Searching for simplicity in the analysis of neurons and behavior

Greg J. Stephens[a,1], Leslie C. Osborne[b], and William Bialek[a]

[a]Joseph Henry Laboratories of Physics, Lewis–Sigler Institute for Integrative Genomics and Princeton Center for Theoretical Sciences, Princeton University, Princeton, NJ 08544; and [b]Department of Neurobiology, University of Chicago, Chicago, IL 60637

Edited by Donald W. Pfaff, The Rockefeller University, New York, NY, and approved January 20, 2011 (received for review October 21, 2010)

What fascinates us about animal behavior is its richness and complexity, but understanding behavior and its neural basis requires a simpler description. Traditionally, simplification has been imposed by training animals to engage in a limited set of behaviors, by hand scoring behaviors into discrete classes, or by limiting the sensory experience of the organism. An alternative is to ask whether we can search through the dynamics of natural behaviors to find explicit evidence that these behaviors are simpler than they might have been. We review two mathematical approaches to simplification, dimensionality reduction and the maximum entropy method, and we draw on examples from different levels of biological organization, from the crawling behavior of *Caenorhabditis elegans* to the control of smooth pursuit eye movements in primates, and from the coding of natural scenes by networks of neurons in the retina to the rules of English spelling. In each case, we argue that the explicit search for simplicity uncovers new and unexpected features of the biological system and that the evidence for simplification gives us a language with which to phrase new questions for the next generation of experiments. The fact that similar mathematical structures succeed in taming the complexity of very different biological systems hints that there is something more general to be discovered.

maximum entropy models | stochastic dynamical systems

The last decades have seen an explosion in our ability to characterize the microscopic mechanisms—the molecules, cells, and circuits—that generate the behavior of biological systems. In contrast, our characterization of behavior itself has advanced much more slowly. Starting in the late 19th century, attempts to quantify behavior focused on experiments in which the behavior itself was restricted, for example by forcing an observer to choose among a limited set of alternatives. In the mid-20th century, ethologists emphasized the importance of observing behavior in its natural context, but here, too, the analysis most often focused on the counting of discrete actions. Parallel to these efforts, neurophysiologists were making progress on how the brain represents the sensory world by presenting simplified stimuli and labeling cells by preference for stimulus features.

Here we outline an approach in which living systems naturally explore a relatively unrestricted space of motor outputs or neural representations, and we search directly for simplification within the data. Although there is often suspicion of attempts to reduce the evident complexity of the brain, it is unlikely that understanding will be achieved without some sort of compression. Rather than restricting behavior (or our description of behavior) from the outset, we will let the system "tell us" whether our favorite simplifications are successful. Furthermore, we start with high spatial and temporal resolution data because we do not know the simple representation ahead of time. This approach is made possible only by the combination of experimental methods that generate larger, higher-quality data sets with the application of mathematical ideas that have a chance of discovering unexpected simplicity in these complex systems. We present four very different examples in which finding such simplicity informs our understanding of biological function.

Dimensionality Reduction

In the human body there are ≈100 joint angles and substantially more muscles. Even if each muscle has just two states (rest or tension), the number of possible postures is enormous, $2^{N_{\mathrm{muscles}}} \sim 10^{30}$. If our bodies moved aimlessly among these states, characterizing our motor behavior would be hopeless—no experiment could sample even a tiny fraction of all of the possible trajectories. Moreover, wandering in a high dimensional space is unlikely to generate functional actions that make sense in a realistic context. Indeed, it is doubtful that a plausible neural system would independently control all of the muscles and joint angles without some coordinating patterns or "movement primatives" from which to build a repertoire of actions. There have been several motor systems in which just such a reduction in dimensionality has been found (1–5). Here we present two examples of behavioral dimensionality reduction that represent very different levels of system complexity: smooth pursuit eye movements in monkeys and the free wiggling of worm-like nematodes. These examples are especially compelling because so few dimensions are required for a complete description of natural behavior.

Smooth Pursuit Eye Movements. Movements are variable even if conditions are carefully repeated, but the origin of that variability is poorly understood. Variation might arise from noise in sensory processing to identify goals for movement, in planning or generating movement commands, or in the mechanical response of the muscles. The structure of behavioral variation can inform our understanding of the underlying system if we can connect the dimensions of variation to a particular stage of neural processing.

Like other types of movement, eye movements are potentially high dimensional if eye position and velocity vary independently from moment to moment. However, an analysis of the natural variation in smooth pursuit eye movement behavior reveals a simple structure whose form suggests a neural origin for the noise that gives rise to behavioral variation. Pursuit is a tracking eye movement, triggered by image motion on the retina, which serves to stabilize a target's retinal image and thus to prevent motion blur (6). When a target begins to move relative to the eye, the pursuit system interprets the resulting image motion on the retina to estimate the target's trajectory and then to accelerate the eye to match the target's motion direction and speed. Although tracking on longer time scales is driven by both retinal inputs and by extraretinal feedback signals, the initial ≈125 ms of the movement is generated purely from sensory estimates of the target's motion,

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Quantification of Behavior" held June 11–13, 2010, at the AAAS Building in Washington, DC. The complete program and audio files of most presentations are available on the NAS Web site at www.nasonline.org/quantification.

Author contributions: G.J.S., L.C.O., and W.B. designed research; G.J.S., L.C.O., and W.B. performed research; G.J.S., L.C.O., and W.B. contributed new reagents/analytic tools; G.J.S., L.C.O., and W.B. analyzed data; and G.J.S., L.C.O., and W.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: gstephen@princeton.edu.

using visual inputs present before the onset of the response. Focusing on just this initial portion of the pursuit movement, we can express the eye velocity in response to steps in target motion as a vector, $\mathbf{v}(t) = v_H(t)\hat{\mathbf{i}} + v_V(t)\hat{\mathbf{j}}$,where $v_H(t)$ and $v_V(t)$ are the horizontal and vertical components of the velocity, respectively. In Fig. 1*A* we show a single trial velocity trajectory (horizontal and vertical components, dashed black and gray lines) and the trial-averaged velocity trajectory (solid black and gray lines). Because the initial 125 ms of eye movement is sampled every millisecond, the pursuit trajectories have 250 dimensions.

We compute the covariance of fluctuations about the mean trajectory and display the results in Fig. 1*B*. Focusing on a 125-ms window at the start of the pursuit response (green box), we compute the eigenvalues of the covariance matrix and find that only the three largest are statistically different from zero according to the SD within datasets (7). This low dimensional structure is not a limitation of the motor system, because during fixation (yellow box) there are 80 significant eigenvalues. Indeed, the small amplitude, high dimensional variation visible during fixation seems to

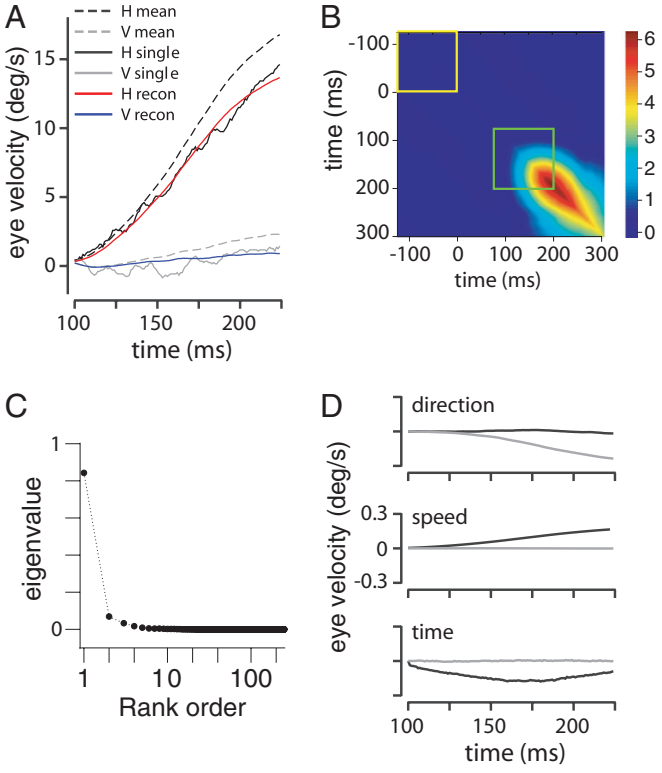

Fig. 1. Low-dimensional dynamics of pursuit eye velocity trajectories (7). (*A*) Eye movements were recorded from male rhesus monkeys (*Macaca mulatta)* that had been trained to fixate and track visual targets. Thin black and gray lines represent horizontal (H) and vertical (V) eye velocity in response to a step in target motion on a single trial; dashed lines represent the corresponding trial-averaged means. Red and blue lines represent the model prediction. (*B*) Covariance matrix of the horizontal eye velocity trajectories. The yellow square marks 125 ms during the fixation period before target motion onset, the green square the first 125 ms of pursuit. The color scale is in deg/s^2. (*C*) Eigenvalue spectrum of the difference matrix $\Delta C(t, t') = C_{\text{pursuit}}(t, t')$ (green square) – $C_{\text{background}}(t, t')$ (yellow square). (*D*) Time courses of the sensory error modes ($\mathbf{v}_{\text{dir}}$, $\mathbf{v}_{\text{speed}}$, $\mathbf{v}_{\text{time}}$). The sensory error modes are calculated from derivatives of the mean trajectory, as in Eq. **1**, and linear combinations of these modes can be used to reconstruct trajectories on single trials as shown in *A*. These modes have 96% overlap with the significant dimensions that emerge from the covariance analysis in *B* and *C* and thus provide a nearly complete description of the behavioral variation. Black and gray curves correspond to H and V components.

be an ever-present background noise that is swamped by the larger fluctuations in movement specific to pursuit. If the covariance of this background noise is subtracted from the covariance during pursuit, the 3D structure accounts for ~94% of the variation in the pursuit trajectories (Fig. 1*C*).

How does low dimensionality in eye movement arise? The goal of the movement is to match the eye to the target's velocity, which is constant in these experiments. The brain must therefore interpret the activity of sensory neurons that represent its visual inputs, detecting that the target has begun to move (at time t_0) and estimating the direction θ and speed v of motion. At best, the brain estimates these quantities and transforms these estimates into some desired trajectory of eye movements, which we can write as $\mathbf{v}(t; \hat{t_0}, \hat{\theta}, \hat{v})$, where $\hat{\cdot}$ denotes an estimate of the quantity $\cdot$. However, estimates are never perfect, so we should imagine that $\hat{t_0} = t_0 + \delta t_0$, and so on, where δt_0 is the small error in the sensory estimate of target motion onset on a single trial. If these errors are small, we can write

$$\mathbf{v}(t) = \mathbf{v}(t; t_0, v, \theta) + \delta t_0 \frac{\partial \mathbf{v}(t; t_0, v, \theta)}{\partial t_0} + \delta\theta \frac{\partial \mathbf{v}(t; t_0, v, \theta)}{\partial \theta} + \delta v \frac{\partial \mathbf{v}(t; t_0, v, \theta)}{\partial \mathbf{v}} + \delta v_{\text{back}}(t), \quad [1]$$

where the first term is the average eye movement made in response to many repetitions of the target motion, the next three terms describe the effects of the sensory errors, and the final term is the background noise. Thus, if we can separate out the effects of the background noise, the fluctuations in $\mathbf{v}$ (**t**) from trial to trial should be described by just three random numbers, δt_0, δθ, and δv: the variations should be 3D, as observed.

The partial derivatives in Eq. **1** can be measured as the difference between the trial-averaged pursuit trajectories in response to slightly different target motions. In fact the average trajectories vary in a simple way, shifting along the t axis as we change t_0, rotating in space as we change θ, and scaling uniformly faster or slower as we change v (7), so that the relevant derivatives can be estimated just from one average trajectory. We identify these derivatives as sensory error modes and show the results in Fig. 1*D*, where we have abbreviated the partial derivative expressions for the modes of variation as $\mathbf{v}_{\text{dir}} \equiv \partial\mathbf{v}/(t;\, t_0,\, v,\, \theta)/\partial\theta$, $\mathbf{v}_{\text{speed}} \equiv \partial\mathbf{v}/(t;\, t_0,\, v,\, \theta)/\partial v$, and $\mathbf{v}_{\text{time}} \equiv \partial\mathbf{v}/(t;\, t_0,\, v,\, \theta)/\partial t_0$. We note that each sensory error mode has a vertical and horizontal component, although some make little contribution. We recover the sensory errors (δθ, δv, δt_0) by projecting the pursuit trajectory on each trial onto the corresponding sensory error mode.

We can write the covariance of fluctuations around the mean pursuit trajectory in terms of these error modes as

$$C_{\text{ij}}(t,t') = \begin{bmatrix} \mathbf{v}^{(\text{i})}_{\text{dir}}(t) \\ \mathbf{v}^{(\text{i})}_{\text{speed}}(t) \\ \mathbf{v}^{(\text{i})}_{\text{time}}(t) \end{bmatrix}^T \begin{bmatrix} \langle\delta\theta\delta\theta\rangle & \langle\delta\theta\delta v\rangle & \langle\delta\theta\delta t_0\rangle \\ \langle\delta v\delta\theta\rangle & \langle\delta v\delta v\rangle & \langle\delta v\delta t_0\rangle \\ \langle\delta t_0\delta\theta\rangle & \langle\delta t_0\delta v\rangle & \langle\delta t_0\delta t_0\rangle \end{bmatrix} \begin{bmatrix} \mathbf{v}^{(\text{j})}_{\text{dir}}(t') \\ \mathbf{v}^{(\text{j})}_{\text{speed}}(t') \\ \mathbf{v}^{(\text{j})}_{\text{time}}(t') \end{bmatrix} + C^{(\text{back})}_{\text{ij}}(t,t'), \quad [2]$$

where the terms $\{\langle\delta\theta\delta\theta\rangle, \langle\delta\theta\delta v\rangle, \ldots\}$ are the covariances of the sensory errors. The fact that C can be written in this form implies not only that the variations in pursuit will be 3D but that we can predict in advance what these dimensions should be. Indeed, we find experimentally that the three significant dimensions of C have 96% overlap, with axes corresponding to $\mathbf{v}_{\text{dir}}$, $\mathbf{v}_{\text{speed}}$, and $\mathbf{v}_{\text{time}}$.

These results strongly support the hypothesis that the observable variations in motor output are dominated by the errors that the brain makes in estimating the parameters of its sensory

inputs, as if the rest of the processing and motor control circuitry were effectively noiseless, or more precisely that they contribute only at the level of background variation in the movement. Further, the magnitude and time course of noise in sensory estimation are comparable to the noise sources that limit perceptual discrimination (7, 8). This unexpected result challenges our intuition that noise in the execution of movement creates behavioral variation, and it forces us to consider that errors in sensory estimation may set the limit to behavioral precision. Our findings are consistent with the idea that the brain can minimize the impact of noise in motor execution in a task-specific manner (9, 10), although it suggests a unique origin for that noise in the sensory system. The precision of smooth pursuit fits well with the broader view that the nervous system can approach optimal performance at critical tasks (11–14).

How the Worm Wiggles. The free motion of the nematode *Caenorhabditis elegans* on a flat agar plate provides an ideal opportunity to quantify the (reasonably) natural behavior of an entire organism (15). Under such conditions, changes in the worm's sinuous body shape support a variety of motor behaviors, including forward and backward crawling and large body bends known as Ω-turns (16). Tracking microscopy provides high spatial and temporal resolution images of the worm over long periods of time, and from these images we can see that fluctuations in the thickness of the worm are small, so most variations in the shape are captured by the curve that passes through the center of the body. We measure position along this curve (arc length) by the variable s, normalized so that $s = 0$ is the head and $s = 1$ is the tail. The position of the body element at s is denoted by x(s), but it is more natural to give an "intrinsic" description of this curve in terms of the tangent angle θ(s), removing our choice of coordinates by rotating each image so that the mean value of θ along the body always is zero. Sampling at $n = 100$ equally spaced points along the body, each shape is described completely by a 100-dimensional vector (Fig. 2 *A* and *B*).

As we did with smooth pursuit eye movements, we seek a low dimensional space that underlies the shapes we observe. In the simplest case, this space is a Euclidean projection of the original high dimensional space so that the covariance matrix of angles, $C(s, s') = \langle(\theta(s) - \langle\theta\rangle)\,(\theta(s') - \langle\theta\rangle)\rangle$, will have only a small number of significant eigenvalues. For *C. elegans* this is exactly what we find, as shown in Fig. 2 *C* and *D*: more than 95% of the variance in body shape is accounted for by projections along just four dimensions ("eigenworms," red curves in Fig. 2*C*). Further, the trajectory in this low dimensional space of shapes predicts the motion of the worm over the agar surface (17). Importantly, the simplicity that we find depends on our choice of initial representation. For example, if we take raw images of the worm's body, cropped to a minimum size (300 × 160 pixels) and aligned to remove rigid translations and rotations, the variance across images is spread over hundreds of dimensions.

The tangent angle representation and projections along the eigenworms provide a compact yet substantially complete description of worm behavior. In distinction to previous work (see, e.g., refs. 16, 18, and 19), this description is naturally aligned to the organism, fully computable from the video images with no human intervention, and also simple. In the next section we show how these coordinates can be also used to explore dynamical questions posed by the behavior of *C. elegans*.

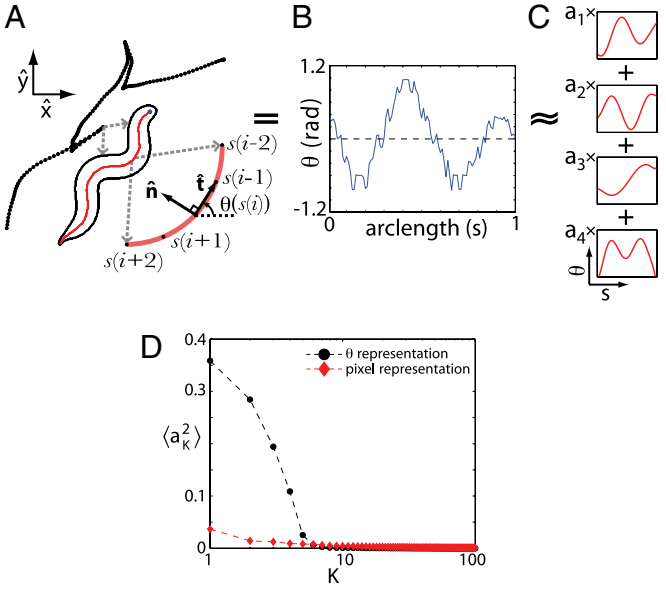

Fig. 2. Low-dimensional space of worm postures (15). (*A*) We use tracking video microscopy to record images of the worm's body at high spatiotemporal resolution as it crawls along a flat agar surface. Dotted lines trace the worm's centroid trajectory, and the body outline and centerline skeleton are extracted from the microscope image on a single frame. (*B*) We characterize worm shape by the tangent angle θ vs. arc length s of the centerline skeleton. (*C*) We decompose each shape into four dominant modes by projecting θ (s) along the eigenvectors of the shape covariance matrix (eigenworms). (*D*, black circles) Fraction of total variance captured by each projection. The four eigenworms account for ≈95% of the variance within the space of shapes. (*D*, red diamonds) Fraction of total variance captured when worm shapes are represented by images of the worm's body; the low dimensionality is hidden in this pixel representation.

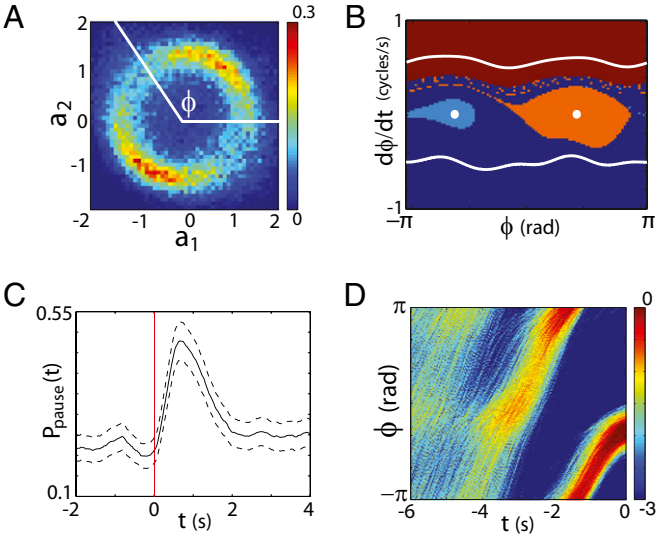

Fig. 3. Worm behavior in the eigenworm coordinates. (*A*) Amplitudes along the first two eigenworms oscillate, with nearly constant amplitude but time-varying phase $\phi = \tan^{-1}(a_2/a_1)$. The shape coordinate $\phi(t)$ captures the phase of the locomotory wave moving along the worm's body. (*B*) Phase dynamics from Eq. **3** reveals attracting trajectories in worm motion: forward and backward limit cycles (white lines) and two instantaneous pause states (white circles). Colors denote the basins of attraction for each attracting trajectory. (*C*) In an experiment in which the worm receives a weak thermal impulse at time $t = 0$, we use the basins of attraction of *B* to label the instantaneous state of the worm's behavior and compute the time-dependent probability that a worm is in either of the two pause states. The pause states uncover an early-time stereotyped response to the thermal impulse. (*D*) Probability density of the phase [plotted as log $P(\phi|t)$], illustrating stereotyped reversal trajectories consistent with a noise-induced transition from the forward state. Trajectories were generated using Eq. **3** and aligned to the moment of a spontaneous reversal at $t = 0$.

Dynamics of Worm Behavior. We have found low dimensional structure in the smooth pursuit eye movements of monkeys and in the free wiggling of nematodes. Can this simplification inform our understanding of behavioral dynamics—the emergence of discrete behavioral states, and the transitions between them? Here we use the trajectories of *C. elegans* in the low dimensional space to construct an explicit stochastic model of crawling behavior and then show how long-lived states and transitions between them emerge naturally from this model.

Of the four dimensions in shape space that characterize the crawling of *C. elegans*, motions along the first two combine to form an oscillation, corresponding to the wave that passes along the worm's body and drives it forward or backward. Here, we focus on the phase of this oscillation, $\phi = \tan^{-1}(a_2/a_1)$ (Fig. 3*A*), and construct, from the observed trajectories, a stochastic dynamical system, analogous to the Langevin equation for a Brownian particle. Because the worm can crawl both forward and backward, the phase dynamics is minimally a second-order system,

$$\frac{d\phi}{dt} = \omega, \quad \frac{d\omega}{dt} = F(\omega, \phi) + \sigma(\omega, \phi)\eta(t), \qquad [3]$$

where ω is the phase velocity and $\eta(t)$ is the noise—a random component of the phase acceleration not related to the current state of the worm—normalized so that $\langle \eta(t)\eta(t')\rangle = \delta(t - t')$. As explained in ref. 15, we can recover the "force" $F(\omega, \phi)$ and the local noise strength $\sigma(\omega, \phi)$ from the raw data, so no further "modeling" is required.

Leaving aside the noise, Eq. **3** describes a dynamical system in which there are multiple attracting trajectories (Fig. 3*B*): two limit cycle attractors corresponding to forward and backward crawling (white lines) and two pause states (white circles) corresponding to an instantaneous freeze in the posture of the worm. Thus, underneath the continuous, stochastic dynamics we find four discrete states that correspond to well-defined classes of behavior. We emphasize that these behavioral classes are emergent—there is nothing discrete about the phase time series $\phi(t)$, nor have we labeled the worm's motion by subjective criteria. Whereas forward and backward crawling are obvious behavioral states, the pauses are more subtle. Exploring the worm's response to gentle thermal stimuli, one can see that there is a relatively high probability of a brief sojourn in one of the pause states (Fig. 3*C*). Thus, by identifying the attractors—and the natural time scales of transitions between them—we uncover a more reliable component of the worm's response to sensory stimuli (15).

The noise term generates small fluctuations around the attracting trajectories but more dramatically drives transitions among the attractors, and these transitions are predicted to occur with stereotyped trajectories (20). In particular, the Langevin dynamics in Eq. **3** predict spontaneous transitions between the attractors that correspond to forward and backward motion. To quantify this prediction, we run long simulations of the dynamics, choose moments in time when the system is near the forward attractor ($0.1 < d\phi/dt < 0.6$ cycles/s), and then compute the probability that the trajectory has not reversed ($d\phi/dt < 0$) after a time τ following this moment. If reversals are rare, this survival probability should decay exponentially, $P(\tau) = \exp(-\tau/\langle\tau\rangle)$, and this is what we see, with the predicted mean time to reverse $\langle\tau\rangle = 15.7 \pm 2.1$ s, where the error reflects variations across an ensemble of worms.

We next examine the real trajectories of the worms, performing the same analysis of reversals by measuring the survival probability in the forward crawling state. We find that the data obey an exponential distribution, as predicted by the model, and the experimental mean time to reversal is $\langle\tau_{\rm data}\rangle = 16.3 \pm 0.3$ s. This observed reversal rate agrees with the model predictions within error bars, and this corresponds to a precision of ~4%, which is quite surprising. It should be remembered that we make our model of the dynamics by analyzing how the phase and phase velocity at the time t evolve into phase and phase velocity at time $t + dt$, where the data determine $dt = 1/32$ s. Once we have the stochastic dynamics, we can use them to predict the behavior on long time time scales. Although we define our model on the time scale of a single video frame (dt), behavioral dynamics emerge that are nearly three orders of magnitude longer ($\langle\tau\rangle/dt \approx 500$), with no adjustable parameters (20).

In this model, reversals are noise-driven transitions between attractors, in much the same way that chemical reactions are thermally driven transitions between attractors in the space of molecular structures (21). In the low noise limit, the trajectories that carry the system from one attractor to another become stereotyped (22). Thus, the trajectories that allow the worm to escape from the forward crawling attractor are clustered around prototypical trajectories, and this is seen both in the simulations (Fig. 3*D*) and in the data (20).

In fact, many organisms, from bacteria to humans, exhibit discrete, stereotyped motor behaviors. A common view is that these behaviors are stereotyped because they are triggered by specific commands, and in some cases we can even identify "command neurons" whose activity provides the trigger (23). In the extreme, discreteness and stereotypy of the behavior reduces to the discreteness and stereotypy of the action potentials generated by the command neurons, as with the escape behaviors in fish triggered by spiking of the Mauthner cell (24). However, the stereotypy of spikes itself emerges from the continuous dynamics of currents, voltages, and ion channel populations (25, 26). The success here of the stochastic phase model in predicting the observed reversal characteristics of *C. elegans* demonstrates that stereotypy can also emerge directly from the dynamics of the behavior itself.

Maximum Entropy Models of Natural Networks

Much of what happens in living systems is the result of interactions among large networks of elements—many amino acids interact to determine the structure and function of proteins, many genes interact to define the fates and states of cells, many neurons interact to represent our perceptions and memories, and so on. Even if each element in a network achieves only two values, the number of possible states in a network of N elements is 2^N, which easily becomes larger than any realistic experiment (or lifetime!) can sample, the same dimensionality problem that we encountered in movement behavior. Indeed, a lookup table for the probability of finding a network in any one state has $\approx 2^N$ parameters, and this is a disaster. To make progress we search for a simpler class of models with many fewer parameters.

We seek an analysis of living networks that leverages increasingly high-throughput experimental methods, such as the recording from large numbers of neurons simultaneously. These experiments provide, for example, reliable information about the correlations between the action potentials generated by pairs of neurons. In a similar spirit, we can measure the correlations between amino acid substitutions at different sites across large families of proteins. Can we use these pairwise correlations to say anything about the network as a whole? Although there are an infinite number of models that can generate a given pattern of pairwise correlations, there is a unique model that reproduces the measured correlations and adds no additional structure. This minimally structured model is the one that maximizes the entropy of the system (27), in the same way that the thermal equilibrium (Boltzmann) distribution maximizes the entropy of a physical system given that we know its average energy.

Letters in Words. To see how the maximum entropy idea works, we examine an example in which we have some intuition for the states of the network. Consider the spelling of four-letter English words (28), whereby at positions $i = 1, 2, 3, 4$ in the word we can chose a variable x_i from 26 possible values. A word is then represented by the combination $x \equiv \{x_1, x_2, x_3, x_4\}$, and we can sample the distribution of words, $P(x)$, by looking through a large corpus of writings,

for example the collected novels of Jane Austen [the Austen word corpus was created via Project Gutenberg (www.gutenberg.org), combining text from *Emma*, *Lady Susan*, *Love and Friendship*, *Mansfield Park*, *Northhanger Abbey*, *Persuassion*, *Pride and Prejudice*, and *Sense and Sensibility*]. If we do not know anything about the distribution of states in this network, we can maximize the entropy of the distribution $P(\mathrm{x})$ by having all possible combinations of letters be equally likely, and then the entropy is $S_0 = -\sum P_0 \log_2 P_0 = 4 \times \log_2(26) = 18.8$bits. However, in actual English words, not all letters occur equally often, and this bias in the use of letters is different at different positions in the word. If we take these "one letter" statistics into account, the maximum entropy distribution is the independent model,

$$P^{(1)}(\mathbf{x}) = P_1(x_1)\, P_2(x_2)\, P_3(x_3)\, P_4(x_4), \qquad [4]$$

where $P_i(x)$ is the easily measured probability of finding letter x in position i. Taking account of actual letter frequencies lowers the entropy to $S_1 = 14.083 \pm 0.001$ bits, where the small error bar is derived from sampling across the $\sim 10^6$ word corpus.

The independent letter model defined by $P^{(1)}$ is clearly wrong: the most likely words are "thae," "thee," and "teae." Can we build a better approximation to the distribution of words by including correlations between pairs of letters? The difficulty is that now there is no simple formula like Eq. **4** that connects the maximum entropy distribution for $\mathbf{x}$ to the measured distributions of letter pairs (x_i, x_j). Instead, we know analytically the form of the distribution,

$$P^{(2)}(\mathbf{x}) = \frac{1}{Z} \exp\left[-\sum_{i>j} V_{ij}(x_i, x_j) \right], \qquad [5]$$

where all of the coefficients $V_{ij}(x, x')$ have to be chosen to reproduce the observed correlations between pairs of letters. This is complicated but much less complicated than it could be—by matching all of the pairwise correlations we are fixing $\sim 6 \times (26)^2$ parameters, which is vastly smaller than the $(26)^4$ possible combinations of letters.

The model in Eq. **5** has exactly the form of the Boltzmann distribution for a physical system in thermal equilibrium, whereby the letters "interact" through a potential energy $V_{ij}(x, x')$. The essential simplification is that there are no explicit interactions among triplets or quadruplets—all of the higher-order correlations must be consequences of the pairwise interactions. We know that in many physical systems this is a good approximation, that is $P \approx P^{(2)}$. However, the rules of spelling (e.g., i before e except after c) seem to be in explicit conflict with such a simplification. Nonetheless, when we apply the model in Eq. **5** to English words, we find reasonable phonetic constructions. Here we leave aside the problem of how one finds the potentials V_{ij} from the measured correlations among pairs of letters (see refs. 29–35) and discuss the results.

Once we construct a maximum entropy model of words using Eq. **5**, we find that the entropy of the pairwise model is $S_2 = 7.471 \pm 0.006$ bits, approximately half the entropy of independent letters S_1. A rough way to think about this result is that if letters were chosen independently, there would be $2^{S_1} \sim 17,350$ possible four-letter words. Taking account of the pairwise correlations reduces this vocabulary by a factor of $2^{S_1 - S_2} \sim 100$, down to effectively ≈ 178 words. In fact, the Jane Austen corpus is large enough that we can estimate the true entropy of the distribution of four-letter words, and this is $S_{\mathrm{full}} = 6.92 \pm 0.003$ bits. Thus, the pairwise model captures $\sim 92\%$ of the entropy reduction relative to choosing letters independently and hence accounts for almost all of the restriction in vocabulary provided by the spelling rules and the varying frequencies of word use. The same result is obtained with other corpora, so this is not a peculiarity of an author's style.

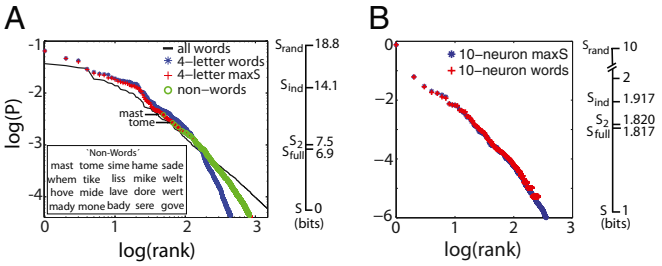

Fig. 4. For networks of neurons and letters, the pairwise maximum entropy model provides an excellent approximation to the probability of network states. In each case, we show the Zipf plot for real data (blue) compared with the pairwise maximum entropy approximation (red). Scale bars to the right of each plot indicate the entropy captured by the pairwise model. (*A*) Letters within four-letter English words (28). The maximum entropy model also produces "nonwords" (*Inset*, green circles) that never appeared in the full corpus but nonetheless contain realistic phonetic structure. (*B*) Ten neuron patterns of spiking and silence in the vertebrate retina (36).

We can look more closely at the predictions of the maximum entropy model in a "Zipf plot," ranking the words by their probability of occurrence and plotting probability vs. rank, as in Fig. 4. The predicted Zipf plot almost perfectly overlays what we obtain by sampling the corpus, although some weight is predicted to occur in words that do not appear in Austen's writing. Many of these are real words that she happened not to use, and others are perfectly pronounceable English even if they are not actually words. Thus, despite the complexity of spelling rules, the pairwise model captures a very large fraction of the structure in the network of letters.

Spiking and Silence in Neural Networks. Maximum entropy models also provide a good approximation to the patterns of spiking in the neural network of the retina. In a network of neurons where the variable x_i marks the presence ($x_i = +1$) or absence ($x_i = -1$) of an action potential from neuron i in a small window of time, the state of the whole network is given by the pattern of spiking and silence across the entire population of neurons, $\mathrm{x} \equiv \{x_1, x_2, \ldots, x_N\}$. In the original example of these ideas, Schneidman et al. (36) looked at groups of $n = 10$ nearby neurons in the vertebrate retina as it responded to naturalistic stimuli, with the

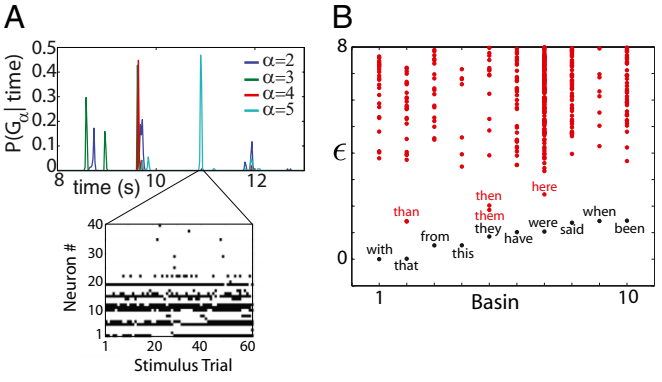

Fig. 5. Metastable states in the energy landscape of networks of neurons and letters. (*A*) Probability that the 40-neuron system is found within the basin of attraction of each nontrivial locally stable state G_α as a function of time during the 145 repetitions of the stimulus movie. *Inset:* State of the entire network at the moment it enters the basin of G_5, on 60 successive trials. (*B*) Energy landscape ($\varepsilon = -\ln P$) in the maximum entropy model of letters in words. We order the basins in the landscape by decreasing probability of their ground states and show the low energy excitations in each basin.

results shown in Fig. 4. Again we see that the pairwise model does an excellent job, capturing ≈90% or more of the reduction in entropy, reproducing the Zipf plot and even predicting the wildly varying probabilities of the particular patterns of spiking and silence (see Fig. 2a in ref. 36).

The maximum entropy models discussed here are important because they often capture a large fraction of the interactions present in natural networks while simultaneously avoiding a combinatorial explosion in the number of parameters. This is true even in cases in which interactions are strong enough so that independent (i.e., zero neuron–neuron correlation) models fail dramatically. Such an approach has also recently been used to show how network functions such as stimulus decorrelation and error correction reflect a tradeoff between efficient consumption of finite neural bandwidth and the use of redundancy to mitigate noise (37).

As we look at larger networks, we can no longer compute the full distribution, and thus we cannot directly compare the full entropy with its pairwise approximation. We can, however, check many other predictions, and the maximum entropy model works well, at least to $n = 40$ (30, 38). Related ideas have also been applied to a variety of neural networks with similar findings (39–42) (however, also see ref. 43 for differences), which suggest that the networks in the retina are typical of a larger class of natural ensembles.

Metastable States. As we have emphasized in discussing Eq. **5**, maximum entropy models are exactly equivalent to Boltzmann distributions and thus define an effective "energy" for each possible configuration of the network. States of high probability correspond to low energy, and we can think of an "energy landscape" over the space of possible states, in the spirit of the Hopfield model for neural networks (44). Once we construct this landscape, it is clear that some states are special because they sit at the bottom of a valley—at local minima of the energy. For networks of neurons, these special states are such that flipping any single bit in the pattern of spiking and silence across the population generates a state with lower probability. For words, a local minimum of the energy means that changing any one letter produces a word of lower probability.

The picture of an energy landscape on the states of a network may seem abstract, but the local minima can (sometimes surprisingly) have functional meaning, as shown in Fig. 5. In the case of the retina, a maximum entropy model was constructed to describe the states of spiking and silence in a population of $n = 40$ neurons as they respond to naturalistic inputs, and this model predicts the existence of several nontrivial local minima (30, 38). Importantly, this analysis does not make any reference to the visual stimulus. However, if we play the same stimulus movie many times, we see that the system returns to the same valleys or basins surrounding these special states, even though the precise pattern of spiking and silence is not reproduced from trial to trial (Fig. 5*A*). This suggests that the response of the population can be summarized by which valley the system is in, with the detailed spiking pattern being akin to variations in spelling. To reinforce this analogy, we can look at the local minima of the energy landscape for four-letter words.

In the maximum entropy model for letters, we find 136 of local minima, of which the 10 most likely are shown in Fig. 5*B*. More than two thirds of the entropy in the full distribution of words is contained in the distribution over these valleys, and in most of these valleys there is a large gap between the bottom of the basin (the most likely word) and the next most likely word. Thus, the entropy of the letter distribution is dominated by states that are not connected to each other by single letter substitutions, perhaps reflecting a pressure within language to communicate without confusion.

Discussion

Understanding a complex system necessarily involves some sort of simplification. We have emphasized that, with the right data, there are mathematical methods that allow a system to "tell us" what sort of simplification is likely to be useful.

Dimensionality reduction is perhaps the most obvious method of simplification—a direct reduction in the number of variables that we need to describe the system. The examples of *C. elegans* crawling and smooth pursuit eye movements are compelling because the reduction is so complete, with just three or four coordinates capturing ≈95% of all of the variance in behavior. In each case, the low dimensionality of our description provides functional insight, whether into origins of stereotypy or the possibility of optimal performance. The idea of dimensionality reduction in fact has a long history in neuroscience, because receptive fields and feature selectivity are naturally formalized by saying that neurons are sensitive only to a limited number of dimensions in stimulus space (45–48). More recently it has been emphasized that quantitative models of protein/DNA interactions are equivalent to the hypothesis that proteins are sensitive only to limited number of dimensions in sequence space (49, 50).

The maximum entropy approach achieves a similar simplification for networks; it searches for simplification not in the number of variables but in the number of possible interactions among these variables. The example of letters in words shows how this simplification retains the power to describe seemingly combinatorial patterns. For both neurons and letters, the mapping of the maximum entropy model onto an energy landscape points to special states of the system that seem to have functional significance. There is an independent stream of work that emphasizes the sufficiency of pairwise correlations among amino acid substitutions in defining functional families of proteins (51–53), and this is equivalent to the maximum entropy approach (53); explicit construction of the maximum entropy models for antibody diversity again points to the functional importance of the metastable states (54).

Although we have phrased the ideas of this article essentially as methods of data analysis, the repeated successes of mathematically equivalent models (dimensionality reduction in movement and maximum entropy in networks) encourages us to seek unifying theoretical principles that give rise to behavioral simplicity. Finding such a theory, however, will only be possible if we observe behavior in sufficiently unconstrained contexts so that simplicity is something we discover rather than impose.

ACKNOWLEDGMENTS. We thank D. W. Pfaff and his colleagues for organizing the Sackler Colloquium and for providing us the opportunity to bring together several strands of thought; and our many collaborators who have worked with us on these ideas and made it so much fun: M. J. Berry II, C. G. Callan, B. Johnson-Kerner, S. G. Lisberger, T. Mora, S. E. Palmer, R. Ranganathan, W. S. Ryu, E. Schneidman, R. Segev, S. Still, G. Tkačik, and A. Walczak. This work was supported in part by grants from the National Science Foundation, the National Institutes of Health, and the Swartz Foundation.

1. Nelson WL (1983) Physical principles for economies of skilled movements. *Biol Cybern* 46:135–147.
2. d'Avella A, Bizzi E (1998) Low dimensionality of supraspinally induced force fields. *Proc Natl Acad Sci USA* 95:7711–7714.
3. Santello M, Flanders M, Soechting JF (1998) Postural hand synergies for tool use. *J Neurosci* 18:10105–10115.
4. Sanger TD (2000) Human arm movements described by a low-dimensional superposition of principal components. *J Neurosci* 20:1066–1072.
5. Ingram JN, Körding KP, Howard IS, Wolpert DM (2008) The statistics of natural hand movements. *Exp Brain Res* 188:223–236.
6. Rashbass C (1961) The relationship between saccadic and smooth tracking eye movements. *J Physiol* 159:326–338.
7. Osborne LC, Lisberger SG, Bialek W (2005) A sensory source for motor variation. *Nature* 437:412–416.
8. Osborne LC, Hohl SS, Bialek W, Lisberger SG (2007) Time course of precision in smooth-pursuit eye movements of monkeys. *J Neurosci* 27:2987–2998.

9. Harris CM, Wolpert DM (1998) Signal-dependent noise determines motor planning. *Nature* 394:780–784.
10. Todorov E, Jordan MI (2002) Optimal feedback control as a theory of motor coordination. *Nat Neurosci* 5:1226–1235.
11. Todorov E (2004) Optimality principles in sensorimotor control. *Nat Neurosci* 7: 907–915.
12. Bialek W (2002) *Physics of Biomolecules and Cells: Les Houches Session LXXV*, eds Flyvbjerg H, Julicher F, Ormos P, David F (EDP Sciences, Les Ulis and Springer-Verlag, Berlin), pp 485–577.
13. Bialek W (1987) Physical limits to sensation and perception. *Annu Rev Biophys Biophys Chem* 16:455–478.
14. Barlow HB (1981) The Ferrier Lecture, 1980. Critical limiting factors in the design of the eye and visual cortex. *Proc R Soc Lond B Biol Sci* 212:1–34.
15. Stephens GJ, Johnson–Kerner B, Bialek W, Ryu WS (2008) Dimensionality and dynamics in the behavior of *C. elegans*. *PLoS Comp Biol* 4:e1000028.
16. Croll N (1975) Components and patterns in the behavior of the nematode *Caenorhabditis elegans*. *J Zool* 176:159–176.
17. Stephens GJ, Bialek W, Ryu WS (2010) From modes to movement in the behavior of *C. elegans*. *PLoS One* 5:e13914.
18. Pierce-Shimomura JT, Morse TM, Lockery SR (1999) The fundamental role of pirouettes in *Caenorhabditis elegans* chemotaxis. *J Neurosci* 19:9557–9569.
19. Gray JM, Hill JJ, Bargmann CI (2005) A circuit for navigation in *Caenorhabditis elegans*. *Proc Natl Acad Sci USA* 102:3184–3191.
20. Stephens GJ, Ryu WS, Bialek W (2009) The emergence of stereotyped behaviors in *C. elegans*, arXiv:0912.5232 [q-bio].
21. Hänggi P, Talkner P, Borkovec M (1990) Reaction-rate theory: Fifty years after Kramers. *Rev Mod Phys* 62:251–341.
22. Dykman MI, Mori E, Ross J, Hunt P (1994) Large fluctuations and optimal paths in chemical kinetics. *J Chem Phys* 100:5735–5750.
23. Bullock TH, Orkand R, Grinnell A (1977) *Introduction to Nervous Systems* (WH Freeman, San Francisco).
24. Korn H, Faber DS (2005) The Mauthner cell half a century later: A neurobiological model for decision-making? *Neuron* 47:13–28.
25. Hodgkin AL, Huxley AF (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* 117:500–544.
26. Dayan P, Abbott LF (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, Cambridge, MA).
27. Jaynes ET (1957) Information theory and statistical mechanics. *Phys Rev* 106:62–79.
28. Stephens GJ, Bialek W (2010) Statistical mechanics of letters in words. *Phys Rev E* 81: 066119.
29. Darroch JN, Ratcliff D (1972) Generalized iterative scaling for log-linear models. *Ann Math Stat* 43:1470–1480.
30. Tkačik G, Schneidman E, Berry MJ, II, Bialek W (2006) Ising models for networks of real neurons, arXiv:q-bio/0611072.
31. Broderick T, Dudik M, Tkačik G, Schapire RE, Bialek W (2007) *Faster solutions of the inverse pairwise Ising problem*, arXiv:0712:2437[q-bio].
32. Ganmor E, Segen R, Schneidman E (2009) How fast can we learn maximum entropy populations? *J Phys Conf Ser* 197:012020.
33. Sessak V, Monasson R (2009) Small–correlation expansions for the inverse Ising problem. *J Phys A: Math Theor* 42:055001.
34. Mézard M, Mora T (2009) Constraint satisfaction problems and neural networks: A statistical physics perspective. *J Physiol Paris* 103:107–113.
35. Roudi Y, Aurell E, Hertz J (2009) Statistical physics of pairwise probability models. *Front Comp Neurosci* 3:22.
36. Schneidman E, Berry MJ, 2nd, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440:1007–1012.
37. Tkačik G, Prentice JS, Balasubramanian V, Schneidman E (2010) Optimal population coding by noisy spiking neurons. *Proc Natl Acad Sci USA* 107:14419–14424.
38. Tkačik G, Schneidman E, Berry MJ, II, Bialek W (2009) Spin glass models for networks of real neurons, arXiv:0912:5409 [q-bio].
39. Shlens J, et al. (2006) The structure of multi-neuron firing patterns in primate retina. *J Neurosci* 26:8254–8266.
40. Tang A, et al. (2008) A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J Neurosci* 28:505–518.
41. Yu S, Huang D, Singer W, Nikolić D (2008) A small world of neuronal synchrony. *Cereb Cortex* 18:2891–2901.
42. Shlens J, et al. (2009) The structure of large-scale synchronized firing in primate retina. *J Neurosci* 29:5022–5031.
43. Ohiorhenuan IE, et al. (2010) Sparse coding and high-order correlations in fine-scale cortical networks. *Nature* 466:617–621.
44. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 79:2554–2558.
45. Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W (1997) *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA).
46. Bialek W, de Ruyter van Steveninck R (2005) Features and dimensions: Motion estimation in fly vision, arXiv:q-bio/0505003.
47. Sharpee T, Rust NC, Bialek W (2004) Analyzing neural responses to natural signals: Maximally informative dimensions. *Neural Comp* 16:223–250.
48. Schwartz O, Pillow JW, Rust NC, Simoncelli EP (2006) Spike-triggered neural characterization. *J Vis* 6:484–507.
49. Kinney JB, Tkačik G, Callan CG, Jr. (2007) Precise physical models of protein-DNA interaction from high-throughput data. *Proc Natl Acad Sci USA* 104:501–506.
50. Kinney JB, Murugan A, Callan CG, Jr., Cox EC (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci USA* 107:9158–9163.
51. Socolich M, et al. (2005) Evolutionary information for specifying a protein fold. *Nature* 437:512–518.
52. Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R (2005) Natural-like function in artificial WW domains. *Nature* 437:579–583.
53. Bialek W, Ranganathan R (2007) Rediscovering the power of pairwise interactions, arXiv:0712.4397[q-bio].
54. Mora T, Walczak AM, Bialek W, Callan CG, Jr. (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* 107:5405–5410.