Cellular Dynamics of Network Memory*

Joaquin M. Fuster

Brain Research Institute and Department of Psychiatry and Biobehavioral Sciences, School of Medicine, University of California at Los Angeles, 760 Westwood Plaza, Los Angeles, CA 90024

Z. Naturforsch. **53c,** 670–676 (1998); received April 20, 1998

Cortex, Memory Networks, Recurrent Model, Active Short-Term Memory, Nonhuman Primate

One example of "emergence" is the development, as a result of neural ontogeny and living experience, of cortical networks capable of representing and retaining cognitive information. A large body of evidence from neuropsychology, electrophysiology and neuroimaging indicates that so-called working memory and long-term memory share the same neural substrate in the cerebral cortex. That substrate consists in a system of widespread, overlapping and hierarchically organized networks of cortical neurons. In this system, any neuron or group of neurons can be part of many networks, and thus many memories. Working memory is the temporary activation of one such network of long-term memory for the purpose of executing an action in the near future. The activation of the network may be brought about by stimuli that by virtue of prior experience are in some manner associated with the cognitive content of the network, including the response of the organism to those stimuli. The mechanisms by which the network stays activated are presumed to include the recurrent re-entry of impulses through associated neuronal assemblies of the network. Consistent with this notion is the following evidence: (1) working memory depends on the functional integrity of cortico-cortical connective loops; and (2) during working memory, remarkable similarities – including "attractor behavior" – have been observed between firing patterns in real cortex and in an artificial recurrent network.

In this paper, I first present briefly some general notions about the cortical topography of long-term memory. Then I focus on a particular aspect of memory dynamics that is essential for sequential behavior, speech and logical reasoning, and that has been the subject of many of our studies in the nonhuman primate. I am referring to the neural manifestations and mechanisms of so-called working memory, a state of memory which I postulate consists basically in the temporary activation of long-term memory at the service of behavior, speech and thinking. In my view, working memory is simply the sustained activation of a network of cortical neurons that, because of its connections between cortical sites, represents all the associated components of long-term memory and procedural knowledge that together constitute a "working memorandum." Behavior, speech and reasoning require the timely and temporary retention of many such memoranda in succession. All of them have wide, though differing, cortical distribution; it now appears that the activated networks straddle distant and extensive portions of posterior and anterior (frontal) cortex.

The first experimental evidence of the distributed character of memory was obtained by Lashley (1950) through cortical lesion studies in rats and monkeys. At the theoretical level, Friedrich Hayek (1952) was the first to propose the existence of cortical networks ("maps") supporting both memory and perception in the human. According to him, memories are stored in wide networks of interconnected cortical neurons. Such networks would be formed by facilitation of synapses through temporal coincidence of inputs, in accord with principles of synaptic change similar to those proposed by Hebb (1949). The reactivation of these networks would be the basis of perception as well as the acquisition of new memory.

* This communication is a contribution to the workshop on "Natural Organisms, Artificial Organisms, and Their Brains" at the Zentrum für interdisziplinäre Forschung (ZiF) in Bielefeld (Germany) on March 8–12, 1998.

Reprint requests to Prof. J. M. Fuster.
Fax: *310-8256792.
E-mail: joaquinf@ucla.edu.

0939–5075/98/0700–0670 $ 06.00

The latter would result from the reactivation and expansion of old memory networks.

A similar concept of a widely distributed substrate for learning and memory was proposed by Edelman and Mountcastle (1978). In their model, the basic unit of representation is the cortical column of cells of equal function, as have been found in primary sensory cortices. Further they postulate that the selectivity of perception and memory is ensured by the re-entry of excitation through recurrent circuits upon the cortical columns that constitute the neural support of a percept or a memory.

The hippocampus and other cortical structures of the medial and inferior temporal lobe probably play a critical role in the acquisition of new memory, that is, in the expansion of old memory networks in the neocortex (Squire and Zola-Morgan, 1988). That memory-making role of temporal-lobe structures is probably mediated by fiber connections between those structures and the neocortex of association (Amaral, 1987). The mechanisms of memory formation are still unknown. They may include certain synaptic processes such as long-term potentiation (LTP) and glutaminergic receptors, such as NMDA receptors (for review, see Fuster, 1995).

Memories are made of innumerable associations of sensorium and of motor action with widely scattered representations in the cerebral cortex. Thus, the networks of long-term memory are distributed over large expanses of cortex. Moreover, since the associated components of any given memory can be also components of other memories, any cortical neuron or group of neurons can be part of many memory networks. Insofar as working memory can be construed as the temporary activation of long-term memory, it is to be expected that the performance of working-memory tasks, whatever the memorandum, will activate large regions of cortex. This expectation has been verified by microelectrode recording in monkeys (Fuster, 1995) and by PET imaging in humans (Swartz *et al.,* 1995).

Those wide cortical networks, however, which are activated in working memory, have certain nodes of preferential or specialized activation depending on the sensory stimuli or motor responses they represent for the short term. Thus, for example, if the task requires the retention of a visual stimulus for the performance of a selective manual response in the short term, then, during the memorization of the visual stimulus, the inferotemporal cortex (visual association cortex) will be most activated, together with frontal areas representing the motor response. The activated network will extend from inferior temporal to prefrontal cortex. It will be maintained active by a complex system of cortico-cortical connections, some local and some distant. One plausible hypothesis is that the sustained activation of the network's neurons is at least in part assured by the recurrent circulation of impulses within it. What follows is an outline of our research to test the validity of this hypothesis.

By cooling one part of cortex while recording from another in the course of memory tasks, we were able to obtain the first evidence of tonic influences from one cortical region upon another. These tonic influences seem to be part of the mechanism for the maintenance of short-term memory in wide cortical networks (Fuster *et al.,* 1985). The monkeys for our experiment were trained to perform a visual memory task: delayed matching to sample with colors. On every trial, the animal had to retain a color in memory for some 10–20 sec (delay) in order to perform a correct color match at the end of the trial. We knew from previous experiments (Fuster and Jervey, 1982; Fuster *et al.* 1982) that the memorization of the color stimulus increases in a sustained manner the activity of large populations of cells in the inferotemporal and prefrontal cortices. Some of the cells, in either cortex, show a different level of activation depending on the color in memory. We also knew that the cooling of either of those cortices, but not parietal cortex, induced a reversible deficit in the performance of the task (Bauer and Fuster, 1976; Fuster *et al.,* 1981).

Thus it appeared, from recording and reversible-lesion data, that the maintenance of a visual memory for the short term was dependent on the activation of a large network of neurons that extended from inferior temporal cortex, through long fiber connections, all the way to the prefrontal cortex. These connections, which are reciprocal (i.e., they go both ways), have been well demonstrated by anatomical studies in the monkey. Consequently, we reasoned that, possibly through those connections, changes in activity in one part of the network should affect activity in another. By cooling inferotemporal cortex we should be able to, per-

haps, depress the memory-elevated discharge of cells in prefrontal cortex, and vice versa.

The results did not turn out to be that simple. Some cells were not at all affected by the distant cortical cooling. Others were excited by the cooling, and still others inhibited. Further, in a small number of inferotemporal and prefrontal cells the effect of distant cooling appeared related to the color that the animal was retaining in short-term memory. The net effect of cooling on these cells was a diminution of their color-dependent differences in firing during the delay, that is, during the memorization period. In other words, under cooling of distant cortex a cell would tend to be similarly activated whether the color in memory was, say, red or green. No cell was found that would show the opposite effect of cooling in its firing, namely, an increase of color discrimination during memory. The loss of the ability of some cells to "discriminate in memory" was accompanied by a descent in the animal's performance. Thus, whether the cooling was inferotemporal or prefrontal, the animal made more errors, as if less capable then at normal temperature to remember the colors through the delays.

These findings indicate that, during short-term mnemonic retention, tonic excitatory influences flow between inferotemporal and prefrontal cortex, and that these influences allow the active retention of the memorandum. This may take place by way of excitatory feedback (re-entrant) loops between distant parts of an extensive memory network that includes neurons in both inferotemporal and prefrontal cortex. The depression of one cortex by cooling lowers the activity of the network as a whole and results in poor retention of the memorandum.

The idea that short-term memory consists in the reverberation of nerve impulses through recurrent networks was first proposed by Hebb (1949). Based on the histological evidence of profuse recurrent fiber connections (Lorente de Nó, 1949), he applied his concept of reverberation to the peristriate cortex as he theorized about the physiology of visual short-term memory. Our cooling experiments, however, suggested that the re-entrant reverberation of visual working memory might be more wide-ranging than Hebb envisioned. It seemed that the sustained activation of a visual memory network required the long cortico-cortical connectivity as well as the local re-entrant circuitry. In any case, the concept of reverberation itself required more empirical support.

To test the reverberation idea in active memory, we have applied to our neuronal data several analytical procedures. The first one was to construct a spiking computer model of a neural network and to train it to retain information for the short term (Zipser *et al.,* 1993). Our purpose was to find out if, in short-term memory, the units of the model behaved like real cortical cells in an active memory network. The architecture of the model was essentially based on re-entry: any unit in the network was connected to all others by re-entrant connections (Fig. 1). The model was trained by the backpropagation method (Rumelhart *et al.,* 1986). This is an error-reducing procedure allowing the network, through successive iterations, to adjust its

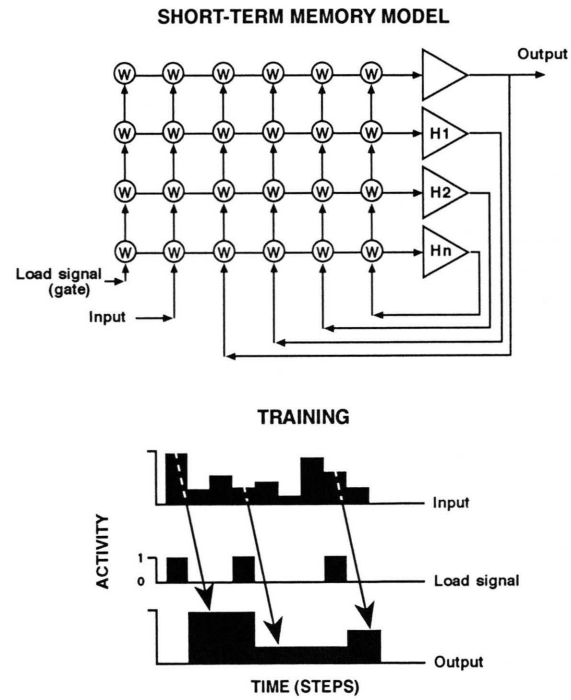

Fig. 1. Recurrent network model of active short-term memory. *Above:* Architecture of the model, basically constituted by a recurrent network with input and output connections. A neuron soma is represented by a triangle with dendrites leading to it from the left. Synaptic weights (w) are established by the training of the network to retain input information in order to emit a later output that is a function of the input. A load signal gates the input. H1, H2, and Hn are hidden units. *Below:* Diagram of the training paradigm (activity refers to firing probability). Arrows mark the input levels that the load signal gates into the network. From Zipser *et al.,* (1993).

synaptic weights to maintain a stable input-output relationship despite fluctuations in input value. The weights stay fixed after training.

Unlike the conventional backpropagation models, our model has a load signal or gate. With open gate (load signal 0), the latest input of a given value is allowed into the network and held at that level. (In the real brain, that gate might be in limbic structures or the prefrontal cortex, signaling and permitting the network to hold the information in that input for subsequent action.) In order to take into account the stochastic nature of real cell firing, all input and output values of the model were assumed to correspond to spiking probabilities (Amit, 1990). It is thereby assumed that in the absence of input or output change a cell will fire randomly with a fixed probability. Changes in input should translate themselves into changes of spike frequency, each with its random distributions of firing intervals. In the fully trained model, single units are substituted by pools of units, to represent the cell assemblies or network components ("netlets") of the real brain. At a given time, the input to a given pool is determined by two vectors: one is a function of the total outputs from network's pools', and the other is determined by external inputs. Two other vectors are the weights of connection with other pools and with external inputs.

In the fully trained network, a trial of the memory task can be simulated by loading an analog input, i.e., the memorandum, and by holding the gate open (load signal 0) through the memory period (delay) until the recall, when the load signal shifts to another level, thus closing again the gate without input. Under these conditions the units of the network behave much like cells in the real cortex during the trials of a memory task, especially if a degree of internal noise is injected in the network (cells in the real brain network also fire in a noisy environment). Output cells behave unremarkably, as they reflect the transfer function from input to output, which is a defining feature of the model. What is truly noteworthy, however, is the behavior of the internal units of the network, the "hidden units," which – after appropriate scaling – exhibit temporal patterns of discharge extraordinarily similar to those exhibited by cortical cells in the memory task (Fig. 2). It is important to note that those patterns of network discharge, which so much resemble those of real

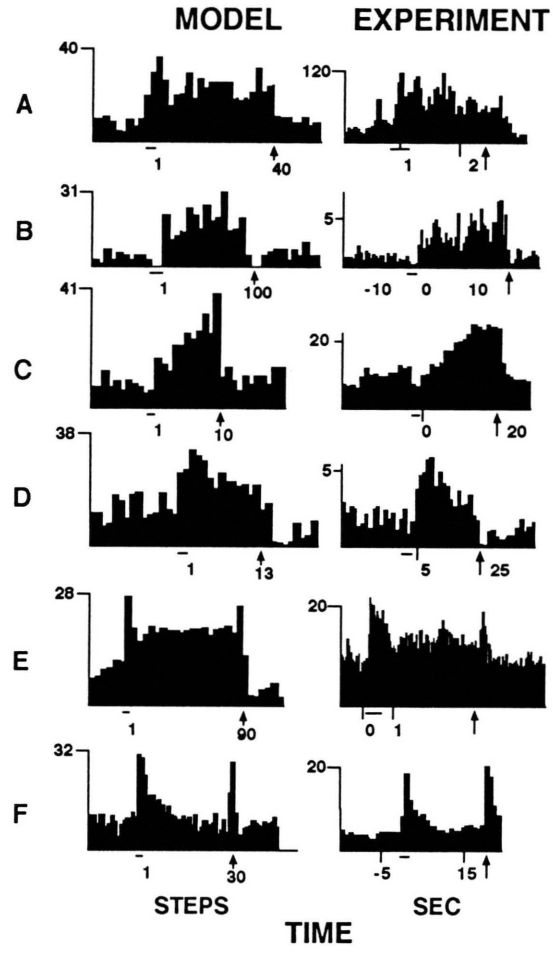

Fig. 2. Comparison of the firing frequency of real cortical neurons with that of hidden units from the model during short-term memory tests. The model's histograms have been scaled to facilitate comparison with those of real cells, which are depicted in the right column. These have been extracted from published microelectrode studies in the monkey. From Zipser *et al.*, (1993).

cells, are part of a repertoire obtained by repeated test of the sample-and-hold operation of the model and generated exclusively by its internal architecture.

In conclusion, the firing patterns of cortical cells in a memory task can be understood as a product of the activation of a fully trained, recurrent memory network with preestablished synaptic weights. The training mechanism by which the weights have been fixed in the model is immaterial to the validity of that conclusion. Indeed, backpropagation has little to do with the hebbian mechanisms

that govern the brain's acquisition of memory. The essential point is that, once the weights have been established, once the memory has been acquired, the short-term activation of the network elicits in its component units similar patterns of firing in the brain as in the model. Thus, the role of recurrence or re-entry in the cerebral dynamics of short-term memory is substantiated by the behavior of the units in a model in which recurrence is an essential feature of functional architecture.

In addition, the model taught us something else of exceptional importance for understanding the discharge patterns of cortical cells in an active memory network. It revealed on close analysis that, in the absence of new inputs or changes in current input, the frequency of the network and its elements will not remain steady, but will drift toward one or several discrete frequencies and will shift between them. A given unit may change repeatedly and in a more or less cyclic manner between several such frequencies. This is especially evident at relatively fine temporal resolutions, as changes occur rapidly, once every few computer-time steps. These changes are not apparent on large temporal scale or smoothed-out frequency histograms.

Although the reasons for this behavior of the model at a fine temporal resolution were unclear, the phenomenon obliged us to take a closer look at the discharge of real – inferotemporal – neurons in the visual memory task. By analyzing spike trains at a fine temporal resolution, as we had not done previously in our experiments, we discovered similar rapid changes between alternating frequencies as those that the model exhibited. We began to think that those changes were inherent in recurrent networks and resulted precisely from their re-entry property. We found literature support for that way of thinking. Cowan (1971) was one of the first to show that recurrent networks of nonlinear units, like ours, will tend to drift toward certain frequencies called fixed-point "attractors."

An unresolved question is whether there are any physiological or anatomical constraints, instead of or in addition to recurrence, that could account for what has been called "attractor dynamics" in cortical networks actively retaining information. One possibility is that attractor dynamics is at least in part determined by intrinsic cellular mechanisms of cortical cells that cannot yet be fully specified (Llinás, 1990). It seems plausible, however, without excluding intrinsic factors, to explore the physiological role of recurrent brain circuitry as the basis for attractor dynamics in active memory. In support of this quest is our previous evidence of cortico-cortical loops in memory (Fuster *et al.,* 1985), which I have mentioned above, and the remarkable similarity of firing patterns in the monkey and in the artificial recurrent network during active short-term memory.

Here our experimental data and deductive reasoning come together. If we assume that a memory network links the associated properties of a memorandum, and if these properties are represented in separate but interconnected cell assemblies of the cortex, it is possible that the activation of the memorandum (i.e., the network) entails the reverberation of impulses between those assemblies. Could that reverberation not be the basis of attractor dynamics in active memory? A neuron's attractor frequency might reflect circulation of impulses between it and other neurons of the network. Each of the cell's attractor frequencies might reflect its functional linkage to a different component of the memorandum. The cell would thus be subject to as many attractors as there are associated features of the memorandum. That functional linkage by re-entry would the a kind of binding in memory, similar to the binding that has been postulated for perceived objects.

Several of these tenets are at least indirectly testable by electrophysiological means. We began by testing the general prediction that any cell in a memory network will exhibit more fluctuations of firing frequency while the network is engaged in active memory than when it is not. Of course, this would be consistent with attractor dynamics. More generally, however, the prediction may be based on the presumption that the cell, as a member of an active ensemble, will be subject to more inputs, excitatory as well as inhibitory, than when the ensemble is in the passive state. Given that cellular action potentials are generated and inhibited in discontinuous fashion by the temporal summation of presynaptic potentials, the more inputs of diverse origin a cell receives the less stable its firing will be, regardless of its average firing frequency. This is especially likely to be the case in neurons of a cortical network representing a complex memorandum.

To test our prediction we chose cells of somatosensory cortex of the monkey during the active memory of the surface features of an object perceived by active touch (Zhou and Fuster, 1996; Bodner *et al.,* 1997). The spike discharge from single cells was recorded as the animal performed the following haptic memory task. Each task trial began with the brief palpation, in the dark, of a sample rod with a special feature (smooth or rough surface, horizontal or vertical edges). After sample palpation, a delay period of 12 sec followed (memorization period), at the end of which the monkey was allowed to palpate two rods and choose for a reward, the one of the two that was identical to the sample. For the analysis, two spike trains were selected on every trial from each cell: the first during the 12-sec baseline period preceding the trial, and the second during the 12-sec delay (memory period).

As in other regions of cortex (Fuster, 1995), "memory cells" can be found in somatosensory cortex. These cells go into higher firing frequency as soon as the animal is obliged to retain a stimulus in short-term memory (delay). Some such cells show a different level of firing depending on the particular stimulus in memory. Included for our analysis, however, were cells whose *average* firing during memorization did not significantly differ from baseline (spontaneous firing between test trials). The objective, in any given cell, was to expose rapid fluctuations of firing – up or down – within the 12 seconds of memorization, irrespectively of the cell's average firing for that period.

Though our analysis of inferotemporal cells had provided us with some general idea of the velocity of "attractor shifts," here in parietal cortex we had no *a priori* notion of that velocity and thus of the time resolution needed to investigate quantitatively that issue in a massive collection of spike trains from many neurons. Furthermore, we had no grounds to assume that all cells in our sample would show frequency fluctuations, much less in the same range of incidence. We needed a kind of multiple filter of cell-firing transients.

One such "filter" is the systematic binary mapping of discrete temporal events – in our case, cell spikes. The method is useful for detecting fluctuations of frequency in time series without preconceived notion of the precise range in which they occur. On each of the spike trains from a somatosensory cell obtained during intertrial baseline and "delay" periods, the following mapping procedure was conducted on a computer (Fig. 3): (a) the 12-sec period was segmented into equal-size bins, and (b) a 1 or a 0 was assigned to each bin depending on whether it contained any spike or not. In the resulting temporal map or binary curve, a transition of frequency was defined as a transition from

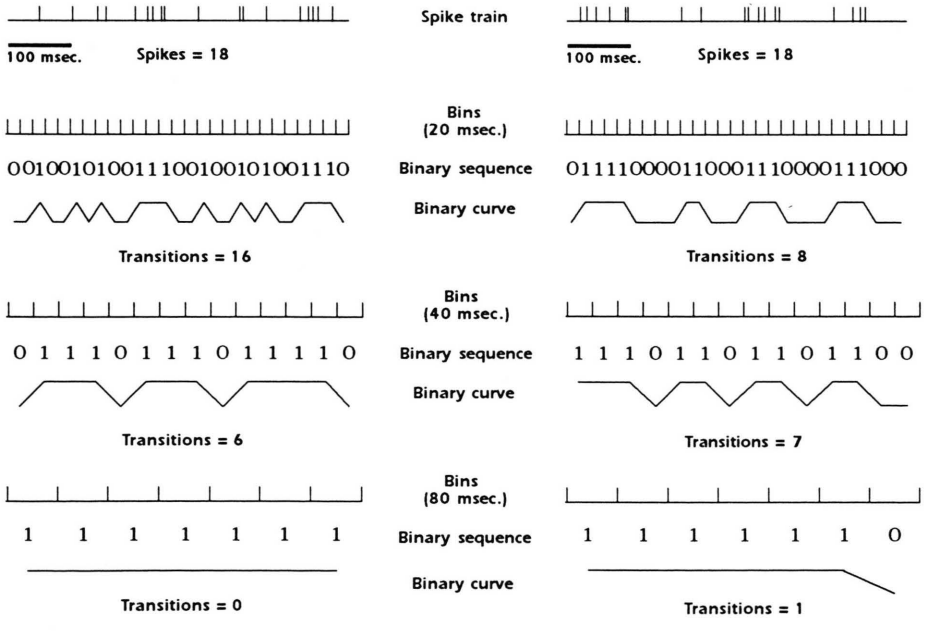

Fig. 3. Binary mapping procedure. For illustration, two spike trains (about 0.5-sec long) with identical number of spikes, but different temporal distribution, are converted into binary sequences by using three levels of time resolution, that is, three different bin sizes (20, 40, and 80 msec). See text for additional explanation. From Bodner *et al.,* (1997).

0 to 1 or vice versa. (By segmenting the record into bins of multiple sizes, the method became essentially a low-pass filter of spike-frequency changes or transitions.) In the analysis of each train we systematically used all bin sizes between 1 and 100 msec in 1-msec increments, thus 100 filters or time resolutions.

Transitions were found more numerous, and at more bin sizes, during "delays" (memory periods) than during inter-trial baseline periods. Thus, in active memory, the majority of cells showed more fluctuations of firing frequency than in baseline periods. As predicted, the cells appeared subject to increased inputs while the monkey memorized the palpated object. The increase in transitions need not be accompanied by average frequency change. Differences in transitions between baseline and delay were most conspicuous at bin sizes between 20 and 50 msec, with a peak around 28 msec.

In conclusion, as expected, cortical cells in active memory appeared to fluctuate more often between different frequencies than in baseline conditions. The analysis does not specify precisely, however, what those frequencies were. That the difference was most evident within a range of resolutions simply implies that the transitions themselves occurred most often at the frequencies corresponding to that range, but it does not signify the cell's discharge between transitions, except within wide limits – from 10 to 50 Hz. An increase of shifts between attractor discharges in that range would be consistent with multiple re-entries from within a memory network upon a given cell. The attractors, however, need not be "fixed-point." Periodic patterns of frequency fluctuations ("limit-cycle attractors") are also possible, and these would be even more consistent with re-entry. Such patterns are currently under investigation.

Amaral D. G. (1987), Memory: Anatomical organization of candidate brain regions. In: Handbook of Physiology; Nervous System, Vol. V: Higher Functions of the Brain, Part 1 (F. Plum, ed.). Amer. Physiol. Soc. Bethesda, pp. 211–294.

Amit D. J. (1990), Attractor neural networks and biological reality: associative memory and learning. Fut. Gen. Comp. Sys. **6,** 111–119.

Bauer R. H. and Fuster J. M. (1976): Delayed-matching and delayed-response deficit from cooling dorsolateral prefrontal cortex in monkeys. J. Comp. Physiol. Psychol. **90,** 293–302.

Bodner M., Zhou Y. and Fuster J. M. (1997), Binary mapping of cortical spike trains in short-term memory. J. of Neurophysiol. **77,** 2219–2222.

Cowan J. D. (1971), Stochastic models of neuroelectric activity. In: Statistical Mechanics (S. A. Ricce, K. F. Fread, and J. C. Light, eds.). University of Chicago, Chicago, pp. 109-127.

Fuster J. M. (1985), The prefrontal cortex and temporal integration. In: Cerebral Cortex, Vol. **4,** (E. G. Jones and A. Peters, eds.). Plenum Press, New York, pp. 151–177.

Fuster J. M. (1995), Memory in the cortex of the primate. Biol. Res. **28,** 59–72.

Fuster J. M., Bauer R. H. and Jervey J. P. (1981), Effects of cooling inferotemporal cortex on performance of visual memory tasks. Exp. Neurol. **71,** 398–409.

Fuster J. M., Bauer R. H. and Jervey J. P. (1982), Cellular discharge in the dorsolateral prefrontal cortex of the monkey in cognitive tasks. Exp. Neurol. **77,** 679–694.

Fuster J. M., Bauer R. H. and Jervey J. P. (1985), Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. Brain Res. **330,** 299–307.

Fuster J. M. and Jervey J. P. (1982), Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. J. Neurosci. **2,** 361–375.

Hayek F. A. (1952), The Sensory Order. University of Chicago Press, Chicago.

Hebb D. O. (1949), The Organization of Behavior. John Wiley & Sons, New York.

Lashley K. S. (1950), In search of the engram. Symp. Soc. Exp. Biol. **4,** 454–482.

Llinás R. (1990), Intrinsic electrical properties of nerve cells and their role in network oscillation. Cold Spr. Harbor Symp. Quant. Biol. **55,** 933–938.

Lorente de Nó R. (1949), Cerebral cortex: Architecture, intracortical connections, motor projections. In: Physiology of the Nervous System (J. Fulton, ed.). University Press, New York, pp. 288–315.

Mountcastle V. B. (1978), An organizing principle for cerebral function. The unit module and the distributed system. In: The Mindful Brain (G. M. Edelman and V. B. Mountcastle, eds.). Plenum Press, New York, pp. 7-50.

Rumelhart D. E., Hinton G. E. and Williams R. J. (1986), Learning representations by back-propagating errors. Nature **323,** 533–536.

Squire L. R. and Zola-Morgan S. (1988), Memory: brain systems and behavior. Trends Neurosci. **11,** 170–175.

Swartz B. E., Halgren E., Fuster J. M., Simpkins F., Gee M. and Mandelkern M. (1995), Cortical metabolic activation in humans during a visual memory task. Cereb. Cortex **3,** 205–214.

Zhou Y.-D. and Fuster J. M. (1996), Mnemonic neuronal activity in somatosensory cortex. Proc. Natl. Acad. Sci. USA **93,** 10533–10537.

Zipser D., Kehoe B., Littlewort G. and Fuster J. (1993), A spiking network model of short-term active memory. J. Neurosci. **13,** 3406–3420.