DERIVATION OF NATURAL STIMULUS FEATURE SET USING A DATA-DRIVEN MODEL

Alexander G. Dimitrov, Tomas Gedeon, Brendan Mumey, Ross Snider, Zane Aldworth, Albert E. Parker and John P. Miller

Center for Computational Biology
Montana State University, Bozeman MT 59717

Abstract: A formal approach for deciphering the information contained within nerve cell ensemble activity patterns is presented. Approximations of each nerve cell's coding scheme is derived by quantizing its neural responses into a small reproduction set, and minimizing an information-based distortion function. During an experiment, the sensory stimulus world presented to the animal is modified to contain a richer set of relevant features, as those features are discovered. A dictionary of equivalence classes is derived, in which classes of stimulus features correspond to classes of spike-pattern code words. We have tested the approach on a simple insect sensory system.

Keywords: Biomedical systems, Coding schemes, Computational methods, Neural activity, Quantization

1. THE NEURAL ENCODING PROBLEM

Two major goals facing neuroscientists are to understand how information is encoded in the activity patterns of neural ensembles and to understand how those activity patterns are decoded by cells at the subsequent processing stages. A formal, general approach toward achieving those goals has been presented in previous work (Dimitrov *et al.,* in press). A significant extension of the technique is presented here, along with a demonstration of the application of that approach to the analysis of neural coding in an insect sensory system. This approach is being refined through the development of a data-driven model of that sensory system.

1.1 A Dictionary for the neural code.

Tools from information theory were used recently to characterize the neural coding scheme of a simple sensory system (Dimitrov and Miller, 2001). That work demonstrated that a coding scheme can be conceptualised as an almost-deterministic relation between clusters of stimulus-response classes, where each class consists of a set of stimuli and a synonymous set of the neural responses elicited by those stimuli. Each "entry" in the dictionary consists of one of these stimulus-responses classes: *i.e.*, all of the stimuli in the stimulus class are treated as being equivalent (*e.g.*, repeated but slightly variant strums of a guitar chord) and all neural responses in the class are also considered to be equivalent (*e.g.*, repeated but slightly variant hand-written notations for that chord). In the context of the sensory system used to illustrate that approach, the stimuli are short-duration (*circa* 20 msec.) segments of sensory input waveforms, and the neural responses are short-duration patterns of action potentials (1-3 APs within 20 ms windows following the stimuli).

A method was developed to find high quality approximations of such a coding scheme (Dimitrov, *et al.*, 2002). The technique involved the quantization of the neural responses to a small reproduction set, and used a minimization of an information-based distortion function to optimize the quantization. In cases involving complex, high-dimensional input stimuli, a model was derived for the stimulus-response relation. Several classes of models were used to provide upper bounds to the information

distortion function used in the optimization problem (Dimitrov, *et al.*, in press). In general, a smaller value of the cost function indicated a better model. All models were variants of Gaussian Mixture Models (GMM) (Bishop, 1998). The differences pertained to the number of parameters used: richer models provided a better bound to the cost function, but needed more data for robust estimates. The applicability of the models were demonstrated by investigating coding properties of several identified neurons in the cricket cercal sensory system (Dimitrov, *et al.*, in press; Dimitrov, *et al.*, 2002).

1.2 What stimulus features are encoded in neural activity patterns?

A typical initiation point for system-identification studies of this nature involve the presentation of Gaussian white noise stimuli. Many earlier studies have used GWN stimuli for the characterization of neural coding characteristics. However, recent results in the cricket cercal sensory system indicate that sensory interneurons show sensitivity to higher-order statistical features that occur very infrequently in GWN stimuli (Roddey *et al.,* 2000). Any characterization of the encoding scheme of a neural system that does not encompass the relevant stimulus regime would be essentially meaningless.

Here, an approach is presented that enables discovery of the set of stimuli to which a cell or ensemble of cells is "tuned", based on the assumption that those cells have become optimised over time to encode and transmit information about that natural stimulus set. This approach also enables a more consistent characterization of the stimulus/response properties of neurons to their natural, behaviourally-relevant stimulus regime. The techniques are introduced and demonstrated within the context of a simple test system: the cricket cercal sensory system.

2. NEUROPHYSIOLOGICAL TESTBED PREPARATION

The preparation used for these studies was the cercal sensory system of the cricket. This system mediates the detection and analysis of low velocity air currents in the cricket's immediate environment. This sensory system is capable of detecting the direction and dynamic properties of air currents with great accuracy and precision (Gnatzy and Heusslein, 1986; Heinzel and Dambach, 1987; Kamper and Kleindienst, 1990; Miller *et al.*, 1991; Shimozawa and Kanou, 1984a,b; Stout *et al.*, 1983; Theunissen and Miller, 1991; Theunissen *et al.,* 1996) and can be thought of as a near-field, low-frequency extension of the animal's auditory system.

Receptor organs. The receptor organs for this modality are two antenna-like appendages called cerci at the rear of the abdomen. Each cercus is covered with approximately 1000 filiform mechanosensory hairs, like bristles on a bottle brush. Each hair is constrained to move along a single axis in the horizontal plane. As a result of this mechanical constraint, an air current of sufficient strength will deflect each hair from its rest position by an amount that is proportional to the cosine of the angle between the air current direction and the hairs movement axis. The 1000 hairs on each cercus are arrayed with their movement axes in diverse orientations within the horizontal plane, insuring that the relative movements of the ensemble of hairs will depend on the direction of the air current. The filiform hairs also display differential sensitivity to aspects of the dynamics of air displacements, including the frequency, velocity, and acceleration of air currents [Osborne, 1997; Roddey and Jacobs, 1996).

Sensory receptor neurons. Each hair is innervated by a single spike-generating mechanosensory receptor neuron. These receptors display directional and dynamical sensitivities that are derived directly from the mechanical properties of the hairs (Kamper and Kleindienst, 1990; Landolfa and Jacobs, 1995; Landolfa and Miller, 1995; Roddey and Jacobs, 1996; Shimozawa and Kanou, 1984a,b). The set of approximately 2000 receptors innervating these filiform hairs have frequency sensitivities spanning the range from about 5 Hz up to about 1000 Hz.

Primary sensory interneurons. The sensory afferents synapse with a group of approximately thirty local interneurons and approximately twenty identified projecting interneurons that send their axons to motor centers in the thorax and integrative centers in the brain. It is a subset of these projecting interneurons that we study here. Like the afferents, these interneurons are also sensitive to the direction and dynamics of air current stimuli (Miller *et al.*, 1991; Theunissen and Miller, 1991; Theunissen *et al.,* 1996). Stimulus-evoked neural responses have been measured in several projecting and local interneurons, using several different classes of air current stimuli. Each of the interneurons studied so far has a unique set of directional and dynamic response characteristics. Previous studies have shown that these projecting interneurons encode a significant quantity of information about the direction and velocity of low frequency air current stimuli with a linear rate code (Clague *et al.*, 1997; Theunissen and Miller, 1991; Theunissen *et al.,* 1996). More recent studies demonstrate that there is also substantial amount of information in the spike trains that cannot be accounted for by a simple linear encoding scheme (Roddey *et al.*, 2000). Evidence suggests the implementation of an ensemble temporal encoding scheme in this system.

3. METHODS

3.1 Experimental approach.

Stimulus-response properties of sensory interneurons were measured using intracellular and extracellular electrodes. Stimuli consisted of controlled air currents

directed across the animals' bodies, and the responses consisted of the corresponding spike trains elicited by those air currents. The cricket preparations were mounted within a miniature wind tunnel, which generated laminar air currents having precisely controlled direction and velocity parameters. Details of the dissection, stimulus generation, and electrophysiological recording procedures are presented in Dimitrov *et al.* (2001).

3.2 Derivation of the stimulus-response equivalence sets.

Details of all analytical techniques, as well as discussions of our computational approaches, are presented in Dimitrov *et al.* (in press.) A brief summary is as follows.

A model of neural processing. The *input signal X* to a neuron (or neural ensemble) may be a sensory stimulus or may be the activity of another set of (pre-synaptic) neurons. We considered the input signal to be produced by a source with a probability $p(x)$. The *output signal Y* generated by that neuron (or neural ensemble) in response to *Y* will be a spike train (or ensemble of spike trains.) We consider the encoding of *X* into *Y* to be a map from one stochastic signal to the other. This stochastic map is the *encoder* $q(y|x)$, which will model the operations of this neuronal layer. The *output signal Y* is induced by $q(y|x)$ by $p(y) = \Sigma_x\, q(y|x)p(x)$.

This view of the neural code, which is probabilistic on a fine scale but deterministic on a large scale, emerges naturally in the context of Information Theory. The Noisy Channel Coding Theorem suggests that, in this context, relations between individual elements of the stimulus and response spaces are not the basic building elements of the system. Rather, the defining objects are relations between *classes* of stimulus-response pairs. There are about $2^{I(X;Y)}$ such equivalence classes (*i.e.,* codeword classes). When restricted to codeword classes, the stimulus-response relation is almost bijective. That is, with probability close to 1, elements of *Y* are assigned to elements of *X* in the same codeword class. This framework naturally deals with lack of bijectivity, by treating it as effective noise. We decode an output *y* as any of the inputs that belong to the same codeword class. Similarly, we consider the neural representation of an input *x* to be any of the outputs in the same codeword class. Stimuli from the same equivalence class are considered indistinguishable from each other, as are responses from within the same class.

Finding the codebook. Given this model of neural function, our task is to recover the codebook. In this context, this equates to identifying the joint stimulus-response classes that define the coding relation. The approach we use is to quantize (*i.e.*, cluster) the response space *Y* to a small reproduction space of finitely many abstract classes, Y_N. This method allows us to study coarse (i.e., small *N*) but highly informative models of a coding scheme, and then to automatically refine them when more data becomes available. This refinement is done by simply increasing the size of the reproduction, *N*.

The mutual information $I(X;Y)$ tells us how many different states on the average can be distinguished in *X* by observing *Y*. If we quantize *Y* to Y_N (a reproduction with N elements), we can estimate $I(X;Y_N)$, which is the mutual information between *X* and the reproduction Y_N. Our information-preserving criterion will then require that we choose a quantizer that preserves as much of the mutual information as possible, *i.e.,* to choose the quantizer $q(Y_N|Y)$ which minimizes the difference

$$D_I(Y;Y_N) = I(X;Y) - I(X;Y_N) \qquad (1)$$

Following examples from rate distortion theory, this problem of optimal quantization can then be formulated as a maximum entropy problem. The reason is that, among all quantizers that satisfy a given set of constraints, the maximum entropy quantizer does not implicitly introduce additional constraints in the problem. Within this framework, the minimum distortion problem is posed as a maximum quantization entropy problem with an appropriate information-theoretic distortion constraint. Complete details are presented in Dimitrov *et al.* (in press). The optimal quantizer $q(y_N|y)$ induces a coding scheme from which is the most informative approximation of the original relation $p(x|y)$ for a fixed size *N* of the reproduction Y_N. Increasing *N* produces a refinement of the approximation, which is more informative (has lower distortion and thus preserves more of the original mutual information I(X;Y)). The elements of Y_N can be interpreted as the labels of the equivalence classes which we want to find. The quantizer $q(y_N|y)$ gives the probability of a response *y* belonging to an equivalence class y_N. Through this approach, we recover an almost complete reproduction of the coding scheme as a relation between stimulus-response equivalence classes. For each neuron, the characteristic stimulus features are represented as the mean voltage waveforms of the stimulus that drove the air currents immediately preceding the elicited spike pattern code words, and the response code words are represented as the actual spike patterns that corresponded to those stimulus features.

3.3 Refinement of the stimulus set.

In previous studies, the encoding properties of nerve cells were studied using single-axis sinusoidal and band-passed gaussian noise stimuli. Here we characterized the INs responses to more complex stimuli, constructed to be more representative of the system's natural stimulus set. During each experiment, the sensory stimulus ensemble presented to the animal was modified to contain a richer set of relevant features, as those features were discovered through on-line analysis of the accumulating stimulus/response data.

During the course of the physiological recording, the system was stimulated initially with air currents drawn from a band-limited (5-500Hz) Gaussian white noise (GWN) source (Theunissen *et al.,* 1996). This broad, non-specific stimulus allowed us to explore a large portion of the input space, and provided sufficient stimulation for a coarse model (*i.e.*, a low-dimensional quantization) of the system. This coarse analysis yielded the preliminary set of response-conditioned mean stimuli leading up to the most frequently observed neural spike-pattern code words (*e.g.*, single spikes, short-interval spike doublets and triplets.)

After the initial model was in place, we modified the stimulus set in two ways. First, we added samples of the response-conditioned stimulus classes derived from the coarse analysis into the GWN stimulus waveforms, to increase the frequency of occurrence of the neural responses for that stimulus class. This allowed us to sample this part of stimulus space more finely, and refine the stimulus-response model, thus lowering the distortion and increasing the informativeness of the reproduction.

Second, we modified the variability of the stimuli along certain stimulus dimensions. Since most of the GMMs project parts of the stimulus to a smaller space, reducing the variability in the discarded subspace does not significantly affect the model. Modifying the variability in the retained subspace does, however, modify the properties of the model. In this case, we modified it in a way that provided a tighter bound in the subsequent analysis, which indicated that we had built a better model. We also modified the variability in directions orthogonal to the ones already presented. The intent of this step was analogous to the initial GWN stimulation: to present parts of stimulus space which the sensory system had not yet perceived.

This refinement procedure was repeated, and the stimulus ensemble was thus iteratively refined from non-specific stimuli to the specific subset of stimuli which elicited a selection of the quantized response classes, which in turn enabled further refinement of the quantizer itself. The method produced a `more` refined dictionary of equivalence classes. In principle, this approach can be continued until the stimulus space is explored in sufficient detail that further refinements of the stimulus do not produce refinements of the response model.

4. RESULTS

This analytical approach was used to characterize the encoding characteristics of single identified sensory interneurons in the cricket cercal system. The specific goal of the experiments and analyses were to discover (jointly) a) the dynamic stimulus waveform features encoded by the cells, and b) the spike train codeword patterns that encoded those features. In figure 1, the stimulus features are represented as the mean voltage waveforms of the stimuli that drove the air currents immediately preceding the elicited spike pattern code words, and the response code words are represented as the actual spike patterns that corresponded to those stimulus features. We use a representation of spike patterns that is similar to a peristimulus time histogram. We call this representation a class conditioned time histogram (CCTH.)

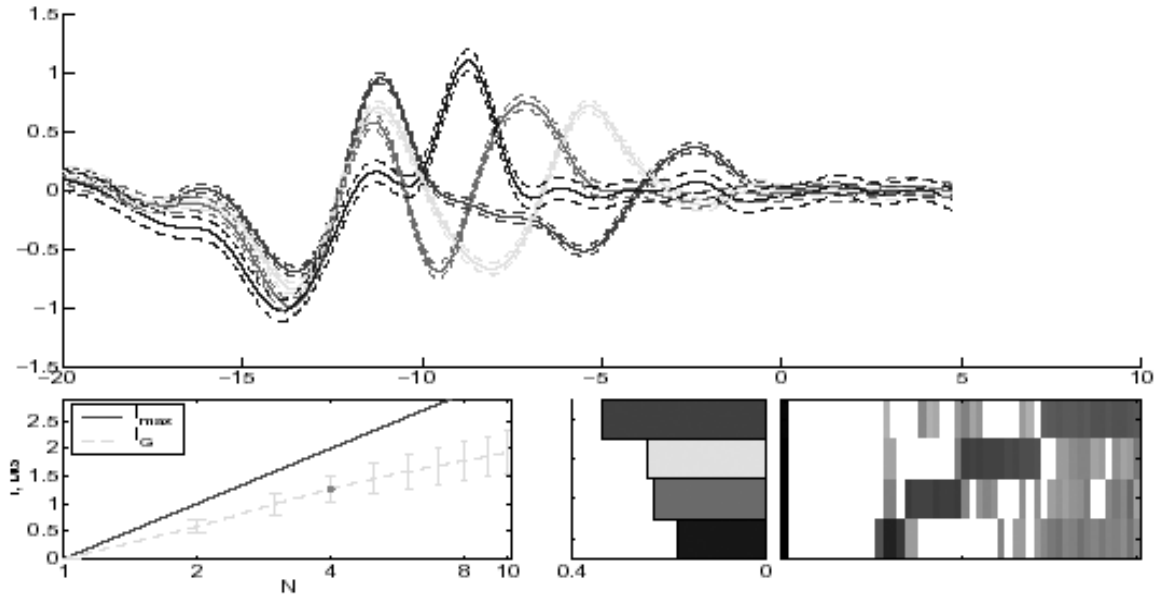

Fig. 1. A quantization with 4 classes. Top panel: the four class-conditioned mean stimulus waveforms. Bottom right panel: the CCTH of a spike at time T given the pattern in a certain class. Lower center panel: relative proportion of spike patterns belonging to the different classes, as GMM priors. Lower left panel: estimate of the lower bound to the mutual information (gray dashed curve) and the absolute upper bound for the same level of quantization (dark solid curve, $\log_2$).

For this cell, the stimulus-response space was quantized into four classes (*i.e.*, N=4). The top panel shows four class-conditioned mean stimulus waveforms, corresponding to the four spike pattern code words derived through this successive quantization and stimulus refinement procedure. These waveforms correspond to the air-current velocity presented to the preparation. The 4 mean waveforms are each plotted in the gray scale of the corresponding class. The horizontal axis of this top plot denotes time, in ms, relative to the occurrence of the first spike in a class. That is, time 0 is the time at which the first spike on the codeword pattern occurred. The dashed lines denote 95% confidence intervals of the means, which depend on the reproduction size, N.

The lower right panel plots the CCTH spike code words for these four classes. These are the classes of spike patterns that served as the basis for extracting the corresponding mean stimulus waveforms. Every class starts with a spike (line at 0ms). We plot the conditional probability of spike occurrence *vs.* time for each pattern on a logarithmic scale, with black indicating a probability of one for the occurrence of a spike, and a lighter shade of gray representing a lower probability. To be precise, we plot the expectation is $\Sigma_y\, y_i p(y_i|y_N)$. The pattern y_i can be considered as the conditional probability $p(t_j|y_i)$ = p(spike occurs at time t_j given that the observed pattern is y_i). This probability is 1 at times when a spike occurs and zero otherwise. In this case, this panel can be interpreted as showing $p(t_j|y_N) = \Sigma_y p(t_j|y_i)p(y_i|y_N)$ - the conditional probability of a spike at time t_i given class y_N. The similarity to a PSTH is that we present the distribution of spikes in time, conditioned on the occurrence of an event. For the PSTH, the event is a particular stimulus. For this representation, the event is a certain response class, hence the name CCTH. These patterns are aligned in time with the mean stimulus waveforms that elicited them, in the panel directly above. That is, the significant portions of the stimulus patterns that correspond to the observed spike patterns occur during the 20 ms preceding the spike code words.

The lower center panel to the left of these CCTH plots show the relative proportion of spike patterns belonging to the different classes, as GMM priors (weights). These bars are gray-scale-coded to indicate the class-conditioned mean stimulus waveform to which the spike pattern to the right corresponds. This particular quantization grouped the spike patterns roughly according to interspike intervals: the top class consisted mostly of doublets with a second spike 7-10 ms after the initial spike (dark gray range to the right), and a few triplets (light gray bars in front), for which the third spike is in the same range. The bottom (black) class consists mostly of short doublets, with a second spike 2.5-3.3 ms after the first spike, and a range of triplets with a third spike 6-10ms after the first spike.

The lower left panel shows the estimate of the lower bound to the mutual information (gray dashed curve) and the absolute upper bound for the same level of quantization (dark solid curve, $\log_2$ N). The error bars mark the uncertainty of the estimate, which depend on the reproduction size. The estimate for the 4-class quantization shown here is denoted with a dark plotted point on the mutual information curve at N=4.

For this cell, application of the iterative stimulus-refinement approach yielded a set of stimulus waveforms that differed significantly from the set obtained when the stimulus regime was limited to Gaussian white noise. Specifically, several of the characteristic stimulus features included multiple cycles of sinusoidal-like oscillations.

5. CONCLUSIONS

This analytical approach offered several significant advantages to our characterization of the neural encoding scheme for this cell than previous approaches. First, the approach enabled a more rapid convergence toward a more accurate and meaningful representation of stimulus–response equivalence classes than did our previous approach. A major reason for this is that the stimulus regime we crafted through the iterative process had a much higher content of waveform segments containing maintained, multiple cycles of sine waves than do Gaussian white noise signals. Such multi-cycle segments are rare occurrences in GWN, and use of GWN to achieve the same level of confidence had required much longer experimental recording sessions. In some other neurons we have studied, preliminary evidence suggests that GWN signals contain such a small fraction of "relevant" stimulus features that a conventional system identification approach would never be practical, given realistic experimental constraints.

The neurobiological results themselves are enlightening, in that the approach demonstrates that non-linear encoding schemes are being used to represent information. In the case shown in Fig. 1, spike multiplets carry a quantity of information about characteristic stimulus features that is greater than the amount that could be extracted by a mean-rate decoder operating on the same stimulus waveforms.

An electronic system to enable execution of this iterative stimulus refinement and quantization analysis in real-time is currently being developed. This test-bed will enable real-time decoding of ensemble neural activity patterns and real-time interactive modulation of those neural activity patterns. The hardware devices supporting these tasks are being developed with advanced Digital Signal Processing and Field Programmable Gate Array technologies.

ACKNOWLEDGEMENTS

Research supported by an NSF EIA-BITS grant (JPM,AGD,RS,TG), and by grants DGE9972824 (ZA,AEP) and MRI9871191, and by NIH grants MH12159 (AGD) and MH57179 (JPM, ZA, AGD).

REFERENCES

Bishop, C.M. (1998). *Neural Networks for Pattern Recognition*. Oxford University Press, New York, New York.

Clague, H., F. Theunissen and J.P. Miller (1997). The effects of adaptation on neural coding by primary sensor interneurons in the cricket cercal system. *J. Neurophysiol.*, **77**: 207–220.

Dimitrov, A.G. and J.P. Miller (2001). Neural coding and decoding: communication channels and quantization. *Network: Computation in Neural Systems* **12**(4): 441-472.

Dimitrov, A.G., J.P. Miller, Z. Aldworth and A. Parker (2002). Spike pattern-based coding schemes in the cricket cercal sensory system. *Neurocomputing* **44-46**: 373-379.

Dimitrov, A.G., J.P. Miller, T. Gedeon, Z. Aldworth and A.E. Parker (in press). Analysis of Neural Coding using Quantization with an information-based distortion measure. *Network: Computation in Neural Systems*.

Gnatzy and Heusslein (1986). Digger wasp against crickets. I. receptors involved in the antipredator strategies of the prey. *Naturwissenschaften*, **73**:212–215.

Heinzel, H.G. and M. Dambach (1987). Traveling air vortex rings as potential communication signals in a cricket. *J. Comp. Physiolo. A.*, 160:**79–88**.

Kamper, G. and H.-U. Kleindienst (1990). Oscillation of cricket sensory hairs in a low frequency sound field. *J. Comp. Physiol. A.*, **167**:193–200.

Landolfa, M.A. and G.A. Jacobs (1995). Direction sensitivity of the filiform hair population of the cricket cercal system. *J. Comp. Physiol. A*, **177**:758–767.

Landolfa, M.A. and J. P. Miller (1995). Stimulus-response properties of cricket cercal filiform hair receptors. *J. Comp. Physiol. A.*, **177**:749–757.

Miller, J.P., G.A. Jacobs and F.E. Theunissen (1991). Representation of sensory information in the cricket cercal sensory system. I. Response properties of the primary interneurons. *J. Neurophysiol.*, **66**:1680–1689.

Osborne, L.C. (1997). *Biomechanical Properties Underlying Sensory Processing in Mechanosensory Hairs in the Cricket Cercal Sensory System.* PhD thesis, University of California, Berkeley, CA.

Roddey, J.C., B.Girish, and J.P. Miller (2000). Assessing the performance of neural encoding models in the presence of noise. *J. Comp. Neurosci.*, **8**:95–112.

Roddey, J.C. and G.A. Jacobs (1996). Information theoretic analysis of dynamical encoding by filiform mechanoreceptors in the cricket cercal system. *J. Neurophysiol.*, **75**:1365–1376.

Shimozawa, T. and M. Kanou (1984a). The aerodynamics and sensory physiology of a range fractionation in the cercal filiform sensilla of the cricket *gryllus bimaculatus*. *J. Comp. Physiol. A.*, **155**:495–505.

Shimozawa, T. and M. Kanou (1984b). Varieties of filiform hairs: range fractionation by sensory afferents and cercal interneurons of a cricket. *J. Comp. Physiol. A.*, **155**:485–493.

Stout, J.F., C.H. DeHaan and R.W. McGhee (1983). Attractiveness of the male acheta domesticus calling song to females. I. Dependence on each of the calling song features. *J. Comp. Physiol.*, **153**:509–521.

Theunissen, F.E. and J.P. Miller (1991). Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning curve width of four primary interneurons. *J. Neurophysiol.*, **66**:1690–1703.

Theunissen, F., J.C. Roddey, S. Stufflebeam, H. Clague, and J.P. Miller (1996). Information theoretic analysis of dynamical encoding by four primary interneurons in the cricket cercal system. *J. Neurophysiol.*, **75**:1345–1364.