Brief Communications

Sparse Representation in the Human Medial Temporal Lobe

Stephen Waydo,[1] Alexander Kraskov,[2] Rodrigo Quian Quiroga,[3] Itzhak Fried,[4,5] and Christof Koch[2]

[1]Control and Dynamical Systems and [2]Computation and Neural Systems, California Institute of Technology, Pasadena, California 91125, [3]Department of Engineering, University of Leicester, Leicester LE1 7RH, United Kingdom, [4]Division of Neurosurgery and Neuropsychiatric Institute, University of California, Los Angeles, Los Angeles, California 90095, and [5]Functional Neurosurgery Unit, Tel-Aviv Medical Center, Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel

Recent experiments characterized individual neurons in the human medial temporal lobe with remarkably selective, invariant, and explicit responses to images of famous individuals or landmark buildings. Here, we used a probabilistic analysis to show that these data are consistent with a sparse code in which neurons respond in a selective manner to a small fraction of stimuli.

Key words: **representation; sparseness; MTL; hippocampus; memory; neural coding**

Introduction

Single-unit recordings from the human medial temporal lobe (MTL) have revealed the existence of highly selective cells that may, for example, respond strongly to different images of a single celebrity but not to 100 pictures of other people or objects (Quian Quiroga et al., 2005). These results suggest a sparse and invariant encoding in MTL and seem to imply the existence of grandmother cells that respond to only a single category, individual, or object (Konorski, 1967; Barlow, 1972; Gross, 2002) [but see criticisms of this view in the study by Quian Quiroga et al. (2005)]. However, because of limitations on the sampling of MTL neurons and on the sampling of the stimulus space, it is unclear how many stimuli a given neuron will respond to on average and, conversely, how many MTL neurons are involved in the representation of a given object. Given the number of stimuli we present and the number of neurons from which we record, we used probabilistic reasoning to explore these issues.

Materials and Methods

Let the sparseness a be the fraction of stimuli a neuron responds to or, alternatively, the probability that a neuron responds to a random stimulus (Treves and Rolls, 1991; Willmore and Tolhurst, 2001; Olshausen and Field, 2004). We assume that a neuron either fires or does not fire to any of the U stimuli making up the universe of stored representations (e.g., Jennifer Aniston, White House, dachshund, iPod). At one extreme (that of a grandmother neuron) $a = 1/U$, whereas at the other extreme a fully distributed representation would have $a = 1/2$, that is, each neuron would respond to half of all represented stimuli. In addition to quantifying how frequently a neuron will respond to a stimulus, this measure has been related to the theoretical storage capacity of autoassociative memory networks (Meunier et al., 1991; Treves and Rolls, 1991).

In the following analysis, we make a few key assumptions. First, we assume the responses of all neurons can be treated in a binary manner, that is, we can define a threshold above which we consider a neuron to have responded (and we examine how our results vary with this threshold). Second, we assume the stimulus presentations are independent, and further that the neuronal responses are independent of one another (aside from any stimulus-induced correlations). The independence assumptions are consistent with our observation of no significant correlations between neurons in the experimental data. Finally, we assume that all of our recorded neurons share the same underlying sparseness a. However, because our results are expressed as a probability density function over this value, the width of the density function can be interpreted as describing the range of sparseness present in the MTL.

Suppose in a single experiment we present S stimuli to a binary neuron and count the number S_r of responses. Let f_a be the probability density function of the sparseness index a; approximately speaking $f_a(\alpha)\delta\alpha$ is the probability that a lies in an interval of size $\delta\alpha$ around some value α. We want to determine $f_a(\alpha \mid S_r = s_r)$, the probability density function for a given the observed data. Our a priori estimate of f_a is simply $f_a(\alpha) = 1$ for $0 \le \alpha \le 1$, that is, a is equally likely to take on any value between 0 and 1. At a particular value of a, the probability that S_r takes on a value s_r (between 0 and S), $P[S_r = s_r \mid a = \alpha]$, follows a binomial distribution, but if a is unknown, all responses are equally likely and so $P[S_r = s_r] = 1/(S + 1)$. Applying Bayes' rule we have the following:

$$f_a(\alpha \mid N_r = n_r \wedge S_r = s_r) = \frac{P[N_r = n_r \wedge S_r = s_r \mid a = \alpha] f_a(\alpha)}{P[N_r = n_r \wedge S_r = s_r]}.$$

The responses of each cell thus yield a curve of the probability density of a given the response pattern of the cell. Figure 1 gives three examples of the curves that could be generated using this method.

A limitation of this approach is that if, for example, two neurons are presented with the same 100 stimuli and neither responds, the true sparseness is likely to be much smaller than that implied by the individual density curves (although the neurons may simply be unresponsive to any stimulus; see below). Because the original data were acquired using 64 microelectrodes, we extend our approach to account for an experiment in which N neurons are recorded simultaneously while S stimuli are presented. We define N_r to be the number of neurons that respond significantly to at least one stimulus and S_r to be the number of stimuli that

Received May 17, 2006; revised Aug. 2, 2006; accepted Aug. 29, 2006.

This work was supported by grants from the National Institute of Neurological Disorders and Stroke, the National Institute of Mental Health, the National Science Foundation, the Defense Advanced Research Projects Agency, the Engineering and Physical Sciences Research Council, the Office of Naval Research, the Gordon Moore Foundation, and the Swartz Foundation for Computational Neuroscience and by a Fannie and John Hertz Foundation fellowship to S.W. We thank Gabriel Kreiman for useful feedback.

Correspondence should be addressed to Stephen Waydo, Control and Dynamical Systems, California Institute of Technology, 1200 East California Boulevard, M/C 107-81, Pasadena, CA 91125. E-mail: waydo@cds.caltech.edu.

DOI:10.1523/JNEUROSCI.2101-06.2006

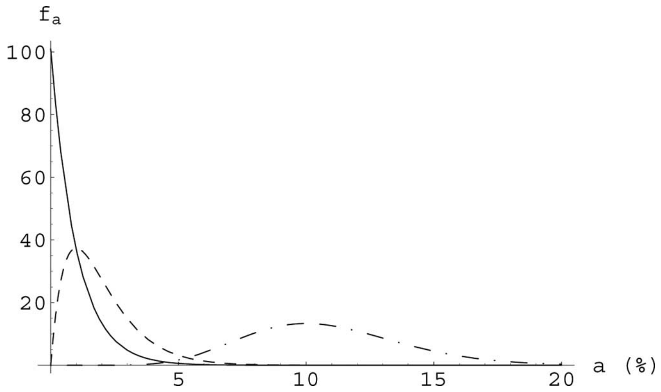

Figure 1. Example probability density functions for sparseness, expressed as a percentage of stimuli that evoke a response, in three scenarios. The dashed curve is that obtained for an idealized grandmother neuron (representing a single object), the preferred stimulus of which was among the 100 images shown, whereas the solid curve would result if the preferred stimulus of the cell was not shown. The dash-dotted curve would result from a cell firing to 10 of the 100 stimuli presented. As the number of stimuli shown to the cell approaches the total number of images stored by the network, the density function will converge to an even narrower curve centered at the true sparseness *a*.

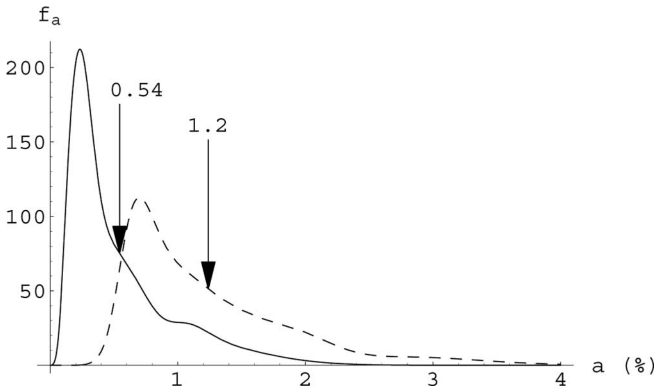

Figure 2. Probability density function for sparseness *a* averaged over 34 experimental sessions that yielded spiking responses from 1425 units. Two different thresholds for defining significant responses are considered: five (solid curve) and three (dashed) SDs above baseline. The means of the distributions, corresponding to the best estimates for *a*, are indicated by arrows, and the values below which *a* is likely to lie with 95% probability are $a = 1.4$ and 2.6%. The peaks of the distributions are at 0.23 and 0.70%. The average number of simultaneously recorded units per session, *N*, is 41.9, and the mean number of images shown to the subjects, *S*, is 88.4.

produce a response in at least one of these. The derivation of the closed-form joint probability distribution of N_r and S_r involves solving a recursive relation for the conditional distribution of S_r given N_r and is described in the supplemental methods (available at www.jneurosci.org as supplemental material). As in the single-neuron example above, we can then apply Bayes' rule to find the probability density function for *a* given the results of a recording session as follows:

$$\begin{aligned} f_a(\alpha \mid S_r = s_r) &= \frac{P[S_r = s_r \mid a = \alpha] f_a(\alpha)}{P[S_r = s_r]} \\ &= \binom{S}{s_r} \alpha^{s_r} (1 - \alpha)^{S - s_r} (S + 1) \,. \end{aligned}$$

Rather than obtaining a single curve for each cell, we now obtain a single curve for each session that takes into account the presence of cells that did not respond to any stimulus or that responded to multiple stimuli.

Results

We use this distribution to determine density functions for *a* from a data set of 1425 MTL units from 34 experimental sessions in 11 patients (Quian Quiroga et al., 2005). To fit the data against our binary model, we considered a response to be significant if it was larger than the mean plus a threshold number of SDs of the baseline (before the onset of the image) and had at least two spikes in the poststimulus time interval considered (0.3–1 s) [as in the previous study by Quian Quiroga et al. (2005)]. Figure 2 depicts the resulting average probability distributions for thresholds of three and five SDs; for lower thresholds, many of the "responses" are a result of random fluctuations in firing rate rather than genuine responses to stimuli. For a threshold of five SDs above baseline, the peaks of the 34 individual distributions lie in the range of 0.16–1.64%, with a mean peak location of 0.51% and a SD of 0.40%. For a threshold of three SDs above baseline, the individual curves peak in the range of 0.52–3.08%, with a mean peak location of 1.21% and a SD of 0.63%. The peaks of the average distributions shown in Figure 2 are at $a = 0.23$ and 0.70% for thresholds of five and three SDs, respectively, whereas the means are at $a = 0.54$ and 1.2%.

From this figure, we conclude that *a* most likely lies in the range of 0.2–1%. Although this is a sparse coding scheme, considering the large number of MTL neurons and the large number of represented stimuli, it still results in a single unit responding to many stimuli and many MTL units responding to each stimulus. This is much sparser than responses obtained from the monkey superior temporal sulcus, where a mean response sparseness of ~33% has been reported (Rolls and Tovee, 1995), as well as responses from the monkey inferotemporal cortex, where sparse population coding has been observed (Young and Yamane, 1992). We assume, however, that all cells we are listening to are involved in the representation of some stimulus, which may not be the case (i.e., some of them could serve a different function altogether) and could cause a downward bias in our estimate. We believe this bias to be small, because repeating the same analysis leaving out half of the unresponsive neurons yields estimates for the mean of *a* of 0.9 and 1.8% at thresholds of five and three SDs, respectively.

We can then estimate the probability of finding such highly selective cells in a given experiment. If the true sparseness is 0.54% (the mean of the distribution with a threshold of 5), in a typical session with $N = 42$ simultaneously recorded units and $S = 88$ test stimuli (the averages from our experiments), we would expect to find on average 15.9 units responding to 17.9 stimuli (with each responsive neuron responding on average to 1.3 images, and each evocative stimulus producing a response in an average of 1.1 neurons). In our experiments, *N* ranged from 18 to 74 and *S* ranged from 57 to 114, and with a five SD threshold, we found on average 7.9 responsive units (range, 3–20) responding to 16.4 stimuli (range, 3–44). As a further check of our methods, we can examine how frequently two or more units responded to the same stimulus. At a five SD threshold, on average, 4.1% of stimuli produced a (simultaneous) response in at least two neurons (range, 0–17.9%; median, 1.6%), compared with a predicted value (at 0.54% sparseness) of 2.2%. Noting that we cannot expect perfect agreement between this prediction and the observed value because of the varying numbers of neurons and stimuli across recording sessions, we see that our model agrees very well with the observed statistics.

Discussion

We developed a method for obtaining a probability distribution for sparseness based on multiple simultaneous neuronal recordings. This distribution allows us to not only examine the average sparseness observed in a given experiment but also the range of sparseness consistent with the data. Averaging these distributions over 34 recording sessions in the human medial temporal lobe, we conclude that highly sparse (although not grandmother) coding is present in this brain region.

To animate this discussion with some numbers, consider 0.54% sparseness level. Assuming on the order of 10^9 neurons in both left and right human medial temporal lobes (Harding et al., 1998; Henze et al., 2000; Schumann et al., 2004), this corresponds to $\sim$5 million neurons being activated by a typical stimulus, whereas a sparseness of 0.23% implies activity in a bit more than 2 million neurons. Furthermore, if we assume that a typical adult recognizes between 10,000 and 30,000 discrete objects (Biederman, 1987), $a = 0.54\%$ implies that each neuron fires in response to 50–150 distinct representations.

This interpretation relies on the assumption that the cells from which we record are part of an object representation system. Instead, it may be possible that these neurons signal the recentness or familiarity rather than the identity of a stimulus. Neurons responding to both novelty and familiarity have been identified in the human hippocampus (Rolls et al., 1982; Fried et al., 1997; Rutishauser et al., 2006; Viskontas et al., 2006). Even if true, however, this view does not invalidate our conclusion that the true sparseness likely lies below 1%. Instead, it would imply that rather than a single neuron responding to dozens of stimuli out of a universe of tens of thousands, such a neuron might respond to only one or a few stimuli out of perhaps hundreds currently being tracked by this memory system, still with millions of neurons being activated by a typical stimulus.

Two significant factors may bias our estimate of sparseness upward. A large majority of neurons within the listening radius of an extracellular electrode are entirely silent during a recording session (e.g., there are as many as 120–140 neurons within the sampling region of a tetrode in the CA1 region of the hippocampus (Henze et al., 2000), but we typically only succeed in identifying 1–5 units per electrode). In rats, as many as two of three cells isolated in the hippocampus under anesthesia may be behaviorally silent (Thompson and Best, 1989), although the reason for their silence is unclear. Thus, the true sparseness could be considerably lower. Furthermore, there is a sampling bias in that we present stimuli familiar to the patient (e.g., celebrities, landmarks, and family members) that may evoke more responses than less familiar stimuli. For these reasons, these results should be interpreted as an upper bound on the true sparseness, and some neurons may provide an even sparser representation.

These results are consistent with Barlow's (1972) claim that "at the upper levels of the hierarchy, a relatively small proportion [of neurons] are active, and each of these says a lot when it is active," and his further speculation that the "aim of information processing in higher sensory centers is to represent the input as completely as possible by activity in as few neurons as possible" (Barlow, 1972).

References

Barlow HB (1972) Single units and sensation: a neuron doctrine for perceptual psychology? Perception 1:371–394.

Biederman I (1987) Recognition-by-components: a theory of human image understanding. Psychol Rev 94:115–147.

Fried I, MacDonald KA, Wilson CL (1997) Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. Neuron 18:753–765.

Gross CG (2002) Genealogy of the grandmother cell. Neuroscientist 8:512–518.

Harding A, Halliday G, Kril J (1998) Variation in hippocampal neuron number with age and brain volume. Cereb Cortex 8:710–718.

Henze DA, Borhegyi Z, Csicsvari J, Mamiya A, Harris KD, Buzsaki G (2000) Intracellular features predicted by extracellular recordings in the hippocampus in vivo. J Neurophysiol 84:390–400.

Konorski J (1967) Integrative activity of the brain; an interdisciplinary approach. Chicago: University of Chicago.

Meunier C, Yanai H, Amari S (1991) Sparsely coded associative memories: capacity and dynamical properties. Network 2:469–487.

Olshausen BA, Field DJ (2004) Sparse coding of sensory inputs. Curr Opin Neurobiol 14:481–487.

Quian Quiroga R, Reddy R, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. Nature 435:1102–1107.

Rolls ET, Tovee MJ (1995) Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J Neurophysiol 73:713–726.

Rolls ET, Perrett DI, Caan AW, Wilson FAW (1982) Neuronal responses related to visual recognition. Brain 105:611–646.

Rutishauser U, Mamelak AN, Schuman EM (2006) Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex. Neuron 49:805–813.

Schumann CM, Hamstra J, Goodlin-Jones BL, Lotspeich LJ, Kwon H, Buonocore MH, Lammers CR, Reiss AL, Amaral DG (2004) The amygdala is enlarged in children but not adolescents with autism: the hippocampus is enlarged at all ages. J Neurosci 24:6392–6401.

Thompson LT, Best PJ (1989) Place cells and silent cells in the hippocampus of freely behaving rats. J Neurosci 9:2382–2390.

Treves A, Rolls ET (1991) What determines the capacity of autoassociative memories in the brain? Network 2:371–397.

Willmore B, Tolhurst DJ (2001) Characterizing the sparseness of neural codes. Netw Comput Neural Syst 12:255–270.

Young MP, Yamane S (1992) Sparse population coding of faces in the inferotemporal cortex. Science 256:1327–1331.

Viskontas IV, Knowlton BJ, Steinmetz PN, Fried I (2006) Differences in mnemonic processing by neurons in the human hippocampus and parahippocampal regions. J Cogn Neurosci, in press.