Journal of Computational Neuroscience, 5, 353–364 (1998)

Adapting a Feedforward Heteroassociative Network to Hodgkin-Huxley Dynamics

WILLIAM W. LYTTON

billl@neurosim.wisc.edu

Department of Neurology and Neuroscience Training Program, Wm. S. Middleton VA Hospital, University of Wisconsin, 1300 University Ave., MSC 1715, Madison, WI 53706-1532

Received May 27, 1997; Revised October 1, 1997; Accepted December 16, 1997

Action Editor: Sejnowski

Abstract. Using the original McCulloch-Pitts notion of simple on and off spike coding in lieu of rate coding, an Anderson-Kohonen artificial neural network (ANN) associative memory model was ported to a neuronal network with Hodgkin-Huxley dynamics. In the ANN, the use of 0/1 (no-spike/spike) units introduced a cross-talk term that had to be compensated by introducing balanced feedforward inhibition. The resulting ANN showed good capacity and fair selectivity (rejection of unknown input vectors). Translation to the Hodgkin-Huxley model resulted in a network that was functional but not at all robust. Evaluation of the weaknesses of this network revealed that it functioned far better using spike timing, rather than spike occurrence, as the code. The algorithm requires a novel learning algorithm for feedforward inhibition that could be sought physiologically.

Keywords: associative memory, inhibition, artificial neural network, hippocampus

1. Introduction

One of the most widely used artificial neural network models for memory in the brain is the Hopfield network (Hopfield, 1982,1984), a symmetrical recurrent network that settles to point attractors that represent individual memories. This type of network has been heavily explored with a goal of altering it so that might more closely fit known neurophysiology. For example, it has been shown that similar models could work without requiring symmetrical connections (Treves and Rolls, 1991, 1994). Other studies have suggested how a point attractor system could play a role in a continuous dynamical system like the brain (Amit, 1989). Variants of the Hopfield network have been directly applied to aid in our understanding of brain areas involved in memory such as hippocampus (Gardner-Medwin, 1976; O'Reilly and McClelland, 1994; Treves and Rolls, 1994) and piriform cortex (Barkai et al., 1994). In the case of the hippocampus, the focus has been on area CA3, since it is there that the necessary combination of high interconnectivity and relatively high firing rates can be found.

If one looks at the hippocampus as a whole, several features that are notable in comparison to other brain areas are precisely the opposite of those required for a Hopfield associative memory. Whereas in most brain areas, forward projections are matched to corresponding back projections, the hippocampus is notable for its apparent one-way flow of information. It is precisely this attribute that has made the classical tri-synaptic pathway (perforant path to dentate granule cell to mossy fiber to CA3 pyramidal cell to Schaeffer

collateral to CA1) so valuable for studying the details of synaptic response. Although this simplified view of connectivity has been called into question in recent years (Amaral et al., 1990; Scharfman, 1994), it does appear that these projections are largely unidirectional. The idea of using feedforward rather than recurrent networks to understand the hippocampus hearkens back to the classical Marr model (Marr, 1971; Willshaw et al., 1969; Willshaw and Buckingham, 1990).

In most artificial neural network models, including the Hopfield network, the state variable associated with the individual unit is assumed to be a correlate of a nearly continuous firing rate. This permits the use of both positive and negative values that correspond to firing rates above and below the rate of ongoing spontaneous activity. However, some cortical cell types show relatively low rates of firing or fire in bursts rather than regularly. In the hippocampus, dentate gyrus granule cells may show extremely low firing rates. Although this may in part be an artifact of *in vitro* slice recording (Fricke and Prince, 1984), the strong adaptation and evidence for powerful feedback inhibition suggests that low firing may be an attribute *in vivo* as well (Buckmaster and Schwartzkroin, 1995). Rate coding would, of course, not be possible for neurons that fire infrequently. Instead, one might want to consider the much simpler view of neuronal function put forth by McCulloch and Pitts (1943). By their account, the neuronal signal was not firing rate but was a simple binary code determined by whether the neuron fired at all.

The current study uses the neural coding scheme proposed by McCulloch and Pitts, in the context of the feedforward networks proposed by Willshaw et al. (1969) and by Marr (1971) and subsequently extended and implemented by Anderson (1972) and Kohonen (1972). The Anderson-Kohonen associative memory model is adapted to a simple no-fire/fire code by extending the standard linear-algebra formulation and then translating it into Hodgkin-Huxley dynamics. A heteroassociative memory is used. This term was coined by Kohonen to describe a network where the input and output vectors are not identical, as they are in an autoassociative memory. The first step of this process requires modification of the algorithm to work with 0/1 (clear bit/set bit) vectors instead of -1/1 vectors. This is necessary because a fire/no-fire code cannot represent negative values. Having demonstrated the functionality and utility of this extension of the Anderson-Kohonen algorithm, the model is translated into a neuronal network using simple compartment neuron models with Hodgkin-Huxley kinetics. This translation leads to an alternative method of neural coding that utilizes spike timing.

2. Methods

Simulations were run in the NEURON simulator (Hines, 1993) with vector extensions provided by Z. Mainen. Algorithms for the linear algebra simulations are given in the results. The compartmental models utilized the following parameters. The individual neuron was represented by a cylindrical single compartment of diameter 10 μ and length 31.8 μ with $\overline{g}_{\mathrm{leak}}$= $3 \cdot 10^{-3}$mS/cm^2; $\mathrm{E}_{\mathrm{leak}}$ was set to produce a steady-state resting membrane potential of -60 mV. Sodium channel and potassium channel were modified from Hodgkin-Huxley using the Borg-Graham parameterization (Borg-Graham, 1991; Lytton and Sejnowski, 1991). The Na^+ channel used standard m^3h state variables with $\overline{g}_{\mathrm{Na}}$=30mS/cm^2, E_{Na}=60 mV, $z_m = 4, \gamma_m = 0.5, \alpha_0 = 4.5, V_{1/2} = -33.5, \tau_{min} = 0.02$, $z_h = -6, \gamma_h = 0.3, \alpha_0 = 0.095, V_{1/2} = -39, \tau_{min} = 0.25$. The K^+ channel used a n^3 state variable with $\overline{g}_{\mathrm{K}}$= 25mS/cm^2, E_{K}= -75 mV, $z_m = 2.8, \gamma_m = 0.7, \alpha_0 = 0.13, V_{1/2} = -18, \tau_{min} = 0.3$.

Generalized glutamatergic and GABA conductances were used for excitatory and inhibitory synapses, respectively. The parameterization used was a modified version of that of Destexhe et al. (Destexhe et al., 1994a; 1994b; Lytton, 1996) with the following parameters. GLU: $\overline{\mathrm{g}}$= 1nS, C_{dur} = 1 ms, α = 1 mM/ms, β = 0.35/ms, $\mathrm{E}_{\mathrm{syn}}$= -20 mV, delay = 0 ms. GABA: $\overline{\mathrm{g}}$= 1nS, C_{dur} = 1 ms, α = 1 mM/ms, β = 0.35/ms, $\mathrm{E}_{\mathrm{syn}}$= -100 mV, delay = 2 ms. Note that the synaptic reversal potentials were situated so as to give them equal and opposite driving force from rest in order to permit setting synaptic conductances without further scaling.

A network with 40 neurons and 300 inputs (12,000 synapses) took 33 seconds to run 1 second of simulated time on a SUN Sparcstation 10. Assessment of network capacity with 1600 patterns in this network took 14 hours, 20 minutes.

3. Results

Holographs, unlike standard photographs, show graceful degradation with damage. This makes them an attractive conceptual model for human memory (Pri-

bram, 1969). Generalization of this concept to heteroassociative memory models suggested that specific neural algorithms could be fashioned from this metaphor (Willshaw et al., 1969; Marr, 1971). Further development of this idea demonstrated the potential of these systems for information storage (Anderson, 1972; Kohonen, 1972).

3.1. Algorithm

The basic algorithm is straightforward and widely applied: the normalized outer-product of the output vector ($\hat{o}$) and the input vector ($\hat{i}$) yields a connection matrix (A, Eq. (1)) that will map the input vector onto the output vector. The normalizing value c is the inverse of the power of the input vector: $1/m_i$ if -1/1 vectors are used (Eq. (2)) and $1/p_i$ for 0/1 vectors, where m_i is input vector length and p_i is the number of set bits in the vector. This works since the dot-product of a vector times itself (represented either as $\hat{i}^T\hat{i}$ or $\hat{i}\cdot\hat{i}$) is the squared norm of the vector:

$$A = c\cdot\hat{o}\hat{i}^T \tag{1}$$

$$A\hat{i} = c\cdot\hat{o}\hat{i}^T\hat{i} = \hat{o}\frac{1}{m_i}(\hat{i}^T\hat{i}) = \hat{o} \tag{2}$$

$$A = c\cdot\sum_{\mu=1}^{N}\hat{o}_\mu\hat{i}_\mu^T \tag{3}$$

$$\hat{o}_* = \Theta\left(A\hat{i}_*\right). \tag{4}$$

This result can be generalized for multiple input-output pairs by simply summing the outer product (Eq. (3)). The network described is a basic feedforward associative network (Eq. (4)), where Θ is a thresholding step function that will produce a set bit (that is, 1) for values above a threshold value θ and a clear bit (that is, 0 for 0/1 vectors or -1 for -1/1 vectors) for values below θ (Anderson, 1972; Kohonen, 1972). An arbitrary input $\hat{i}_*$ will be mapped onto output $\hat{o}_*$. If $\hat{i}_*$ is the input for one of the preselected input-output pairs, then $\hat{o}_*$ should be the corresponding output. If the input-output pairs are chosen with $\hat{i}_k = \hat{o}_k$, the network is autoassociative; if input and output differ within each pair, the network is heteroassociative.

A convenient symmetry is present when using binary vectors that take on values of -1 or 1. Many physical quantities, such as the light of a hologram, the charge on a transistor, the neurotransmitter in a synapse, or the firing frequency of a neuron, do not have this symmetry. Instead these phenomena are positive values – physical quantities that can be measured. In the case of a neural spike, the natural phenomenon is either there or not there (all or none). This is the simple view of the neuron originally proposed by McCulloch and Pitts, who represented these states as 1 and 0, respectively.

While the single-pair correlation matrix of Eq. 2 perfectly reproduces the desired output, the more useful multiple-pair matrix of Eq. 3 introduces crosstalk that adds extraneous signal to the output. For a given vector-pair $\hat{i}_k : \hat{o}_k$, the response is the sum of the single-pair response and the crosstalk from additional vectors:

$$A\hat{i}_k = c\cdot\hat{o}_k\hat{i}_k^T\hat{i}_k + c\cdot\sum_{\mu\neq k}\hat{o}_\mu\hat{i}_\mu^T\hat{i}_k. \tag{5}$$

Using 0/1 vectors, the input vectors must be orthogonal in order to produce 0 crosstalk. This orthogonality requirement explains why many such networks are designed to work only with extremely sparse input vectors. The problem with sparse vectors is that there are not very many of them. Obviously, if only 1 input line is to be active out of m, there are only m possible input vectors. More generally, if p lines are active, there are only m/p orthogonal vectors possible, only a tiny proportion of the ${}_mC_p = m!/((m-p)!\cdot p!)$ possible combinations that might otherwise be used (specifically, $1/{}_{m-1}C_{p-1}$). A 100-axon input, for example, would only provide 50 orthogonal 2:98 vectors, rather insignificant compared to the $\approx 10^{29}$ possible 50:50 vectors.

This paucity of orthogonal vectors suggests the use of nonorthogonal vectors. In this case, the use of -1/1 binary vectors permits one to minimize crosstalk (Anderson, 1972). Restricting ourselves to vectors with half the bits set ($p = m/2$), randomly chosen -1/1 vectors will tend to produce minimal crosstalk since the dot-products of randomly chosen vectors will tend toward 0. The expectation value of crosstalk can be expressed for a/b vectors where a is the value of the clear bit (generally 0 or -1) and b the value of the set bit (generally 1):

$$\langle\hat{i}_j^T\hat{i}_k\rangle = \frac{p!^2}{(2p)!}\cdot\sum_{u=1}^{p}({}_pC_u)^2\cdot\left(2uab + (p-u)\cdot(a^2+b^2)\right). \tag{6}$$

Assuming that vectors are randomly chosen, this will approximate the dot-product interference $\hat{i}_\mu^T\hat{i}_k$ from

Eq. 5. For -1/1 vectors this is 0, while for 0/1 vectors of length 100 it is 25. Not only does the use of 0/1 vectors substantially increase the crosstalk, but it also decreases the signal power since $\hat{i}_k^T \hat{i}_k = 100$ for -1/1 vectors of length 100 but only 50 for 0/1 vectors. Looked at another way, the squared signal to noise ratio goes from being infinite to 4.

Ignoring the thresholding step and the normalizing constant c for the time being, we want $A\hat{i} = \hat{o}$ Eq. (2). Instead, we have $A\hat{i} = \hat{o} + \text{residual}$. We can remove this residual by subtracting off the expected value of the crosstalk, combining Eq. (4), (5) and (6). We turn (5) around to put the desired output on the left side, replacing both the dot-product and the crosstalk term with their expected values $\langle kk \rangle \equiv \langle \hat{i}_k^T \hat{i}_k \rangle; \langle jk \rangle \equiv \langle \hat{i}_j^T \hat{i}_k \rangle$ (from Eq. (6)):

$$A\hat{i}_k \cong \langle kk \rangle \cdot \hat{o}_k + \langle jk \rangle \cdot \sum_{\mu \neq k} \hat{o}_\mu \tag{7}$$

$$\langle kk \rangle \cdot \hat{o}_k \cong A\hat{i}_k - \langle jk \rangle \cdot \left[\left(\sum_{all\ \mu} \hat{o}_\mu \right) - \hat{o}_k \right] \tag{8}$$

$$(\langle kk \rangle - \langle jk \rangle) \cdot \hat{o}_k \cong A\hat{i}_k - \langle jk \rangle \cdot \sum_{\mu} \hat{o}_\mu \tag{9}$$

$$\hat{o}_k \cong \frac{1}{f} A\hat{i}_k - \frac{\langle jk \rangle}{f} \cdot \sum_{\mu} \hat{o}_\mu, \tag{10}$$

where
$f \equiv (\langle kk \rangle - \langle jk \rangle)$. In the simulations, all weights were normalized by number of patterns N, and the product was thresholded at $\theta = (a + b)/2$ to arrive at the final output vector.

"Training" of the network was done to produce an excitatory matrix A/f and an inhibitory projection $(\langle jk \rangle / f) \cdot \sum_\mu \hat{o}_\mu$, that is effectively feedforward, since it is activated directly by the inputs rather than by the excitatory units. The sum of outer products used to create A is equivalent to an iterative Hebbian process (see Anderson and Rosenfeld, 1998, Chap. 14). The inhibitory feedforward projection could also be created using a Hebbian process. This will be further explored in the Discussion (Section 4.2).

3.2. Simulations: Performance and Capacity

Implementation of the algorithm showed generally good ability to reproduce the desired output vector in response to the corresponding input vector (Fig. 1). Errors were generally infrequent but tended to cluster in particular output vectors. Errors for each vector pair could be summarized by looking at percent error. (Percent error is calculated by first calculating the Hamming distance– the number of bits that are wrong; this is normalized by the number of output bits (m_o) to obtain percent Hamming distance or percent error.) The majority of vector pairs were learned perfectly or very well, but those vector pairs that were not well learned tended to show many errors (Fig. 2).

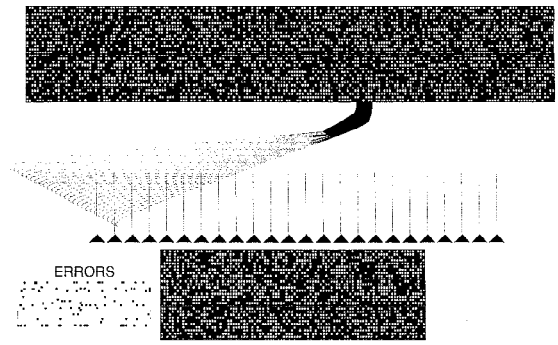

Fig. 1. Schematic of typical input/output results with schematic showing network layout (m_i=200 bits,m_o=100 bits,N=30 patterns). Each row in the input set is paired with the same number row in the output set with each vector arrayed horizontally showing 0 (black) or 1 (white). Errors in the output set are shaded; their locations are also illustrated in the inset. Average percent Hamming distance for output in this simulation was 2.6%.

Performing many simulations, it appeared that network capacity was related to both the number of synapses (input vector length times output vector length, $S = m_i * m_o$) and to convergence (input vector length divided by output vector length, $C = m_i/m_o$) in a complex way. The following empirical formula gave consistent predictions of capacity across a variety of scales and convergence values tested:

$$N = \frac{C^{0.2} \cdot S^{0.5}}{r}, \tag{11}$$

where r is a scaling factor whose choice determines N, the number of patterns that could be stored and still produce a consistent accuracy level.

Increasing the number of stored patterns yielded progressively poorer output accuracy (Fig. 2B). For example, a network with input vectors of length 200 and output vectors of length 100 ($S = 2 \cdot 10^4, C = 2$) learned 162 patterns (r=1) with an average percent error of 20±8%. The large standard deviation is due to the fact that, as noted, a few vectors were generally stored poorly (Fig. 1, 2A). Slightly better results could

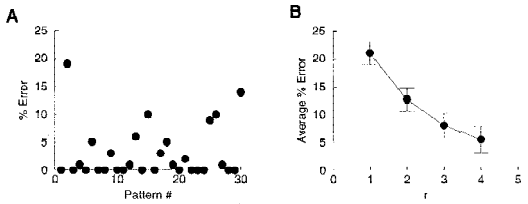

Fig. 2. Capacity and accuracy of the network with 0/1 vectors. **A** In a given network, some patterns were poorly stored. Same network shown in Fig. 1 with m_i=200,m_o=100,C=2,N=30. 23 input-output pairs were learned well, yielding 0-1% error. However, 5 pairs gave 9-19% errors. **B** Average and standard deviation of percent error as a function of r Eqn. 11. In each case 100 different networks were tested varying convergence C from 1 to 4 and output vector length m_o from 50 to 100.

be obtained using -1/1 vectors, while worse results occurred using vectors with a clear bit value greater than 0.

3.3. *Simulations: Pattern Completion and Recognition*

In addition to assessing capacity, the ability of the network to perform pattern completion and recognition was also evaluated. Unlike a recurrent dynamic network of the Hopfield type, this network algorithm does not proceed to a point attractor. Therefore, vectors that were not in the input set would generally be expected to produce outputs that are not in the output set. In fact, a small amount of completion was seen, particularly in networks with large convergence. As an input vector was gradually degraded by 1 to 5% from its original form, the output vectors would remain fairly close to the corresponding output (Fig. 3A). However, degradation of greater than 10% produced increasingly large percent Hamming distance from the desired output. Additionally, completion was extremely inconsistent. A single input that completed well with one random degradation might show no completion using another random degradation of identical Hamming distance.

The tendency of degraded input vectors to produce extremely variable output vectors also limited this network's ability to recognize an input– that is, to determine whether this was a known input. In general however, the power of the output dropped off as a function of the distance of the associated input from a known input, giving a clue as to whether the output was a useful association to a known memory (Fig. 3A). This fall-

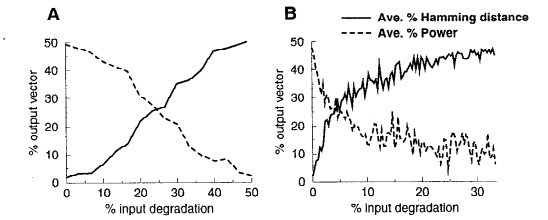

Fig. 3. Recognition of a vector could be assessed by measuring the power of a vector (sum of the bit values=number of 1s). **A** In the basic model, power fell off gradually as the input vector was gradually degraded from a known vector. (Input power was kept constant.) **B** Requiring agreement in 4 of 5 convergent parallel networks formed with identical output but different input vector sets, summing units provided more rapid fall-off in power.

off in output power occurred as long as input power remained at 50%, the power used for all of the known inputs. Inputs with greater or lesser power would, as expected, produce outputs with increased or decreased power, respectively.

To provide recognition in the face of input power variation, a second layer of recognition units was used, one for each output bit. Each recognition unit received convergent input from the corresponding output lines from each of five separate network matrices. Each separate network matrix produced identical outputs but utilized different input vectors. With a known input, each of the five inputs to a recognition input would tend to "vote" in the same way. With an unknown input vector, the output from each of the networks would be largely uncorrelated, leading to no activation of the recognition unit. The multiple converging networks could be seen as corresponding to the multimodal inputs seen by entorhinal cortex coming from different sensory channels. For example, the separate matrices would correspond to pathways specialized for visual, auditory, and somatosensory input. Each matrix would produce the same internal memory index in response to evocation of a memory via a particular sensory modality (Teyler and DiScenna, 1986).

By adjusting the threshold of the summing recognition units, specificity was adjusted for the known inputs, producing a rapid drop-off in power as inputs become progressively more degraded (Fig. 3B). The power and distance curves cross at about 25%. This 25% power point could be used as the cut-off for recognition, since output vectors with less than 25% power tend to be more than 25% distant from the desired

associated output vector. Adjusting recognition unit threshold would alter the point where average Hamming distance equaled total activity, thus altering system sensitivity and specificity. For example, a crossing value of 10% would give a system with lower sensitivity (ability to complete from a degraded input) but higher specificity (ability to distinguish known inputs from unknown inputs).

3.4. Translation into Hodgkin-Huxley dynamics

To translate into a simple Hodgkin-Huxley one-compartment neuron model network, the values of the connectivity matrix A (Eq. (3)) and the feedforward inhibitory input (Eq. (10), second term) were mapped directly onto the respective excitatory and inhibitory synaptic conductances. For simplicity, only two types of synapses were utilized, a generalized glutamatergic excitatory synapse and a generalized GABAergic inhibitory synapse (see Methods). With careful adjustment of synaptic conductances ($\overline{g}_{\mathrm{GLU}}$ and $\overline{g}_{\mathrm{GABA}}$) and membrane leak current ($\overline{g}_{\mathrm{leak}}$) it was possible to produce a working network (Fig. 4). However, the parameter space for these three parameters was narrow and did not always generalize to another set of input-output pairs. In fact, most parameter choices would give one of two results. Either most neurons would be quiescent regardless of input or most neurons would fire regardless of input. While examining these traces, it was noted that in cases where most or all of the neurons spiked, those that represented set bits (1) tended to spike considerably earlier than those representing clear bits (0), suggesting that a suitable coding scheme for this network would be timing rather than simple spike occurrence (Hopfield, 1995).

Timing thresholds gave much better performance that was generally comparable or better than that of the original linear algebra algorithm. Figure 5 shows activity for the 40 units in a network with convergence of 7.5. Each individual unit trace is tagged with a square showing the value of the target bit: filled square for 1 and open square for 0. Timing threshold in each column is shown by the dashed line. In general, spikes representing 1s occur before the dashed line while those representing 0s occur after it. Spike timing therefore reconstructs the bit code. In this figure, unit 7 is incorrect, spiking before the time threshold, and unit 27 is indeterminate, occurring within 1 ms of the threshold. All other spikes are at least 5 ms away from the threshold.

Hopfield pointed out that a timing code would be expected to be robust to alterations in stimulus strength (Hopfield, 1995). This could be tested in our model by globally altering the synaptic strength of the inputs. As strength was increased, firing tended to occur earlier in all the units, but the relative firing times that made up the code were preserved. The timing code was also resistant to parameter changes such as alterations in channel conductances. Again, although absolute timing of spikes changed, the required relative timing remained. This held true for other global parameters as well. A critical scaling parameter in the original matrix memory was the expected value of dot products between vectors defined in Eq. 6). This complex expression, requiring the global summation of various factorial terms, would be impossible to calculate using local processing rules in a neural computation. The factor f given in Eq. 10), a scaling factor for both excitatory and inhibitory synapses, can be eliminated entirely without compromise of performance. The factor $\langle jk \rangle$ of Eq. (10) determines relative strength of excitatory and inhibitory input. Although it could not be eliminated, it could be varied by up to 15%.

Interpretation of the timing code would not, of course, depend on the prior assignment of a specific timing threshold, but on the fact that the set-bit signals arrive first and also tend to be more tightly clustered in time. A measure of network capacity would therefore involve the distinguishability of set-bit and clear-bit spike arrival times for individual patterns. This was assessed using the Mann-Whitney U test. Utilizing all of the data from all neurons and all input-output pairs, this statistical method paradoxically suggested an increasing difference between the two populations with increasing number of input-output patterns, even though the spike-time means moved progressively closer. The increase in number of samples with increasing number of patterns produced this increase in significance. When the number of samples was fixed at 100, a number that might be attainable in physiological experimentation, z-scores dropped with increasing number of patterns, corresponding to an increase in p-value (Fig. 6). Using this measure, the highest capacity fully assessed (1,800 patterns) still showed a significant difference with $p < 0.001$.

Recurrent connections within the network could permit formation of stable attractors that would allow traditional rate coding. To test this, a bidirectional asso-

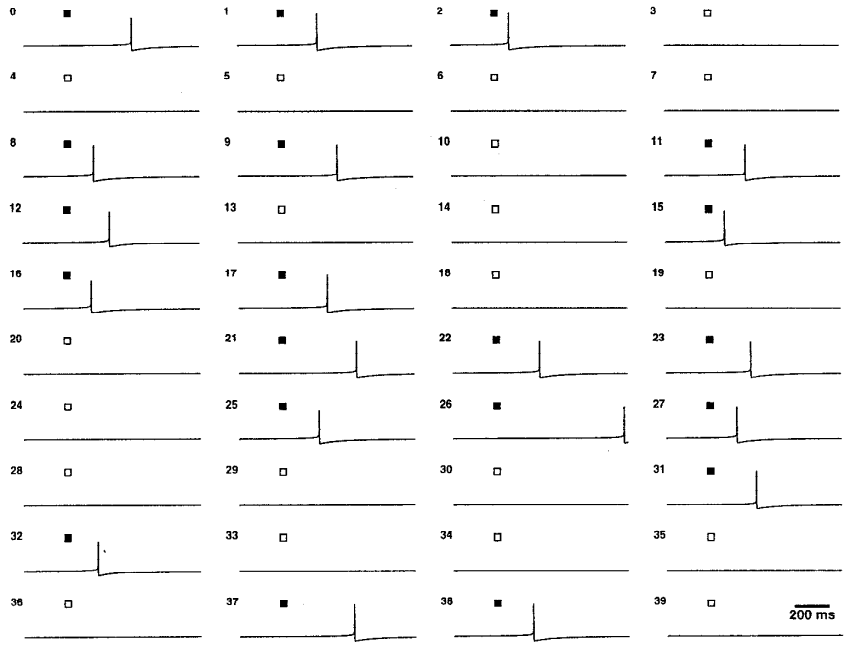

Fig. 4. Hodgkin-Huxley variant of the feedforward network showed desired behavior: no firing for clear bits (open squares) and 1 spike for set bits (filled squares). Each trace shows 1 sec of simultaneous simulated time for a single unit. Overall performance was poor (29% average percent Hamming distance) and performance was very sensitive to parameter change.

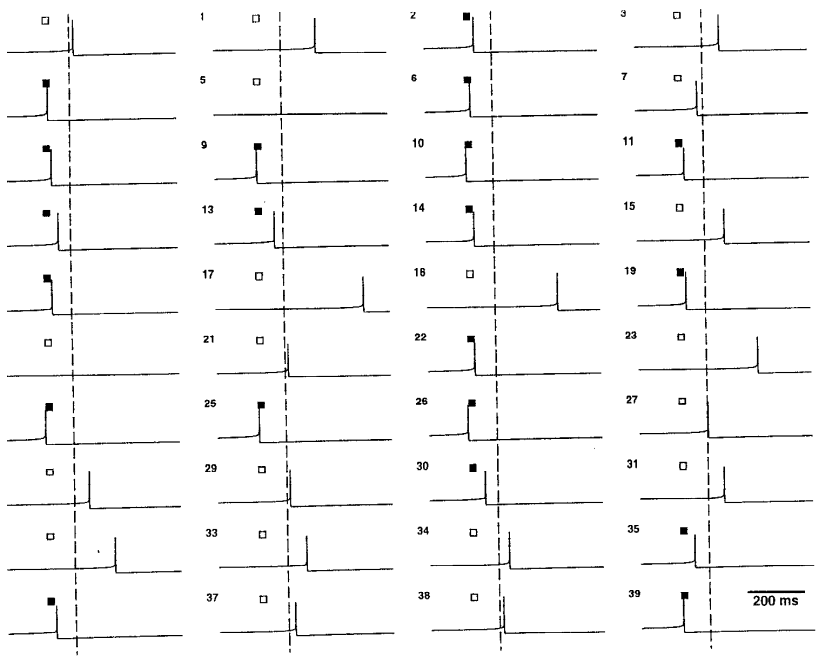

Fig. 5. With increased excitatory tone, almost all neurons fired. Spike timing, relative to a a time threshold (dashed lines), could be used as a code with early spikes representing set bits (filled squares) and late spikes representing clear bits (open squares). Average percent Hamming distance in this case was 2.5% (error for unit 7; note that unit 27 was also very near threshold). Changes in parameters did not have a large effect on accuracy but only changed the location of the timing threshold.

ciative memory (BAM) was implemented for orthogonal 0/1 vectors by connecting two heteroassociative memories together. By increasing excitatory strength,

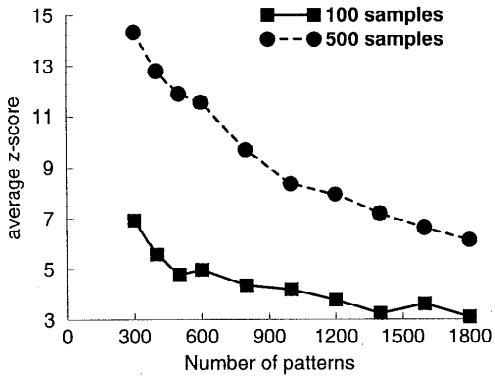

Fig. 6. Application of Mann-Whitney U test to assess network capacity gave widely different results depending on the number of units sampled. In this case a 300/40 network was assessed with varying numbers of input/output pairs (abscissa). All z-scores shown correspond to a $p < 0.001$ distinction between spikes times corresponding to 0s (group A) and 1s (group B). In each case several random samplings were done (6 for 100 samples, 10 for 500 samples) and z-scores were averaged.

this network could produce sustained firing in the set bit neurons with no firing in the clear bits. This implementation has not yet been extended to the case of -1/1 vectors.

4. Discussion

Our study illustrates the fruitfulness of considering both the algorithmic and implementation levels of investigation when trying to understand a particular brain function. These levels, identified by Marr, correspond to what now might be considered artificial neural networks (ANN) and "realistic" neural networks (RNN), respectively (Abbott and Kepler, 1990). The former connects more readily with the level of function, while the latter connects with the brain matter (the hardware) that actually has to make the system work. The bottom-up RNN approach complements the top-down ANN approach, both limiting and extending concepts that arise through the algorithm.

4.1. Coding: Spike Rate or Spike Time

In artificial neural network models the simplified thresholding or sigmoid units have state variables that stand for neuronal firing rates (rate coding). Steven's "slow potential theory" explains how temporal integration would permit synaptic transduction of such signals (Stevens, 1966). However, as Stevens notes, there is effectively no response to most inhibitory events unless the neuron has a basal spontaneous firing rate. In this case, negative state values stand for the amount of decrease from the spontaneous rate. Since negative weight values stand for inhibitory synapse strength, a negative state can combine with a negative weight to produce a positive effect: a reduction in presynaptic inhibitory firing will give increased firing in the postsynaptic neuron.

These idealizations undoubtedly work well for some cell types– namely, those that show a relatively rapid, relatively regular, spontaneous firing rate. However, some cell types do not show either the regularity or the underlying rate required to fit the assumptions of slow potential theory. The most extreme case, probably rare in nature, is that studied here: the neuronal signal is a single spike. However, any neuron that shows rare spike or burst firing will also be unable to directly transduce negative values. In cases where firing frequency is not used as a signal, spike timing may be the critical variable transmitting information (Hopfield, 1995).

The recent revival of interest in spike timing as a primary neural code can be traced to two separate findings. First, the early emphasis of electroencephalographers on oscillatory rhythms of the brain (Bullock, 1993, 1997) has found new life in the work of Gray and Singer (1989; Gray et al., 1989) and others (deCharms and Merzenich, 1996; Gray, 1994; Murthy and Fetz, 1992) who have correlated firing synchrony during rapid gamma oscillations with measures of perceptual coherence. Second, Thorpe et al. (1996; Celebrini et al., 1993) and others (Tovée et al., 1993, 1994) have shown that visual scenes can be processed at a rapid rate inconsistent with the temporal requirements of attractor networks or of the rate code integration of slow potential theory. Spike timing is also clearly used in the auditory system (Joris et al. 1994).

Most computer models of temporal coding have been developed utilizing an underlying oscillatory carrier wave (Fukai, 1996; Lisman and Idiart, 1995; Fransen and Lansner, 1995; Lansner and Fransen, 1992, 1995; Menschik and Finkel in press). Other models, like the current study, stress the importance of spike arrival time (Maass, 1997; Thorpe, 1990). Oscillatory and feedforward aspects of temporal codes can, and most likely do, coexist (Parodi et al., 1996). The coordination of brain oscillation with sniffing in rodents and the role of oscillatory motor signals suggest that temporal senso-

rimotor integration could assist in coordinating signal reception with the phase requirements of the receiver to provide a fine-grained "active" vision (or other sensory modality). Stochastic resonance may also play a role in optimizing information carrying capacity (Stemmler, 1996).

4.2. Learning connectivity

The learning algorithm, embodied in the outer product calculation of Eq. 1, is Hebbian. In the context of the 0/1 vectors used in this study, this involves only the concept originally identified by Hebb– augmenting connection strength with coincident pre- and postsynaptic activity. With the use of -1/1 vectors, this would also involve the reasonable corollary of this rule– decrementing when activity is paired with inactivity– as well as the more questionable corollary: incrementing with paired inactivity. Since a feedforward network has no internal connections, Hebbian learning does not require suppression of associative synapses with iterative learning as it does in a recurrent model (Barkai et al., 1994; Lytton, 1997).

The 0/1 algorithm required calculation of a scaling factor given by $\langle jk \rangle / f$ in Eq. (10), with $\langle jk \rangle$ calculated from Eq. (6). The complexity of this expression would make it impossible to calculate directly using local neural mechanisms. However, as pointed out above, use of a timing code allows parameter flexibility that eliminates the need to precisely calculate this factor, which determines the relative strength of excitatory and inhibitory input to the network. Inhibitory synaptic strength could be modulated globally by a neuromodulator or diffusible substance in order to produce a balance of excitation and inhibition where firing occurred in nearly all cells in response to input.

Although the scaling brought about by $\langle jk \rangle / f$ can be approximated, the overall inhibitory pattern determined by $\sum_\mu \hat{o}_\mu$ of Eq. (10) is required for the functioning of the network. The creation of this feedforward inhibitory weight vector suggests the possibility of a novel learning algorithm that would teach feedforward inhibition based on the average strength of feedback signals integrated over pattern presentation time. Two possible algorithms suggest themselves– one requiring feedforward and feedback inhibition mediated by the same interneurons and the other postulating retrograde learning via a diffusible messenger.

In feedback teaching, the excitatory output of the network would instruct the feedforward inhibition strengths over the presentation of all input patterns. By using the same inhibitory interneurons for both the feedforward and feedback inhibition associated with a particular excitatory cell, a Hebbian mechanism could effect a strength increase in the shared inhibitory synapse onto the excitatory cell. During learning, feedback activation would provide the coincident presynaptic (interneuron) and postsynaptic (excitatory neuron) activity required. With learning completed, the same interneurons would provide appropriately scaled feedforward inhibition during the recall phase. The feedback inhibition would not itself be needed during the recall phase and might be deactivated by modulatory processes. If not deactivated, the feedback inhibition might improve the speed of the network by permitting stronger excitatory inputs to be used. Extension of the feedback inhibitory matrix to include recurrent cross-connections could actually improve the function of the network by allowing early-firing, set-bit neurons to further inhibit late-firing, clear-bit neurons. These hypotheses remain to be tested.

An alternative method for inhibitory learning in this network would be to have the output of the principal neurons directly modulate the strength of inhibitory connections via a retrograde transmitter or diffusible signal. In this case, activity in the principal neuron would increase the strength of inhibitory synapses onto that neuron. Because of segregation between feedforward and feedback inhibition and possible involvement of different subtypes of $GABA_A$ receptors, this learning could be selective (Kapur et al., 1997b).

These learning algorithms are experimentally testable in slice. The retrograde-signal learning algorithm suggests that repeated activation of a principal neuron by current injection would lead to strengthening of feedforward inhibitory strength. This could be tested using intracellular driving as the teaching signal and extracellular stimulation to test the strength of inhibitory input. The implication is that highly active neurons directly increase their own inhibitory load. Previous data, although utilizing sustained depolarization rather than spiking activity, does not support this (Alger et al., 1996; Pitler and Alger, 1994).

The feedback-inhibition teaching algorithm postulates a subset of interneurons that are active in both feedforward and feedback mode. Such cells would be reciprocally connected with the neighboring principal neuron and could also be directly activated by extracel-

lular stimulation of feedforward pathways. This could be demonstrated by finding evidence of stereotyped timing of paired IPSPs that could both be blocked with focal bicuculline application at a single location (Kapur et al., 1997a). This type of circuitry has been previously described in *Aplysia* (Blazis et al., 1993; Fischer and Carew, 1993).

4.3. Application of the Model to the Hippocampus

Although this is a highly abstract model that was not directly based on particular neurophysiological data, this work was motivated by considerations of hippocampal function. As such, it is useful to consider how well the model fits what is know about the hippocampus and to speculate as to what implications this model might have for hippocampal function.

As noted above, the feedforward system presented here transduces a vector pattern into another vector pattern, the essence of a heteroassociative memory. In this way, it can be used to instantiate Teyler and DiScenna's "memory indexing theory" (Teyler and DiScenna, 1986). Under this theory, the hippocampus produces an index of back-pointers when a memory is stored in cortex. The hippocampus then serves both to recognize a memory and to recall it by reactivating the previously indexed locations in neocortex. This idea is similar to that proposed by Damasio in that recall involves distributed reactivation of multiple sites throughout cortex (Damasio, 1990). In the context of the parallel convergent model of Figure 3B, the inputs to the feedforward network would represent sensory cues reminiscent of one or more episodic or semantic memories. The network would produce a distinct output vector that, through back-projection from entorhinal cortex to neocortex, would activate selected cortical areas so as to evoke appropriate memories. Information from sensory input, processed in cortical areas, would be relayed via entorhinal cortex, around the trisynaptic pathway and subiculum back to the entorhinal cortex, for eventual transmission to select appropriate associations in neocortex. The feedforward network presented here is a model of only a single set of synapses in this pathway.

Each of the three synapses of the trisynaptic pathway represent feedforward connections that would be candidate areas to be explained by this model. Of these, the dentate gyrus granule cells seem to have the lowest firing rates and therefore might most closely fit the ideal of single-spike signaling addressed in this article (Lytton et al., 1998). Hence, the mossy fiber projection to CA3 would be a candidate for a memory transduction mechanism of the type shown here. The presynaptic dentate granule cells fire rarely and powerfully activate CA3 pyramidal cells. Continued firing of CA3 pyramidal cells following activation could represent their involvement in an attractor network of the Hopfield type, as previously proposed (O'Reilly and McClelland, 1994; Treves and Rolls, 1994). This would complement the feedforward network by providing pattern completion of degraded outputs.

Other areas with more rapidly firing neurons could also be utilized in a feedforward information matrix. As Hopfield has pointed out, the timing signal in a regularly firing cell could be signaled by initial coordinated firing of the presynaptic inputs, which represent the leading edge of a wave of activation (Hopfield, 1995). This would permit rapid parallel information processing that could complement a more thorough, but slower, attractor processing occurring subsequently.

4.4. Marr's Three Levels Revisited

David Marr's model of hippocampal function was an early effort at applying the notion of an associative memory to a brain structure. Marr's modeling was based on a top-down notion of how the brain could be understood. He stated that there were three levels of investigation– the computational problem, the algorithm, and the implementation (Marr, 1982). He believed that it was always necessary to start at the most abstract level, that of determining what kind of problem a particular brain area has to solve, and then work one's way slowly down to the level of implementation, understanding the actual neurons that do things in the brain. This approach has fallen into disfavor as increased knowledge about the brain has suggested that it does not always solve problems the way an engineer would (Van Essen and Maunsell, 1983).

By contrast, some have suggested that implementation alone can explain everything and that an understanding of function will flow from this bottom-up perspective (Hille, 1996). Instead, it appears likely that neither function nor implementation can be explained in isolation: the brain and its activities must both be taken into account in developing useful models of brain functioning. By working within the limi-

tations of quasi-realistic neuronal elements, we allow the neurons to show what they can and cannot do.

Acknowledgments

I wish to thank Peter Lipton for many helpful discussions, Dan Uhlrich and Karen Manning for assistance with statistical analysis, and Lew Haberly for reading the manuscript. This research was supported by the Office of Research and Development, Medical Research Service of the Department of Veterans Affairs, and by the National Institute of Neurological Disease and Stroke.

References

Abbott LF, Kepler TB (1990) Model neurons: from Hodgkin-Huxley to Hopfield. Paper presented at the Eleventh Sitges Conference on Neural Networks, Sitges, Spain.

Alger BE, Pitler TA, Wagner JJ, Martin LA, Morishita W, Kirov SA, Lenz RA (1996) Retrograde signalling in depolarization-induced suppression of inhibition in rat hippocampal ca1 cells. *J Physiol (Lond)* 496: 197–209.

Amaral DG, Ishizuka N, Claiborne B (1990) Neurons, numbers and the hippocampal network. In: Progress in Brain Research. Elsevier Science Publishers.

Amit D (1989) Modelling Brain Function. Cambridge University Press, Cambridge.

Anderson JA (1972) A simple neural network generating an interactive memory. *Math. biosci.* 14: 197–220.

Anderson JA, Rosenfeld E (1988) Neurocomputing: Foundations of research. MIT Press, Cambridge, MA.

Barkai E, Bergman RE, Horwitz G, Hasselmo ME (1994) Modulation of associative memory function in a biophysical simulation of rat piriform cortex. *J. Neurophysiol.* 72: 659–677.

Blazis DEJ, Fischer TM, Carew TJ (1993) A neural network model of inhibitory information processing in *aplysia*. *Neural Computation* 5: 213–227.

Borg-Graham LJ (1991) Modeling the non-linear conductances of excitable membranes. In: Wheal H, Chad J, eds. Cellular and Molecular Neurobiology: A Practical Approach, Oxford, New York, pp. 247–275.

Buckmaster PS, Schwartzkroin PA (1995) Interneurons and inhibition in the dentate gyrus in vivo. *Neurosci* 15: 774–789.

Bullock TH (1993) Integrative systems research on the brain: resurgence and new opportunities. *Annual Review of Neuroscience* 16: 1–15.

Bullock TH (1997) Signals and signs in the nervous system: the dynamic anatomy of electrical activity is probably information-rich. *Proceedings of the National Academy of Sciences of the United States of America* 94: 1–6.

Celebrini S, Thorpe S, Trotter Y, Imbert M (1993) Dynamics of orientation coding in area V1 of the awake primate. *Visual Neuroscience* 10: 811–825.

Damasio AR (1990) Category-related recognition defects as a clue to the neural substrates of knowledge. *Trends Neurosci* 13: 95–98.

deCharms RC, Merzenich MM (1996) Primary cortical representation of sounds by the coordination of action-potential timing. *Nature* 381: 610–613.

Destexhe A, Mainen ZF, Sejnowski TJ (1994) Synthesis of models for excitable membranes, synaptic transmission and neuromodulation using a common kinetic formalism. *J. Comput. Neurosci.* 1: 195–230.

Destexhe A, Mainen ZF, Sejnowski TJ (1994a) An efficient method for computing synaptic conductances based on a kinetic model of receptor binding. *Neural Comput.* 6: 14–18.

Fischer TM, Carew TJ (1993) Activity-dependent potentiation of recurrent inhibition: a mechanism for dynamic gain control in the siphon withdrawal reflex of aplysia.. *J. Neurosci.* 13: 1302–1314.

Fransen E, Lansner A (1995) Low spiking rates in a population of mutually exciting pyramidal cells. *Network* 6: 271–288. 2660.

Fricke RA, Prince DA (1984) Electrophysiology of dentate gyrus granule cells. *J. Neurophysiol.* 51: 195–209.

Fukai T (1996) Competition in the temporal domain among neural activities phase-locked to subthreshold oscillations. *Biol. Cybernetics* 75: 453–461.

Gardner-Medwin AR (1976) The recall of events through the learning of associations between their parts. *Proc. R. Soc. Lond. B* 194: 375–402.

Gray CM (1994) Synchronous oscillations in neuronal systems: Mechanisms and functions. *J. of Comput. Neurosci.* 1: 11–38.

Gray CM, Konig P, Engel AK, Singer W (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338: 334–337.

Gray C, Singer W (1989) Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl. Acad. Sci. USA* 86: 1698–1702.

Hille B (1996) A K^+ channel worthy of attention. *Science* 273: 1677.

Hines M (1993) NEURON – A program for simulation of nerve equations. In: Eeckman, F, ed. Neural systems: Analysis and modeling, Kluwer Academic Publishers, Boston, MA, pp. 127–136.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Nat. Acad. Sci. USA* 79: 2554–2558.

Hopfield JJ (1984) Neurons with graded response have collective computation abilities. *Proc. Nat. Acad. Sci. USA* 81: 3088–3092.

Hopfield JJ (1995) Pattern recognition computation using action potential timing for stimulus representation. *Nature* 376: 33–36.

Joris PX, Carney LH, Smith PH, Yin TC (1994) Enhancement of neural synchronization in the anteroventral cochlear nucleus I responses to tones at the characteristic frequency. *J. of Neurophysiol.* 71: 1022–1036.

Kapur A, Lytton WW, Ketchum K, L Haberly (1997b) Regulation of the NMDA component of EPSPs by different components of postsynaptic GABAergic inhibition: A computer simulation analysis in piriform cortex. *J. Neurophysiol.* 78: 2546–2559.

Kapur A, Pearce R, Lytton WW, L Haberly (1997a) $GABA_A$-mediated IPSCs in piriform cortex have fast and slow components with different properties and locations on pyramidal cells: Study with physiological and modeling methods. *J. Neurophysiol.* 78: 2531–2545.

Kohonen T (1972) Correlation matrix memories. *IEEE Transactions on Computers* C-21: 353–359.

Lansner A, Fransen E (1992) Modelling hebbian cell assemblies comprised of cortical neurons. *Network* 3: 105–119. 2659.

Lansner A, Fransen E (1995) Improving the realism of attractor models by using cortical columns as functional units. In: J. Bower, ed. The Neurobiology of Computation: Proceedings of the Third Annual Computation and Neural Systems Conference, Kluwer Academic Publishers, Boston, MA.

Lisman JE, Idiart MA (1995) Storage of 7 ± 2 short-term memories in oscillatory subcycles. *Science* 267: 1512–1515.

Lytton WW (1996) Optimizing synaptic conductance calculation for network simulations. *Neural Comput.* 8: 501–510.

Lytton WW (1997) Brain organization: from molecules to parallel processing. In: Trimble, M and Cummings, J, eds. Contemporary Behavioral Neurology, Butterworth–Heinemann, Newton, MA, pp. 5–28.

Lytton WW, Hellman KM, Sutula TP (1998) Computer models of hippocampal circuit changes of the kindling model of epilepsy. *Artificial Intelligence in Medicine*.

Lytton WW, Sejnowski TJ (1991) Inhibitory interneurons may help synchronize oscillations in cortical pyramidal neurons. *J. Neurophysiol.* 66: 1059–1079.

Maass W (1997) Fast sigmoidal networks via spiking neurons. *Neural Comput.* 9: 279–304.

Marr D (1971) Simple memory a theory for archicortex. *Phil. Trans. R. Soc. Lond. B* 262: 23–81.

Marr D. (1982) Vision. Freeman, San Francisco, CA.

McCulloch W, Pitts W (1943) A logical calculus of ideas immanent in nervous activity. *Bull. Math. Biophysics* 5: 115–133.

Menschik ED, Finkel LH (in press) Neuromodulatory control of hippocampal function: Towards a model of Alzheimer's disease. *Artificial Intelligence in Medicine*.

Murthy V, Fetz E (1992) Coherent 25- to 35-Hz oscillations in the sensorimotor cortex of awake behaving monkeys. *Proc. Natl. Acad. Sci. USA* 89: 5670–5674.

O'Reilly RC, McClelland JL (1994) Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus* 4: 661–682.

Parodi O, Combe P, Ducom JC (1996) Temporal coding in vision: Coding by the spike arrival times leads to oscillations in the case of moving targets. *Biol. Cybernetics* 74: 497–509.

Pitler TA, Alger BE (1994) Depolarization-induced suppression of gabaergic inhibition in rat. *Neuron* 13: 1447–1455.

Pribram KH (1969) The limbic systems, efferent control of neural inhibition and behavior. *Scientific Am.* 220: 73–86.

Scharfman HE (1994) Evidence from simultaneous intracellular recordings in rat hippocampal slices that area CA3 pyramidal cells innervate dentate hilar mossy cells. *J. Neurophysiol.* 2: 2167–2180.

Stemmler M (1996) A single spike suffices - the simplest form of stochastic resonance in model neurons. *Network-Comput. in Neural Sys.* 7: 687–716. 2655.

Stevens CF (1966) Neurophysiology: A Primer. Wiley, New York.

Teyler TJ, DiScenna P (1986) The hippocampal memory indexing theory. *Behavioral Neurosci.* 100: 147–154.

Thorpe S (1990) Spike arrival times: a highly efficient coding scheme for neural networks. In: Eckmiller, R, Hartman, G, and Hauske, G, eds. Parallel processing in neural systems, Elsevier, Amsterdam, The Netherlands, pp. 91–94.

Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. *Nature* 381: 520–522.

Tovée MJ (1994) Neuronal processing: How fast is the speed of thought?. *Current Biol.* 4: 1125–1127.

Tovee MJ, Rolls ET, Treves A, Bellis RP (1993) Information encoding and the responses of single neurons in the primate temporal visual cortex. *J. of Neurophysiol.* 70: 640–654.

Treves A, Rolls ET (1991) What determines the capacity of associative memories in the brain. *Network: Comput. and Neural Sys.* 2: 371–397.

Treves A, Rolls ET (1994) Computational analysis of the role of the hippocampus in memory. *Hippocampus* **4**: 374–391.

Van Essen DC, Maunsell JHR (1983) Hierarchical organization and functional streams in the visual cortex. *Trends Neurosci.* **6**: 370–375.

Willshaw DJ, Buckingham JT (1990) An assessment of Marr's theory of the hippocampus as a temporary memory store. *Phil. Trans. R. Soc. Lond. B* **329**: 205–215.

Willshaw DJ, Buneman OP, Longuet-Higgins HC (1969) Non-holographic associative memory. *Nature* **222**: 960–962.