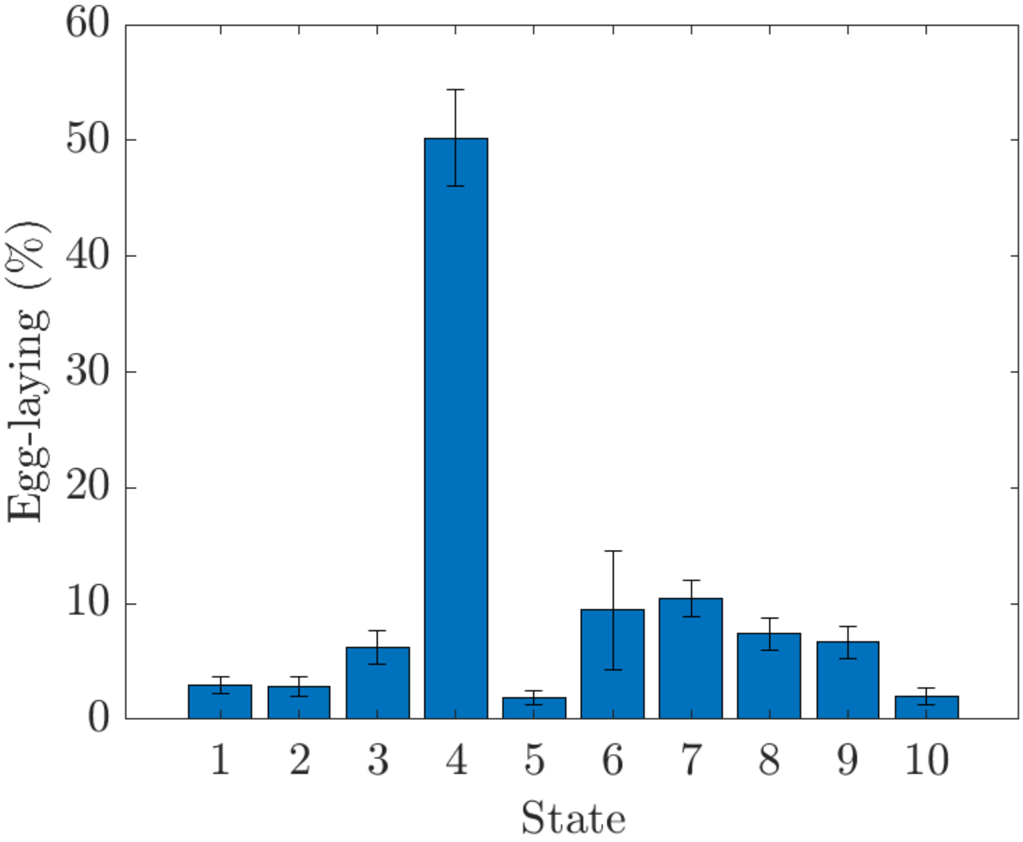

Figure 3-10. Percentage of eggs laid in each state: The percentage of eggs animals laid in each state (mean ± SEM)

While we have established that State 4 of the model corresponds to the roaming behavioral state, further work is needed to fully characterize the remaining states. Each of these states exhibits slower velocity and increased angular speed. Preliminary analysis suggests that these states are part of the dwelling state. Most likely, these nine states capture different aspects of dwelling.

3.4 Applications and future directions

3.4.1 Application of analysis

In this section, we apply the classifications found through the analytical method as previously described to mutant *C. elegans* with known behavioral phenotypes. In particular, we consider *cat-2*, *tbh-1*, and *tph-1* mutants. For this application, we use the same posture classifications and hidden Markov model found with the wild type data. We perform this analysis to confirm that this classification is generalizable and that known differences in behavioral states can be detected through this method. Additionally, we analyze egg-laying and locomotion coordination as these behaviors can be easily quantified and have been previously shown to be coordinated. Moreover, little is known about the mechanism governing egg-laying and locomotion coordination, so this coordination of behaviors is of particular interest.

First, we consider a strain with a mutation in the *cat-2* gene. The *cat-2* gene encodes tyrosine hydroxylase, an enzyme involved in the conversion of tyrosine to L-DOPA [32, 33]. As such, *cat-2* mutants have reduced levels of dopamine [33, 34, 35]. Behaviorally, *cat-2* mutants are known to exhibit increased roaming behavior [35]. We now analyze these *cat-2* mutants (N=10) using the state-analysis previously performed on wild type *C. elegans*.[2] The analysis indicates that *cat-2* mutants do in fact spend more time in a roaming state (Figure 3-11) than wild type *C. elegans*

[2] Strain: MT15620 *cat-2(n4547)* [35].

(Figure 3-6). Additionally, we see that these mutants have an increased velocity (Figure 3-12) when compared to the wild type (Figure 3-7), which is consistent with previous findings.

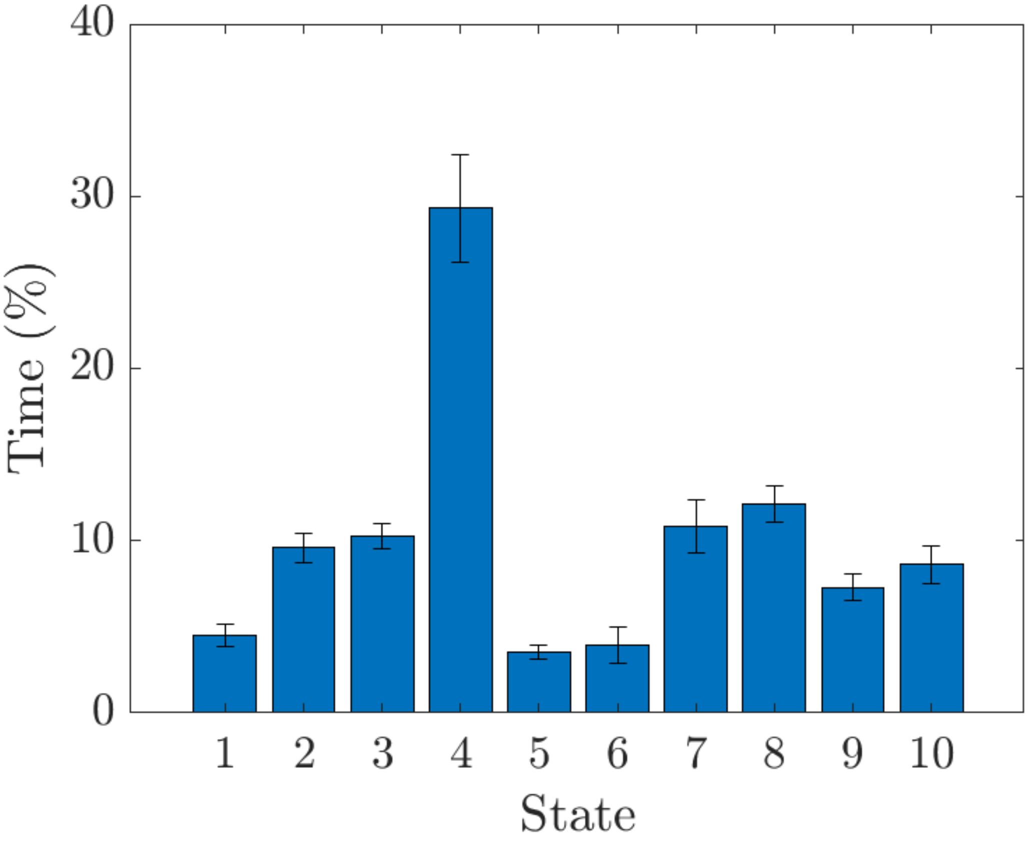

Figure 3-11. Percentage of time in states for *cat-2* mutants: The percentage of a recording the animals spent in each state (mean ± SEM)

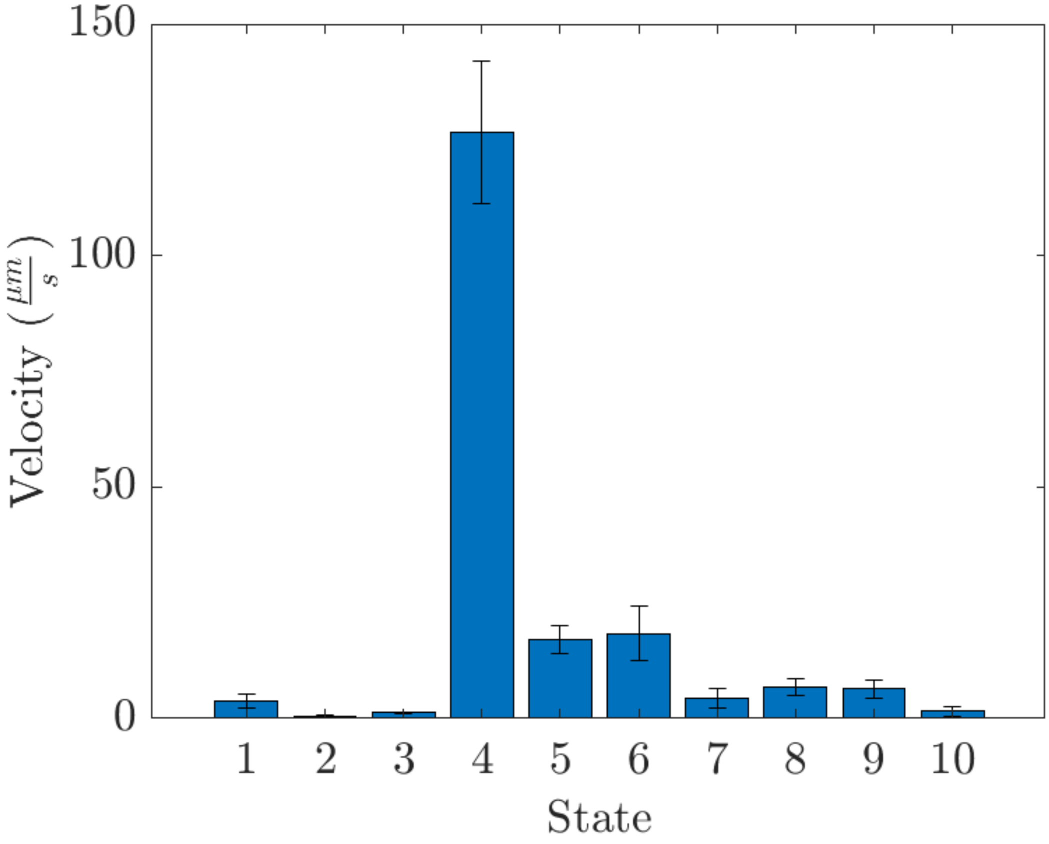

Figure 3-12. Velocity per state for *cat-2* mutants: The velocity animals exhibited in each state, given in μm per second (mean ± SEM)

Interestingly, we also see that egg-laying behavior and locomotion are no longer properly coordinated (Figure 3-13) when compared to the wild type (Figure 3-9). Specifically, egg-laying is reduced in the roaming state for *cat-2* mutants and the percentage of eggs laid in the roaming state is reduced (Figure 3-14) when compared to the wild type (Figure 3-10). To compensate, the percentage of eggs laid while dwelling is increased. This suggests that *cat-2* mutants display abnormal locomotion and egg-laying coordination.

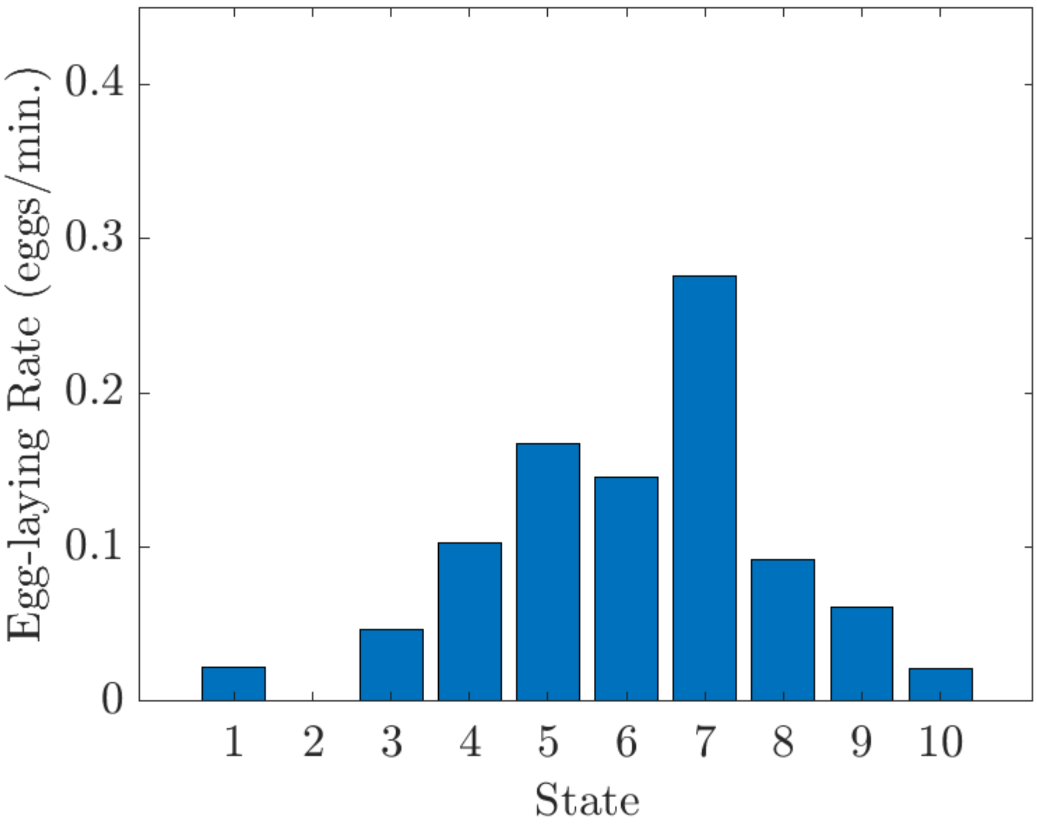

Figure 3-13. Egg-laying rate per state for *cat-2* mutants: The average egg-laying rate animals exhibited in each state, given in eggs per minute.

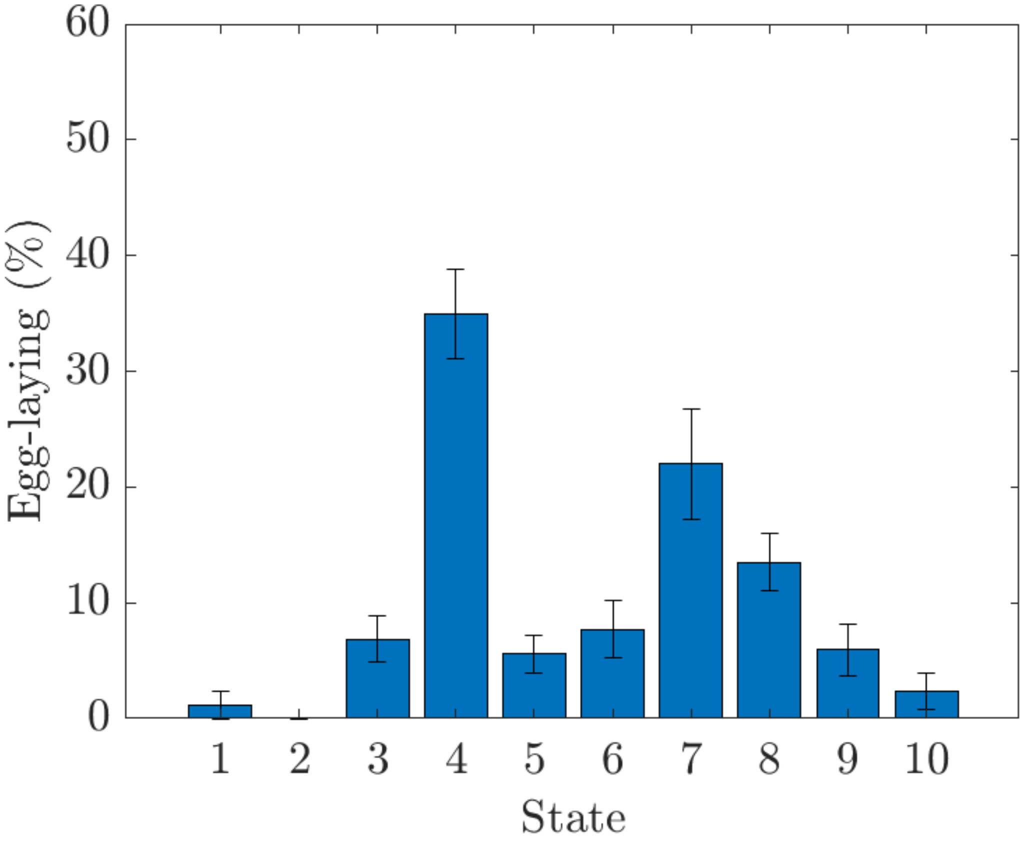

Figure 3-14. Percentage of eggs laid in each state for *cat-2* mutants: The percentage of eggs animals laid in each state (mean ± SEM)

Now, we analyze *tbh-1* mutants (N=10) using this framework.[3] The *tbh-1* gene encodes a protein that is involved in the conversion of tyramine to octopamine. Because egg-laying in *C. elegans* is inhibited by tyramine, *tbh-1* mutants display decreased egg-laying [36]. Indeed, we see this decrease in egg-laying in Figure 3-15, where there is an overall decrease in the egg-laying rates across the states. Unlike *cat-2*, however, the egg-laying rate continues to be highest in the roaming state (Figure 3-16) in a manner comparable to the wild type. Therefore, we do not detect the same abnormal

[3] Strain: MT9455 *tbh-1(n3247)*.

locomotion and egg-laying behavior coordination in *tbh-1* mutants.

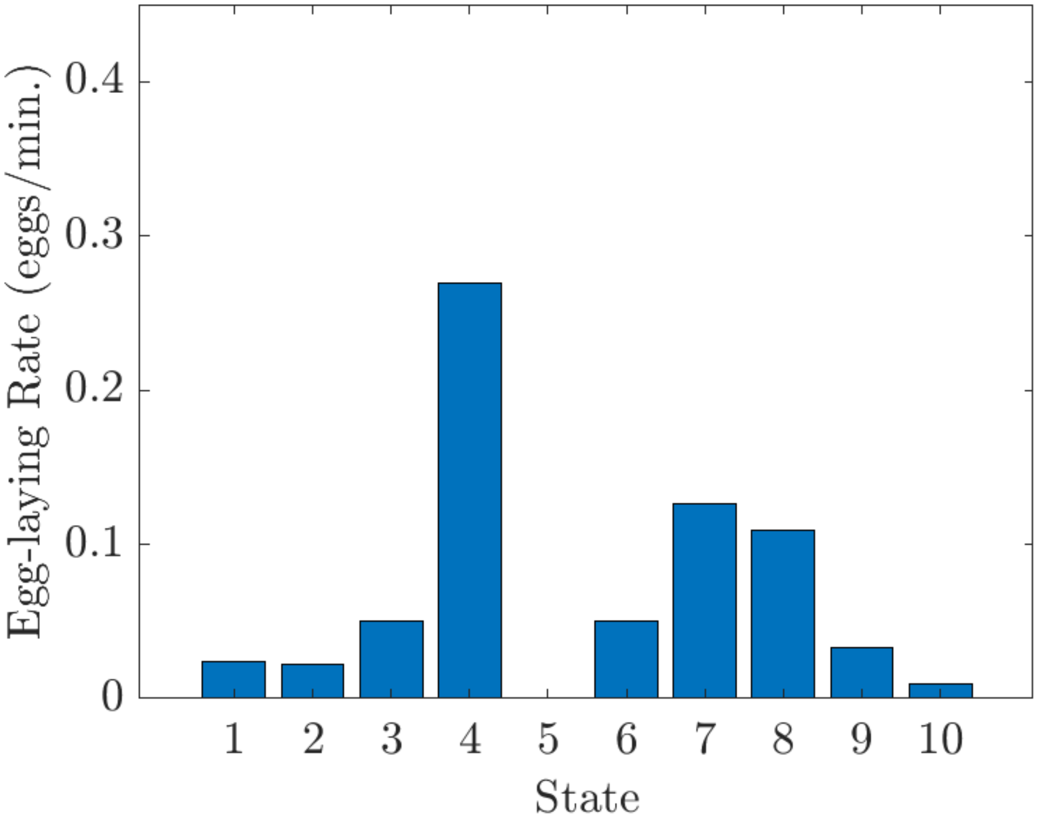

Figure 3-15. Egg-laying rate per state for *tbh-1* mutants: The average egg-laying rate animals exhibited per state, given in eggs per minute.

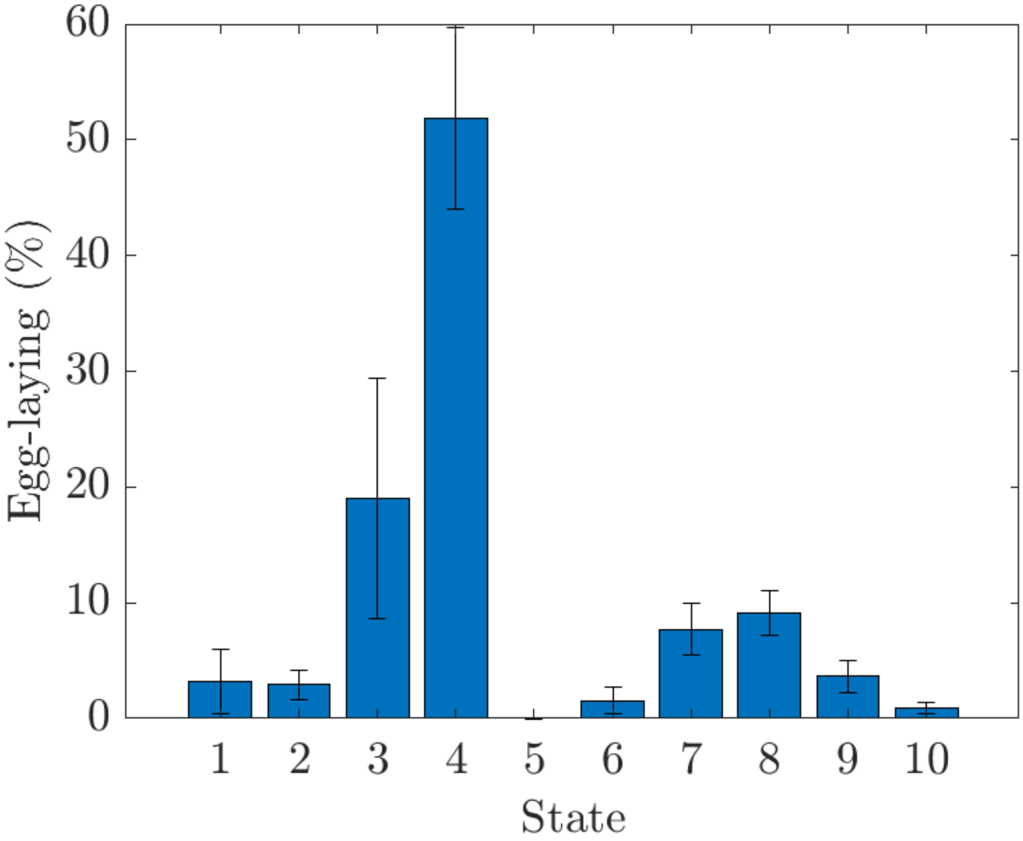

Figure 3-16. Percentage of eggs laid in each state for *tbh-1* mutants: The percentage of eggs animals laid in each state (mean ± SEM)

Finally, we analyze *tph-1* mutants (N=11).[4] The *tph-1* gene encodes a rate limiting enzyme for the synthesis of serotonin [37]. Behaviorally, *tph-1* mutants demonstrate increased exploratory behavior [38]. This behavioral phenotype is also captured by our analysis. These mutants spend more time in the roaming (Figure 3-17) state and demonstrate an increased roaming velocity (Figure 3-18). Furthermore, *tph-1* mutants display a reduction in egg-laying rate (Figure 3-19). This reduction in egg-laying appears to be fairly consistent across states and thus does not indicate

[4] Strain: MT15434 *tph-1(mg280)*.

abnormal locomotion and egg-laying coordination.

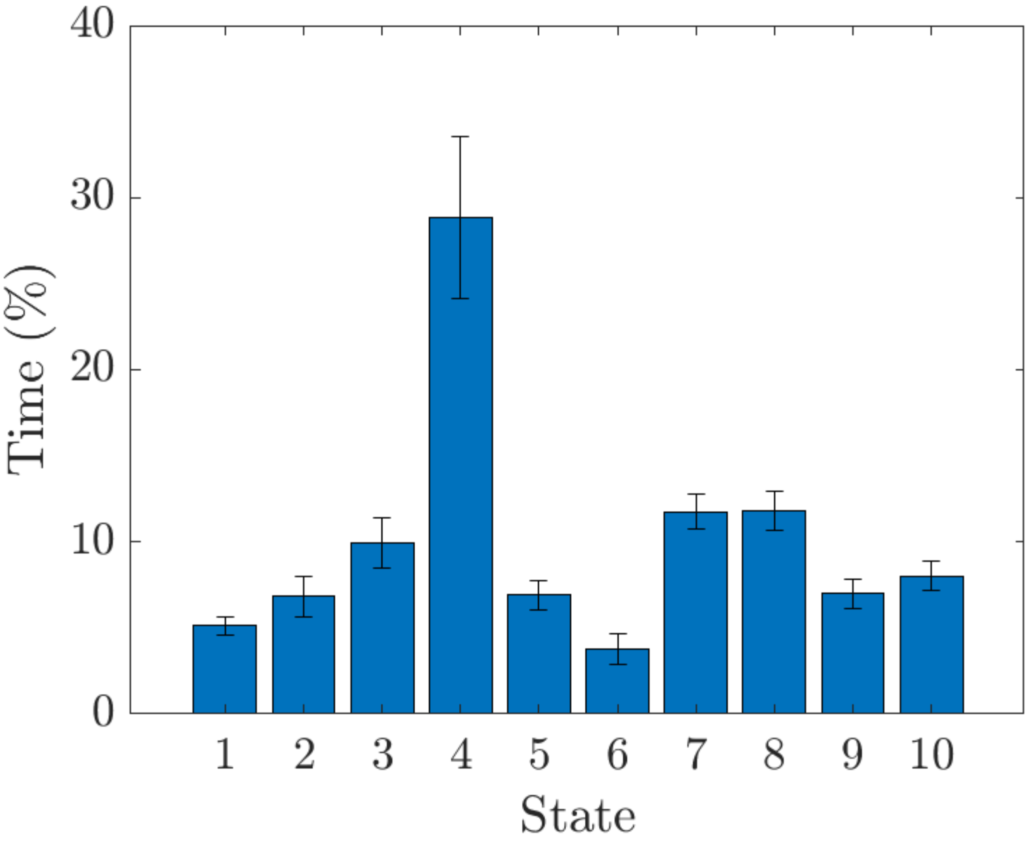

Figure 3-17. Percentage of time in states for *tph-1* mutants: The percentage of a recording the animals spent in each state (mean ± SEM)

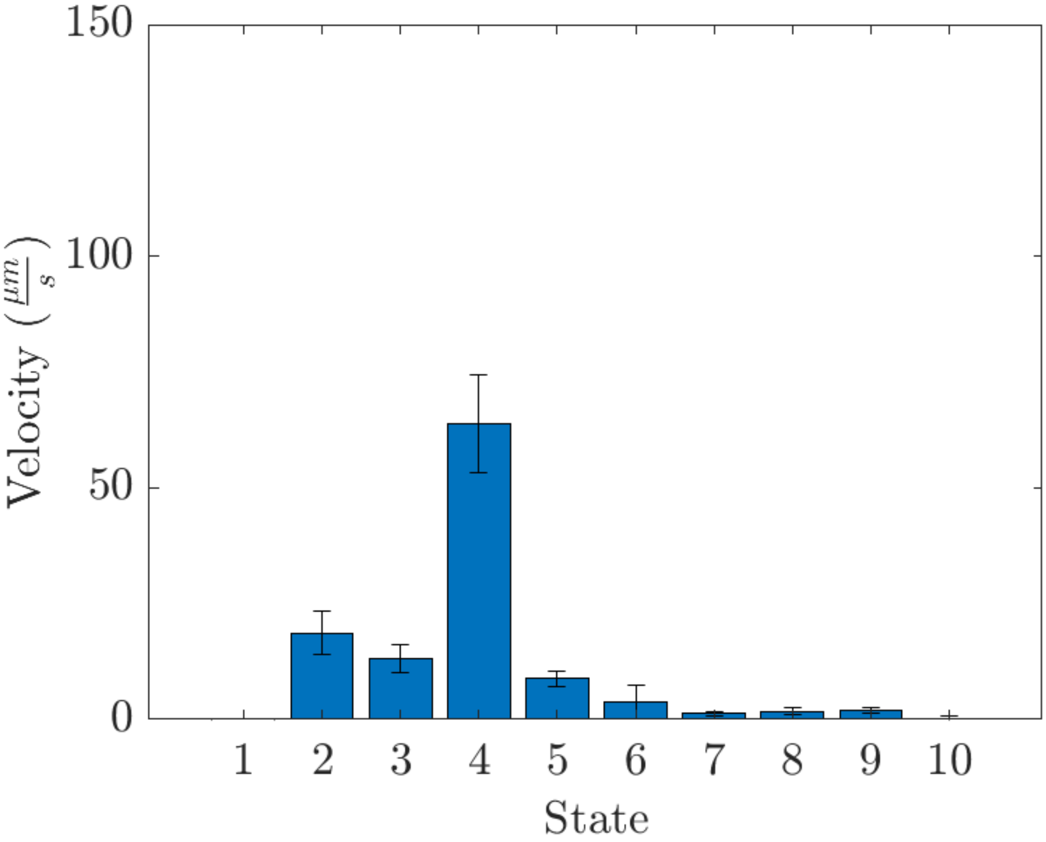

Figure 3-18. Velocity per state for *tph-1* mutants: The velocity animals exhibited in each state, given in μm per second (mean ± SEM)

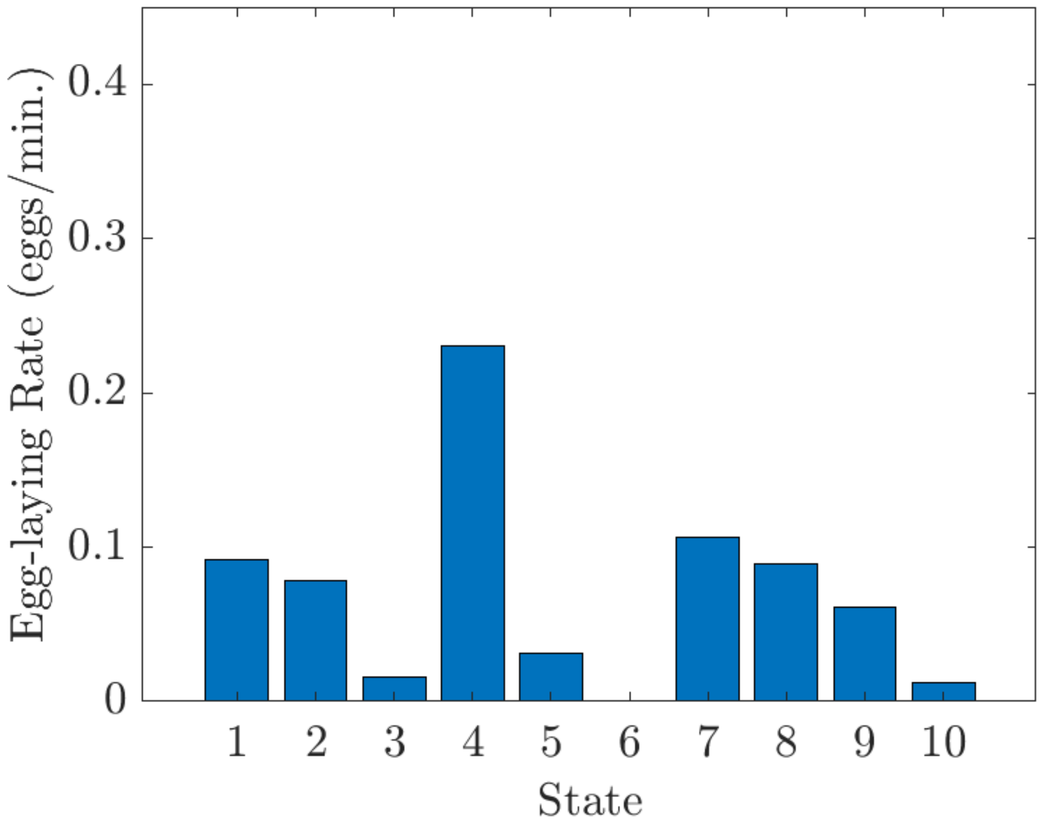

Figure 3-19. Egg-laying rate per state for *tph-1* mutants: The average egg-laying rate animals exhibited in each state, given in eggs per minute.

3.4.2 Future work

In this work, we have demonstrated a framework by which behavioral states can be detected for further behavior coordination study. The model extracted a roaming behavioral state as well as different dwelling state types, though more work is needed to fully classify these dwelling states. In the previous section, we applied our analytical framework to mutants with previously established behavioral phenotypes to confirm that our model extracts these same behaviors.

Using this method, we detected unusual locomotion and egg-laying behavior coordination in *cat-2* mutants wherein there was abnormal egg-laying behavior in the roaming state. Specifically, there was a decreased egg-laying rate in the roaming state and an increased egg-laying rate in dwelling states. These preliminary results suggest that dopamine may be involved in the coordination of egg-laying and locomotion.

The framework presented in this thesis can be used to analyze behavioral states in a data-driven, analytical manner. Specifically, this framework can detect interesting aspects of behavior coordination that would otherwise be difficult to extract. Such analyses may serve to facilitate and direct research regarding the biological processes underlying behavior coordination.

Bibliography

[1] Wood, William B. *The Nematode Caenorhabditis elegans.* Cold Spring Harbor Laboratory Press, 1988.

[2] Brenner, Sydney. The genetics of *Caenorhabditis elegans*. *Genetics*, 77(1):71–94, 1974.

[3] Chalfie, Martin, Robert Horvitz, and John E Sulston. Mutations that lead to reiterations in the cell lineages of *C. elegans*. *Cell*, 24(1):59–69, 1981.

[4] Trent, Carol, Nancy Tsung, and Robert Horvitz. Egg-laying defective mutants of the nematode *Caenorhabditis elegans*. *Genetics*, 104(4):619–647, 1983.

[5] Brenner, Sydney. Nobel lecture: nature's gift to science. *Bioscience reports*, 23(5):225–237, 2003.

[6] Horvitz, H Robert. Worms, life, and death (nobel lecture). *Chembiochem*, 4(8):697–711, 2003.

[7] Sulston, John E. *Caenorhabditis elegans*: the cell lineage and beyond (nobel lecture). *Chembiochem*, 4(8):688–696, 2003.

[8] Chalfie, Martin, Yuan Tu, Ghia Euskirchen, William W Ward, and Douglas C Prasher. Green fluorescent protein as a marker for gene expression. *Science*, 263(5148):802–805, 1994.

[9] Herculano-Houzel, Suzana. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in human neuroscience*, 3:31, November 2009.

[10] White, John, Erica Southgate, J. Nichol Thomson, and Sydney Brenner. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 314:1–340, 1986.

[11] Fujiwara, Manabi, Piali Sengupta, and Steven L. McIntire. Regulation of body size and behavioral state of *C. elegans* by sensory perception and the egl-4 cgmp-dependent protein kinase. *Neuron*, 36(6):1091–1102, 2002.

[12] Shtonda, Boris Borisovich, and Leon Avery. Dietary choice behavior in *Caenorhabditis elegans*. *Journal of experimental biology*, 209(1):89–102, 2006.

[13] Arous, Juliette Ben, Sophie Laffont, and Didier Chatenay. Molecular and sensory basis of a food related two-state behavior in *C. elegans*. *PloS one*, 4(10):e7584, 2009.

[14] Hardaker, Laura Anne, Emily Singer, Rex Kerr, Guotong Zhou, William R Schafer. Serotonin modulates locomotory behavior and coordinates egg-laying and movement in *Caenorhabditis elegans*. *Journal of neurobiology*, 49(4):303–313, 2001.

[15] Rankin, Catherine H, Christine DO Beck, and Catherine M Chiba. *Caenorhabditis elegans*: a new model system for the study of learning and memory. *Behavioural brain research*, 37(1):89–92, 1990.

[16] Wen, Joseph YM, Namit Kumar, Glenn Morrison, Gloria Rambaldini, Susan Runciman, Joyce Rousseau, and Derek van der Kooy. Mutations that prevent associative learning in *C. elegans*. *Behavioral neuroscience*, 111(2):354, 1997.

[17] Roussel, Nicolas, Christine A Morton, Fern P Finger, and Badrinath Roysam. A computational model for *C. elegans* locomotory behavior: application to multiworm tracking. *IEEE transactions on biomedical engineering*, 54(10):1786–1797, 2007.

[18] Stephens, Greg J, Bethany Johnson-Kerner, William Bialek, and William S Ryu. Dimensionality and dynamics in the behavior of *C. elegans*. *PLoS computational biology*, 4(4):e1000028, 2008.

[19] Brown, André, Eviatar I Yemini, Laura J Grundy, Tadas Jucikas, and William R Schafer. A dictionary of behavioral motifs reveals clusters of genes affecting *Caenorhabditis elegans* locomotion. *Proceedings of the National Academy of Sciences*, 110(2):791–796, 2013.

[20] Schwarz, Roland F, Robyn Branicky, Laura J Grundy, William R Schafer, and André EX Brown. Changes in postural syntax characterize sensory modulation and natural variation of *C. elegans* locomotion. *PLoS computational biology*, 11(8):e1004322, 2015.

[21] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016.

[22] MATLAB. *version 9.4.0 (R2018a).* The MathWorks Inc., Natick, Massachusetts, 2018.

[23] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[24] Ng, Andrew Y, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[25] Arthur, David and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[26] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[27] Baum, Leonard E and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[28] Baum, Leonard. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3:1–8, 1972.

[29] Rabiner, Lawrence R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[30] Viterbi, Andrew. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

[31] Schwarz, Gideon. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[32] Lints, Robyn and Scott W Emmons. Patterning of dopaminergic neurotransmitter identity among *Caenorhabditis elegans* ray sensory neurons by a tgfbeta family signaling pathway and a hox gene. *Development*, 126(24):5819–5831, 1999.

[33] Nagatsu, Toshiharu, Morton Levitt, and Sidney Udenfriend. Tyrosine hydroxylase the initial step in norepinephrine biosynthesis. *Journal of Biological Chemistry*, 239(9):2910–2917, 1964.

[34] Sulston, J, M Dew, and S Brenner. Dopaminergic neurons in the nematode *Caenorhabditis elegans. Journal of Comparative Neurology*, 163(2):215–226, 1975.

[35] Omura, Daniel T, Damon A Clark, Aravinthan DT Samuel, and H Robert Horvitz. Dopamine signaling is essential for precise rates of locomotion by *C. elegans. PLoS One*, 7(6):e38649, 2012.

[36] Alkema, Mark J, Melissa Hunter-Ensor, Niels Ringstad, and H Robert Horvitz. Tyramine functions independently of octopamine in the *Caenorhabditis elegans* nervous system. *Neuron*, 46(2):247–260, 2005.

[37] Sze, Ji Ying, Martin Victor, Curtis Loer, Yang Shi, and Gary Ruvkun. Food and metabolic signalling defects in a caenorhabditis elegans serotonin-synthesis mutant. *Nature*, 403(6769):560, 2000.

[38] Flavell, Steven W, Navin Pokala, Evan Z Macosko, Dirk R Albrecht, Johannes Larsch, and Cornelia I Bargmann. Serotonin and the neuropeptide pdf initiate and extend opposing behavioral states in c. elegans. *Cell*, 154(5):1023–1035, 2013.