Predictability and hierarchy in *Drosophila* behavior

Gordon J. Berman[a,b,1,2], William Bialek[a,b], and Joshua W. Shaevitz[a,b]

[a]Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544; and [b]Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544

Edited by Nigel Goldenfeld, University of Illinois at Urbana–Champaign, Urbana, IL, and approved September 6, 2016 (received for review May 11, 2016)

Even the simplest of animals exhibit behavioral sequences with complex temporal dynamics. Prominent among the proposed organizing principles for these dynamics has been the idea of a hierarchy, wherein the movements an animal makes can be understood as a set of nested subclusters. Although this type of organization holds potential advantages in terms of motion control and neural circuitry, measurements demonstrating this for an animal's entire behavioral repertoire have been limited in scope and temporal complexity. Here, we use a recently developed unsupervised technique to discover and track the occurrence of all stereotyped behaviors performed by fruit flies moving in a shallow arena. Calculating the optimally predictive representation of the fly's future behaviors, we show that fly behavior exhibits multiple time scales and is organized into a hierarchical structure that is indicative of its underlying behavioral programs and its changing internal states.

behavior | hierarchy | *Drosophila* | information bottleneck

Animals perform a vast array of behaviors as they go about their daily lives, often in what appear to be repeated and nonrandom patterns. These sequences of actions, some innate and some learned, have dramatic consequences with respect to survival and reproductive function: from feeding, grooming, and locomotion to mating, child rearing, and the establishment of social structures. Moreover, these patterns of movement can be viewed as the final output of the complicated interactions between an organism's genes, metabolism, and neural signaling. As a result, elucidating the principles that govern the generation of behavioral sequences provides a window into the biological mechanisms underlying an animal's movements, appetites, and interactions with its environment, potentially allowing for broader insights into how behaviors evolve.

The prevailing theory for the temporal organization of behavior, rooted in work from neuroscience, psychology, and evolution, is that the pattern of actions performed by animals is hierarchical (1–3). In such a framework, actions are nested into modules on many scales, from simple motor primitives to complex behaviors to sequences of actions. In the case of a fly grooming itself, for example, small movements of the leg and wing muscles are organized into grooming modules for a particular location of the body. These modules are then orchestrated into patterns that exhibit their own complicated dynamics, and this whole pattern is only a small part of the entirety of the animal's activities (4, 5). Additionally, neural architectures related to behavior, such as the motor cortex, are anatomically hierarchical, supporting the idea that animals use a hierarchical representation of behavior in the brain (6–9). Hierarchical organization is also a hallmark of human design, from the layout of cities to the wiring of the internet, and its potential use in various biological contexts has been proposed as an organizing principle (2).

Despite the theoretical attractiveness of behavioral hierarchy, measurements showing that a particular animal's behavioral repertoire is organized in this manner often are limited in their scope. Typically, observations of hierarchy in the ordering of movement have considered a single behavioral type, such as grooming, ignoring relationships between more varied behavioral motifs (4, 10–13). Perhaps more problematic is that most analyses of behavior make use of methods, such as hierarchical clustering, that implicitly or explicitly impose a hierarchical structure onto the data without showing that such a representation is accurate. Lastly, to our knowledge, all measurements of a hierarchical organization of behavior limit their analysis to behavioral dynamics at a single time scale. This scale is often given by the results of fitting a Markov model, where the next step in a behavioral pattern only depends on the animal's current state. Even in the simplest of animals, however, there are many internal states such as hunger, reproductive drive, etc., and sequences of behaviors possess an effective memory of an animal's behavioral state that persists well into the future, a result noted in a wide variety of systems (14–17).

In this paper, we study the behavioral repertoire of fruit flies (*Drosophila melanogaster*), attempting to characterize the temporal organization of their movements over the course of an hour. Decomposing the flies' movements into a set of stereotyped motions without making any a priori behavioral definitions (18), we find that their behavior exhibits long time scales, far beyond what would be predicted from a Markovian model. Applying methods from information theory, we show that a hierarchical representation of actions optimally predicts the future behavioral state of the fly. These results show that the best way to understand how future actions follow from the current behavioral state is to group these current behaviors in a hierarchically nested manner, with fine grained partitions being useful in predicting the near future, and coarser partitions being sufficient for predicting the relatively distant future. These results show that these animals control their movement via a hierarchy of behaviors at varying time scales, affirming and making precise a key concept in ethology.

Experiments and Behavioral States

As a testbed for probing questions of behavioral organization and hierarchy, we sought to measure the entire behavioral repertoire of a population of *D. melanogaster* in a specific environmental context. Specifically, we observed individual, ground-based fruit flies in a largely featureless circular arena for approximately 1 h using a 100-Hz camera. Although they are prevented from jumping or flying, the flies here display many complex behaviors, including locomotion and grooming, that involve multiple parts of their bodies interacting at varying time scales. We recorded videos of 59 male

Significance

How an animal chooses to order its activities—moving, grooming, resting, and so on—is essential to its ability to survive, adapt, and reproduce. Here we investigate the temporal pattern of behaviors performed by fruit flies, finding that their movements are organized in a hierarchical manner that exhibits long time scales. This organization is likely advantageous for adaptability and ease of neural representation and provides hints as to the form of the fly's internal representations of behavioral programs.

Author contributions: G.J.B., W.B., and J.W.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]Present address: Department of Biology, Emory University, Atlanta, GA 30322.

[2]To whom correspondence should be addressed. Email: gordon.berman@emory.edu.

flies using a custom-built tracking setup, producing more than 21 million images (18).

These data were used to generate a 2D map of fly behavior based on an unsupervised approach that automatically identifies stereotyped actions (Fig. 1*A*; for full details, see ref. 18). Briefly, this approach takes a set of translationally and rotationally aligned images of the flies and decomposes the dynamics of the observed pixel values into a low-dimensional basis set describing the flies' posture. Time series are produced by projecting the original pixel values onto this basis set, and the local spectrogram of these trajectories is then embedded into two dimensions (19). Each position in the behavioral map corresponds to a unique set of postural dynamics, with nearby points representing similar motions, i.e., those involving related body parts executing similar temporal patterns.

In the resulting behavioral space, $\mathbf{z}$, we estimate the probability distribution function $P(\mathbf{z})$ and find that it contains a set of peaks corresponding to short segments of movement that are revisited multiple times by multiple individuals (Fig. 1*A*). Pauses in the trajectories through this space, $\mathbf{z}(t)$, are interspersed with quick movements between the peaks. These pauses in $\mathbf{z}(t)$ at a particular peak correspond to the fly performing one of a large set of distinct, stereotyped behaviors such as right wing grooming, proboscis extension, or alternating tripod locomotion (18). In all, we identify 117 unique stereotyped actions, with similar behaviors, i.e., those that use similar body parts at similar frequencies, located near each other in the behavioral map. A watershed algorithm, combined with a threshold on $d\mathbf{z}(t)/dt$, is used to separate the peaks and to segment each movie into a sequence of discrete, stereotyped behaviors.

In this paper, we treat pauses at these peaks to be our states, the lowest level of description of behavioral organization, and investigate the pattern of behavioral transitions among these states over time. We count time in units of the transitions between states, so we have a description of behavior as a discrete variable $S(n)$ that can take on $N = 117$ different values at each discrete time n. Note that because we count time in units of transitions, we always have $S(n+1) \neq S(n)$. Combining data from all 59 flies, we observe a mean residency time in a behavioral state of 0.21 s and an average transition time between pauses at behavioral space peaks of 0.13 s. In total, we observe $\approx 6.4 \times 10^5$ behavioral transitions, or about 10^4 per experiment.

Transition Matrices and Non-Markovian Time Scales

To investigate the temporal pattern of behaviors, we first calculated the behavioral transition matrix over different time scales

$$[\mathbf{T}(\tau)]_{i,j} \equiv p(S(n+\tau) = i | S(n) = j), \quad [1]$$

which describes the probability that the animal will go from state j to state i after τ transition steps. We expect that this distribution becomes less and less structured as τ increases because we lose the ability to make predictions of the future state as the horizon of our predictions extends further. In addition, it will be useful to think about these matrices in terms of their eigendecompositions

$$[\mathbf{T}(\tau)]_{i,j} = \sum_{\mu} \lambda_\mu(\tau) u_i^\mu(\tau) v_j^\mu(\tau), \quad [2]$$

where $\mathbf{u}^\mu \equiv \{u_i^\mu\}$ and $\mathbf{v}^\mu \equiv \{v_i^\mu\}$ are the left and right eigenvectors, respectively, and $\lambda_\mu(\tau)$ is the eigenvalue with the μth largest modulus. Because probability is conserved in the transitions, the largest eigenvalue $\lambda_1(\tau) = 1$, and $\mathbf{v}^1(\tau)$ is proportional to the stationary distribution over states at long times. All of the other eigenvalues have magnitudes less than 1, $|\lambda_{k\neq 1}(\tau)| < 1$, and describe the loss of predictability over time, as shown in more detail below.

The matrix $\mathbf{T}(\tau = 1)$ describes the probability of transitions from one state to the next, the most elementary steps of behavior (Fig. 1*B*). To the eye, this transition matrix appears modular, with most transitions out of any given state only going to one of a handful of other states. By appropriately organizing the states in Fig. 1*B*, $\mathbf{T}(\tau = 1)$ takes on a nearly block-diagonal structure, which can be broken up into modular clusters using the information bottleneck

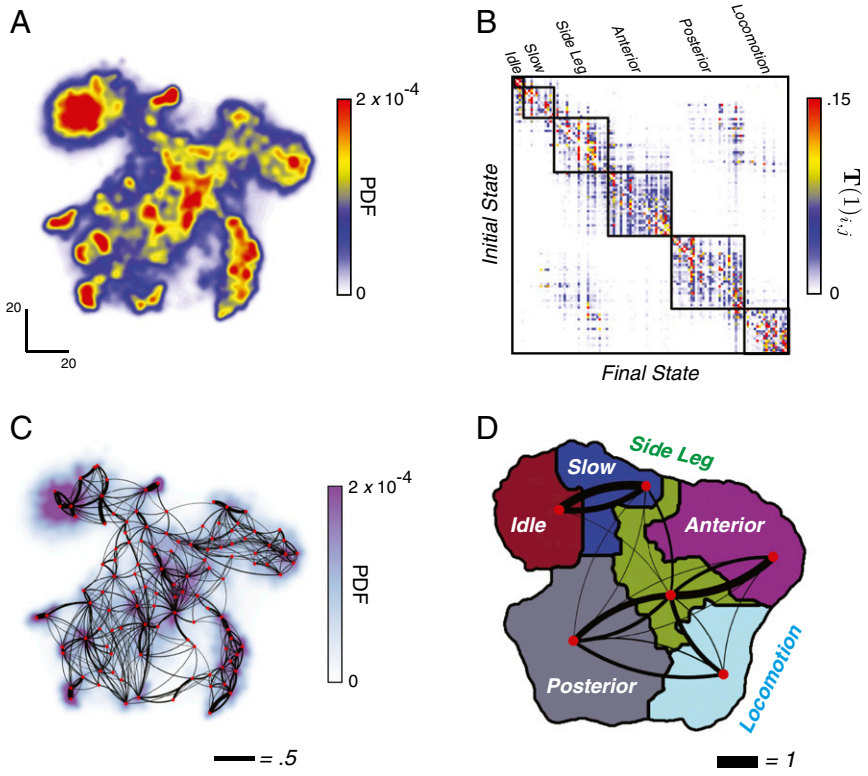

Fig. 1. Transition probabilities and behavioral modularity. (*A*) Behavioral space probability density function (PDF). Here, each peak in the distribution corresponds to a distinct stereotyped movement. (*B*) One-step Markov transition probability matrix $\mathbf{T}(\tau = 1)$. The 117 behavioral states are grouped by applying the predictive information bottleneck calculation and allowing six clusters (Eq. 4). Black lines denote the cluster boundaries. (*C*) Transitions rates plotted on the behavioral map. Each red point represents the maximum of the local PDF, and the black lines represent the transition probabilities between the regions. Line thicknesses are proportional to the corresponding value of $\mathbf{T}(\tau=1)_{ij}$, and right-handed curvature implies the direction of transmission. For clarity, all lines representing transition probabilities of less than 0.05 are omitted. (*D*) The clusters found using the information bottleneck approach (colored regions) are contiguous in the behavioral space. Behavioral labels associated with each partitioned graph cluster from *B* are shown. Black line thicknesses represent the conditional transition probabilities between clusters. All transition probabilities less than 0.05 are omitted.

formalism (*Materials and Methods*). Plotting this matrix on the behavioral map itself (Fig. 1*C*), we see that the transitions are largely localized, with nearly all large probability transitions occurring between nearby behaviors. Furthermore, the transition clusters are contiguous in the behavioral space, defining gross categories of motion including locomotion, behaviors involving anterior parts of the body, etc. (Fig. 1*D*).

It is important to note that $\mathbf{T}(\tau=1)$ does not directly contain information about the location of behavioral states in the 2D map, and hence any relationship we observe between the transition structure and the patterning of behaviors in the map is a consequence of the animal's behavior and not the way we construct the analysis. We thus conclude that behavioral transitions are mostly restricted to occur between similar actions—e.g., grooming behaviors are typically followed by other grooming behaviors of close body parts and animals transition between locomotion gates systematically by changing gate speed and velocity. These observations are consistent with classical ideas of postural facilitation and previous findings that transitions largely occur between similar behaviors (4, 20–22).

We begin to see the necessity of looking at longer time scales as we measure the transition matrices for $\tau \gg 1$. If the observed dynamics are purely Markovian, then the transitions from one state to the next do not depend on the history of behavior, and $\mathbf{T}(\tau=1)$ provides a complete characterization of the system. In particular, if the behavior is Markovian, then we can calculate the transition matrix after τ state just by iterating the matrix from one step

$$\mathbf{T}_M(\tau) \equiv [\mathbf{T}(1)]^\tau = \sum_\mu \left[\lambda_\mu(1)\right]^\tau \mathbf{u}^\mu(1)\mathbf{v}^\mu(1). \qquad [3]$$

Because $|\lambda_\mu(1)| < 1$ for all but the leading eigenvalue, the contributions from the $\mu > 1$ terms decay to zero exponentially as $\tau \to \infty$. For very long times, therefore, $\mathbf{T}_M(\tau)$ loses all information about the current state and instead reflects the average probabilities of performing any particular behavior. Thus, in a Markovian system, the slowest time scale in the system is determined by $|\lambda_2(1)|$, resulting in a characteristic decay time $t_2 = -1/\log|\lambda_2(1)|$. Calculating these eigenvalues for each fly and averaging, we find $\langle\lambda_2(1)\rangle = 0.953 \pm 0.004$, or $\langle t_2\rangle = 29 \pm 2$ transitions. Thus, any memory that extends beyond ≈ 30 transitions into the future is direct evidence for hidden states that carry a memory over longer times and modulate behavior.

Initial evidence for long-time structure in $\mathbf{T}(\tau)$ comes by comparing the lack of structure within $\mathbf{T}_M(100)$ to that within $\mathbf{T}(\tau)$ for $\tau = 100$ and $\tau = 1{,}000$ (Fig. 2 *A*–*C*). After 100 transitions, ($\approx 3\langle t_2\rangle$), the Markov model retains essentially no information, as demonstrated by the similarity between all of the rows, implying that all transitions have been randomized. Conversely, although some of the block–diagonal structure from Fig. 1*B* has

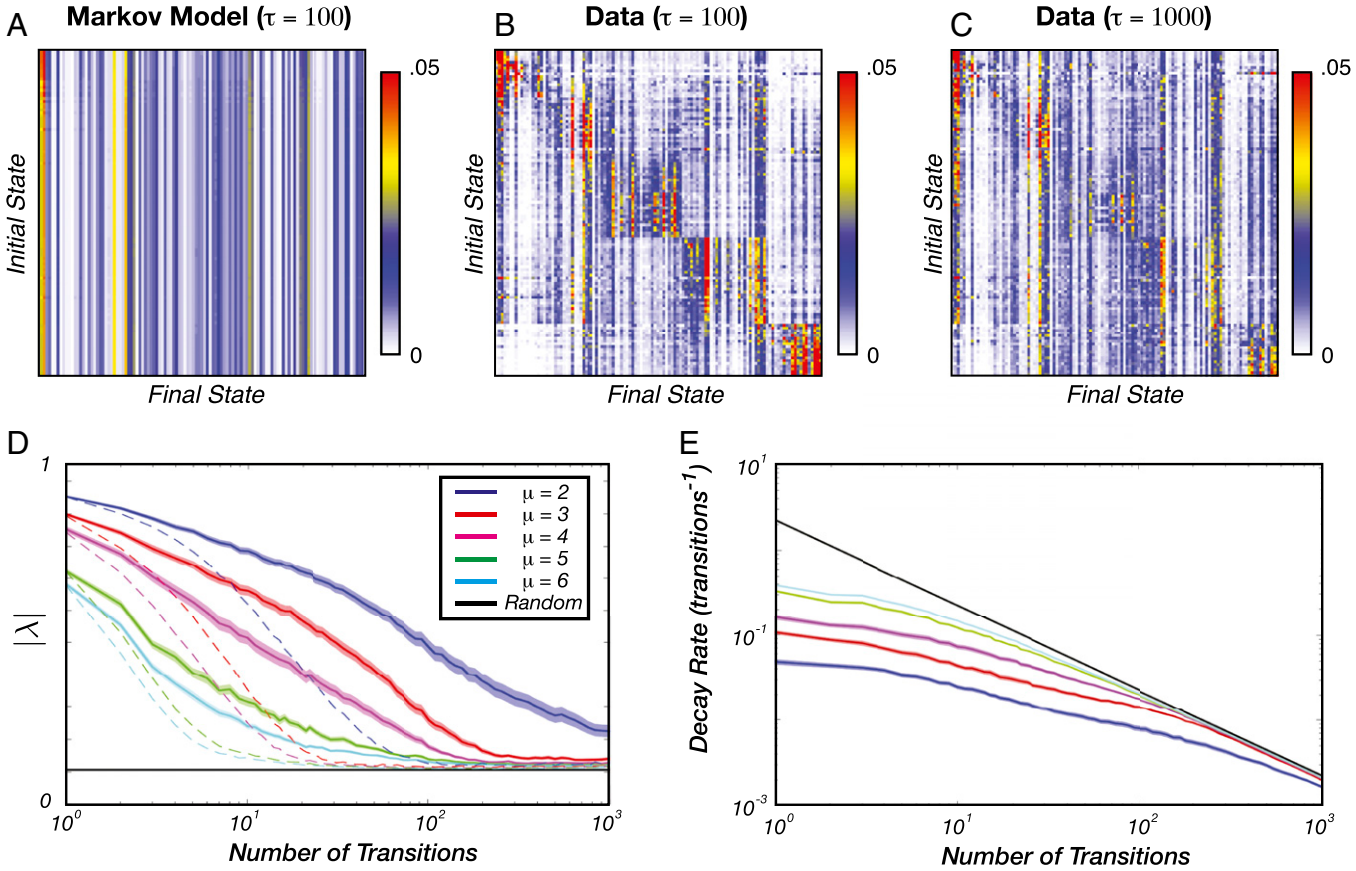

Fig. 2. Long time scale transition matrices and non-Markovian dynamics. (*A*) Markov model transition matrix for $\tau = 100$, $\mathbf{T}_M(100)$, from Eq. 3. (*B* and *C*) Transition matrices for $\tau = 100$ and $\tau = 1{,}000$, respectively, from Eq. **1**. (*D*) Absolute values of the leading eigenvalues of the transition matrices $\mathbf{T}(\tau)$ as a function of τ. The curves represent the average over all flies, and thicknesses represent the SEM. Dashed lines are the predictions for the Markov model $\mathbf{T}_M(\tau)$. The black line is a noise floor, corresponding to the typical value of the second largest eigenvalue in a transition matrix calculated from random temporal shuffling of our finite data set. (*E*) Eigenmode decay rates, $r_m u(\tau) \equiv -\log|\lambda_\mu(\tau)|/\tau$, as a function of the number of transitions. Line colors represent the same modes as in *D* and the black line again corresponds to a "noise floor," in this case, the largest decay rate that we can resolve above the random structures present in our finite sample.

dissipated, we see that $\mathbf{T}(100)$ and $\mathbf{T}(1{,}000)$ retain a great deal of nonrandomness.

This observation can be made more precise by looking at the eigenvalue spectra of the transition matrices. In Fig. 2*D*, we plot $|\lambda_\mu(\tau)|$ as a function of τ for $\mu = 2-6$ (solid color lines) in addition to the predictions from the Markov model of Eq. **3** based on $\mathbf{T}(1)$ (colored dashed lines). In a Markovian system, it would be more natural to plot these results with a logarithmic axis for λ, but here we see that structure extends over such a wide range of time scales that we need a logarithmic axis for τ. We can make this difference more obvious by measuring the apparent decay rate, $r_\mu(\tau) = -\log|\lambda_\mu(\tau)|/\tau$, which should be constant for a Markovian system. For the leading mode, the apparent decay rate falls by nearly two orders of magnitude before the corresponding eigenvalue is lost in the noise (Fig. 2*E*). Similar patterns appear in higher modes, but we have more limited dynamic range for observing them.

These results are direct evidence that many time scales are required to model behavioral sequences, even in this simple context where no external stimuli are provided. Accordingly, we can infer that the organism must have internal states that we do not directly observe, even though we are making rather thorough measurements of the motor output. Roughly speaking, the appearance of decay rates $\approx 10^{-3}$ means that the internal states must hold memory across at least $\approx 10^3$ behavioral transitions, or ~20 min—much longer than any time scale apparent in the Markov model.

Predictability and Hierarchy

The modular structure of the flies' transition matrix, combined with the observed long time scales of behavioral sequences, suggests that we might be able to group the behavioral states into clusters that preserve much of the information that the current behavioral state provides about future actions (predictive information) (23). Furthermore, we should be able to probe whether this results in a hierarchical organization of behaviors. If the states are grouped into a hierarchy, we expect that increasing the number of clusters will largely subdivide existing clusters rather than mix behaviors from two different clusters.

To make this idea more precise, we hope to map the behaviors into groups, $S(n) \to Z$, that compress our description in a way that preserves information about a state τ transitions in the future, $S(n+\tau)$. Mathematically, this means that we should maximize the information about the future, $I(Z;S(n+\tau))$, while holding fixed the information that we keep about the past, $I(Z;S(n))$.

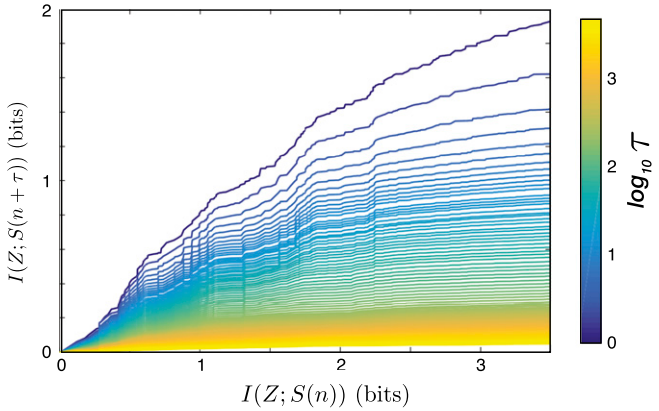

Fig. 3. Optimal trade-off curves for lags from $\tau = 1$ to $\tau = 5{,}000$. For each time lag τ, number of clusters, and β, we optimize Eq. **4** and plot the resulting complexity of the partitioning, $I(Z;S(n))$, vs. the predictive information, $I(Z;S(n+\tau))$.

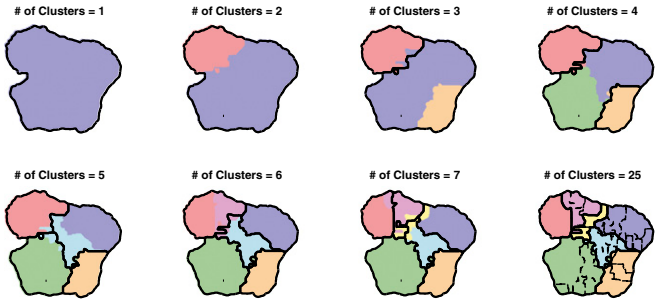

Fig. 4. Information bottleneck partitioning of behavioral space for $\tau = 67$ (approximately twice the longest time scale in the Markov model). Borders from the previous partitions are shown in black. For 25 clusters (*Bottom Right*), the partitions, still contiguous, are denoted by dashed lines.

Introducing a Lagrange multiplier to hold $I(Z;S(n))$ fixed, we wish to maximize

$$\mathcal{F} = I(Z;S(n+\tau)) - \beta I(Z;S(n)). \quad [4]$$

At $\beta = 0$ we retain the full complexity of the 117 behavioral states, and as we increase β, we are forced to tighten our description into a more and more compressed form, thus losing predictive power. This maximization of $\mathcal{F}$ is an example of the information bottleneck algorithm (24). Changing β and the number of clusters allows us to move along a curve that trades complexity of description against predictive power (see *Materials and Methods* for numerical details).

The tradeoff curves resulting from optimizing Eq. **4** for a variety of time scales are shown in Fig. 3. As expected, the optimal curves move downward as the time lag increases, implying that the ability to predict the behavioral state of the animal decreases as we look further into the future. We also observe a relatively rapid decrease in the height of these curves for small τ, followed by increasingly closely spaced optimal curves as the lag length increases. This slowing indicates the presence of long time scales in behavior.

Along each of these tradeoff curves lie partitions of the behavioral space that contain an increasing number of clusters. We can make several observations about these data. First, in agreement with our investigation of the single-step transition matrix, we find that the clusters are spatially contiguous in the behavioral map as exemplified in Fig. 4 for $\tau = 67$. Thus, even when we add in the long time-scale dynamics, we find that transitions predominantly occur between similar behaviors. Second, these spatially contiguous clusters separate hierarchically as we increase the number of clusters, i.e., new clusters largely result from subdividing existing clusters instead of emerging from multiple existing clusters. One example of this can be seen in Fig. 5, where the probability flow between partitions of increasing size subdivide in a tree-like manner. It is important to note that these results are not an artifact of the information bottleneck algorithm; we can solve the bottleneck problem for different numbers of clusters independently, and hence (in contrast to hierarchical clustering), this method could have found nonhierarchical evolution with new clusters comprised of behaviors from many other clusters. That this does not happen is strong evidence that fly behavior is organized hierarchically.

We can go beyond this qualitative description, however, by quantifying the degree of hierarchy in our representation as the number of clusters increases using a "treeness" metric, $\mathcal{T}$ (Fig. 6). The idea behind this metric, which is similar to the one introduced by Corominas-Murtra et al. (25), is that if our representation is perfectly hierarchical, then each cluster has precisely one "parent" in a partitioning with a smaller number of clusters. Thus, the better our ability to distinguish the lineage of a cluster as it splits through increasingly complex partitionings, the higher

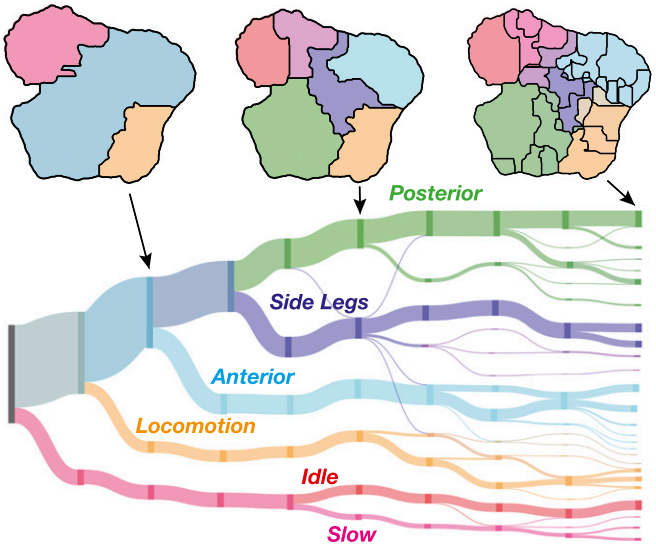

Fig. 5. Hierarchical organization for optimal solutions with lag $\tau = 100$ ranging from 1 cluster to 25. The displayed clusterings are those that have the largest value of $I(Z; S(n+\tau))$ for that number of clusters. The length of the vertical bars are proportional to the percentage of time a fly spends in each of the clusters, and the lines flowing horizontally from left to right are proportional in thickness to the flux from the clustering on the left to the clustering on the right. Fluxes less than 0.01 are suppressed for clarity.

$\mathcal{T}$ becomes. More precisely, the treeness index is given by the relative reduction in entropy going backward rather than forward through the tree

$$\mathcal{T} = \frac{\mathcal{H}_f - \mathcal{H}_b}{\mathcal{H}_f}, \qquad [5]$$

where $\mathcal{H}_f$ and $\mathcal{H}_b$ are the entropies over all possible paths going forward and backward, respectively. This metric is bounded between 0 and 1, and $\mathcal{T} = 1$ implies a perfect hierarchy.

We find that the partitionings derived from the information bottleneck algorithm are much more tree-like than random partitions of the behavioral space (Fig. 6*B*). This result holds even when we attempt to optimally predict behavioral states thousands of transitions into the future. Thus, by finding optimally predictive representations that best explain the relationship between states over long time scales, we have uncovered a hierarchical ordering of actions, supporting decades-old theory without relying on hierarchical clustering, Markov models, or limiting the measured behavioral repertoire.

Conclusions

We measured the behavioral repertoires for dozens of fruit flies, paying particular attention to the structure of their behavioral transitions. We find that these transitions exhibit multiple time scales and possess memory that persists for tens of minutes, indicative of internal states that carry memory across thousands of observable behavioral transitions. Using an information bottleneck approach to find the compressed representations that optimally predict our observed dynamics, we find that behaviors are organized in a hierarchical fashion, with fine grained representations being able to predict short time structure and coarser representations being sufficient to predict the fly's actions that are further removed in time. This finding is fundamentally different from previous measurements of hierarchy in behavior, which were more limited in the types of behaviors they measured, the time scales over which the hierarchy was modeled, and/or relied on hierarchical clustering and other types of analyses that only yield hierarchical outputs.

The type of organization we observe is reminiscent of the functional clustering seen in mouse and primate motor cortex, where groupings of neurons from millimeter scales down to single cells have been found to exhibit increasing temporal correlation as the distance between them decreases (6, 8). Although no such correlation has been specifically found in *Drosophila*, our results suggest that such neuronal patterns may exist: perhaps by combining descending commands from the brain with local circuitry within and emerging from the ventral nerve cord. As circuits for different behavioral modules are uncovered, our results suggest that such hierarchical neuroanatomical organization will also be found in the fly, serving as a general principle that may apply across organisms to provide insight toward how the brain controls behavior and adapts to a complex environment.

Materials and Methods

Experiments. We imaged 59 individual male flies (*D. melanogaster*, Oregon-R strain) for 1 h each, following the protocols originally described in ref. 18. All flies were within the first 2 wk after eclosion during the filming session. Flies were placed into the arena via aspiration and were subsequently allowed 5 min for adaptation before data collection. All recording occurred between the hours of 9:00 AM and 1:00 PM. The temperature during all recordings was 25 ± 1 °C.

Behavioral States. The observed behavioral space was generated following the methods originally described in ref. 18, including image segmentation and alignment, projection of image data onto a set of postural eigenmodes, wavelet transformation, and low-dimensional embedding using t-distributed Stochastic Neighbor Embedding (19). Behaviors were assigned by smoothing the embedded points and performing a watershed transform (26) on the inverse of the density. Behavioral epochs were defined

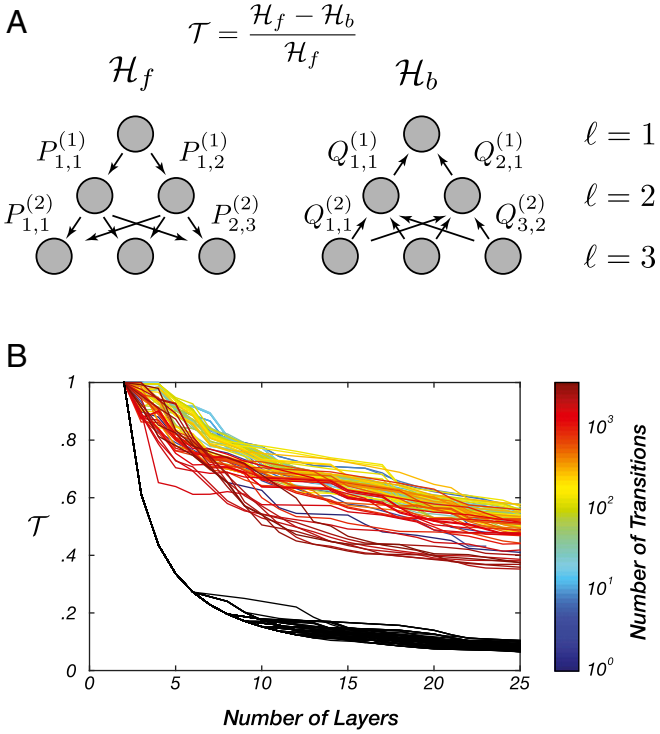

Fig. 6. Partitionings are tree-like over all measured time scales. (*A*) Definition of the treeness metric, $\mathcal{T}$; see *Materials and Methods* for details. (*B*) $\mathcal{T}$ as a function of the number of transitions in the future and the number of clusters in the most fine-grained partition. Colored lines represent values of $\mathcal{T}$ for partitions at varying times in the future, and black lines are values for randomized graphs generated from partitionings that were assigned randomly.

as time periods of low speed in the behavioral space (as defined by a Gaussian mixture model) and were required to be at least 0.05 s in duration.

Generating Markovian Models. Markovian model data were generated by randomly selecting a state, and then finding another, randomly chosen, instance in the measured data set where the fly was performing that behavior. The behavior performed immediately after that behavior is chosen, and the process is iterated until the generated sequence is equivalent in size to the original dataset, similar to the first-order alphabets generated in Shannon's original work on information theory (27).

Predictive Information Bottleneck. The solution to the information bottleneck problem, Eq. **4**, obeys a set of self-consistent equations that can be iterated in a manner equivalent to the Blahut–Arimoto algorithm in rate–distortion theory (24, 28). For a given $|Z| = K$ and inverse temperature β, a random initial condition for $p(z|x)$ is chosen, and the following self-consistent equations are iterated until the convergence criterion $[(\mathcal{F}_t - \mathcal{F}_{t+1})/\mathcal{F}_t < 10^{-6}]$ is met

$$p(z|x) = \frac{p(z)}{\mathcal{Z}(\beta, x)} \exp[-\beta D_{KL}(p(y|x) \| p(y|z))], \quad [6]$$

$$p(z) = \sum_x p(z|x)p(x), \quad [7]$$

$$p(y|z) = p(y|x)p(z|x)p(x), \quad [8]$$

where $x \in S(n)$, $y \in S(n+\tau)$, $z \in Z$, and D_{KL} is the Kullback–Leibler divergence between two probability distributions, and $\mathcal{Z}(\beta, x)$ is a normalizing function.

As this study focuses on hard clusterings of behavioral states, we find solutions by starting at $\beta = 0.1$ and annealing with 40 exponentially spaced values up to $\beta = 500$. Starting from a random initial condition at the initial value of β, the optimization is performed at that value until the convergence criterion is met and that solution is used as the initial condition for the next value of β. All intermediate solutions, $p_\ell^{(n)}(z|x)$ are stored so they can potentially be included in the found Pareto front. We perform 24 replicates of this, with different random initial conditions, for $K = 2, \ldots, 25$ and for 81 time lag values between $n = 1$ and $n = 5{,}000$.

Provided the set of solutions for a given lag, we take the deterministic limit of each clustering ($p(z|x) = \delta_{z,\mathrm{argmax}_{z'} p(z'|x)}$) and recalculate $I(Z;S(n))$ and $I(Z;S(n+\tau))$ accordingly. The Pareto front, $\xi^{(n)}$, is defined as the set of all solutions, $p_\ell^{(n)}(z|x)$, such that no other solution for that given lag results in a smaller value for $I(Z;S(n))$ and a larger value for $I(Z;S(n+\tau))$. Between 150 and 350 solutions were found for all of the fronts. When choosing a clustering for a fixed number of clusters, here, we always pick the representation along the optimal front that has the highest value of $I(Z;S(n+\tau))$.

Treeness Index. To calculate the treeness index, $\mathcal{T}$, we construct a directed, acyclic graph that connects the partitions as the number of clusters increases for a given time lag with values $P_{ij}^{(\ell)}$. These values are the probability that a state contained in one cluster, i, in the partitioning with ℓ clusters also belongs to cluster j in the partitioning with $\ell+1$ clusters. Similarly, we can create the backward graph, $Q_{ij}^{(\ell)}$, that links clusters in the opposite direction; $Q_{ij}^{(\ell)}$ is the probability that a state in cluster i in the partitioning with $\ell+1$ clusters also belongs to the cluster j in the partitioning containing ℓ clusters.

Thus, we can calculate the entropy of picking a path, $\pi^{(f)}$ in the forward direction vs. the entropy of picking a path, $\pi^{(b)}$ in the backward direction. These probabilities can be calculated via $p(\pi_{\mathbf{v}}^{(f)}) = \prod_{\ell=1}^{N-1} P_{v_\ell, v_{\ell+1}}^{(\ell)}$ and $\mathrm{p}(\pi_{\mathbf{v}}^{(b)})$, with $\mathbf{v}$ being a chosen sequence of clusters. We define the forward and backward entropies as follows:

$$\mathcal{H}_f = -\sum_{\mathbf{v} \in \mathbf{V}} p\left(\pi_{\mathbf{v}}^{(f)}\right) \log p\left(\pi_{\mathbf{v}}^{(f)}\right), \quad [9]$$

$$\mathcal{H}_b = \left\langle -\sum_{\mathbf{w} \in \mathbf{W}_r} p\left(\pi_{\mathbf{w}}^{(b)}\right) \log p\left(\pi_{\mathbf{w}}^{(b)}\right) \right\rangle_r, \quad [10]$$

where $\mathbf{V}$ is the set of all possible paths, and $\mathbf{W}_r$ is the set of all paths ending at cluster r in the most fine-grained partitioning. $\langle \cdots \rangle_r$ denotes an average over each end state. $\mathcal{T}$ is then calculated as the relative reduction in entropy between backward and forward path probability distributions, as given by Eq. **5**.

ACKNOWLEDGMENTS. We thank Ugne Klibaite, David Schwab, and Thibaud Taillefumier for discussions and suggestions. J.W.S. and G.J.B. also acknowledge the Aspen Center for Physics, where many ideas for this work were formulated. This work was funded through awards from the NIH (GM098090 and GM071508), National Science Foundation (PHY-1305525, PHY-1451171, and CCF-0939370), the Swartz Foundation, and the Simons Foundation.

1. Tinbergen N (1951) *The Study of Instinct* (Oxford Univ Press, Oxford, UK).
2. Dawkins R (1976) *Hierarchical Organization: A Candidate Principle for Ethology in Growing Points in Ethology*, eds Bateson P, Hinde R (Cambridge Univ Press, Cambridge, UK), pp 7–54.
3. Simon HA (1962) The architecture of complexity. *Proc Am Philos Soc* 106(6):467–482.
4. Dawkins R, Dawkins MS (1976) Hierarchical organization and postural facilitation: Rules for grooming in flies. *Anim Behav* 24(4):739–755.
5. Seeds AM, et al. (2014) A suppression hierarchy among competing motor programs drives sequential grooming in *Drosophila*. *eLife* 3:e02951.
6. Graziano MSA, Aflalo TN (2007) Mapping behavioral repertoire onto the cortex. *Neuron* 56(2):239–251.
7. Bassett DS, et al. (2008) Hierarchical organization of human cortical networks in health and schizophrenia. *J Neurosci* 28(37):9239–9248.
8. Dombeck DA, Graziano MS, Tank DW (2009) Functional clustering of neurons in motor cortex determined by cellular resolution imaging in awake behaving mice. *J Neurosci* 29(44):13751–13760.
9. Chen CH, et al. (2012) Hierarchical genetic organization of human cortical surface area. *Science* 335(6076):1634–1636.
10. Davis WJ, Mpitsos GJ, Siegler MV, Pinneo JM, Davis KB (1974) Neuronal substrates of behavioral hierarchies and associative learning in Pleurobranchaea. *Am Zool* 14(3):1037–1050.
11. Lefebvre L (1981) Grooming in crickets: Timing and hierarchical organization. *Anim Behav* 29(4):973–984.
12. Lefebvre L (1982) The organization of grooming in budgerigars. *Behav Processes* 7(2): 93–106.
13. Lefebvre L, Joly R (1982) Organization rules and timing in Kestrel grooming. *Anim Behav* 30(4):1020–1028.
14. Miller GA (2003) The cognitive revolution: A historical perspective. *Trends Cogn Sci* 7(3):141–144.
15. Heiligenberg W (1973) Random processes describing the occurrence of behavioural patterns in a cichlid fish. *Anim Behav* 21(1):169–182.
16. Jin DZ, Kozhevnikov AA (2011) A compact statistical model of the song syntax in Bengalese finch. *PLOS Comput Biol* 7(3):e1001108.
17. Dawkins R, Dawkins M (1973) Decisions and the uncertainty of behaviour. *Behaviour* 45(1):83–103.
18. Berman GJ, Choi DM, Bialek W, Shaevitz JW (2014) Mapping the stereotyped behaviour of freely moving fruit flies. *J R Soc Interface* 11(99):20140672.
19. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. *J Mach Learn Res* 9(2579-2605):85.
20. Takahata M, Yoshino M, Hisada M (1981) The association of uropod steering with postural movement of the abdomen in crayfish. *J Exp Biol* 92(1):341–345.
21. Ackermann H, Scholz E, Koehler W, Dichgans J (1991) Influence of posture and voluntary background contraction upon compound muscle action potentials from anterior tibial and soleus muscle following transcranial magnetic stimulation. *Electroencephalogr Clin Neurophysiol Evoked Potentials Sect* 81(1):71–80.
22. Hopkins B, Rönnqvist L (2002) Facilitating postural control: Effects on the reaching behavior of 6-month-old infants. *Dev Psychobiol* 40(2):168–182.
23. Bialek W, Nemenman I, Tishby N (2001) Predictability, complexity, and learning. *Neural Comput* 13(11):2409–2463.
24. Tishby N, Pereira FC, Bialek W (1999) The information bottleneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, eds Hajek B, Sreenivas RS (Univ of Illinois Press, Urbana-Champaign, IL), pp 368–377.
25. Corominas-Murtra B, Rodríguez-Caso C, Goñi J, Solé R (2011) Measuring the hierarchy of feedforward networks. *Chaos* 21(1):016108–016111.
26. Meyer F (1994) Topographic distance and watershed lines. *Signal Process* 38(1):113–125.
27. Shannon CE (1948) The mathematical theory of communication. *Bell Syst Tech J* 27(3): 379–423.
28. Blahut RE (1972) Computation of channel capacity and rate-distortion functions. *IEEE Trans Inf Theory* IT-18(4):460–473.