A NEURAL NETWORK MODEL OF MEMORY AND HIGHER COGNITIVE FUNCTIONS

David D. Vogel, Ph.D.

Ross University
Portsmouth
Commonwealth of Dominica

Correspondence:

David Vogel, Ph.D.
Ross University
Portsmouth
Commonwealth of Dominica

E-mail: dvogel@rossmed.edu.dm

Telephone: 1-767-445-5355 Ext. 287

Fax: 1-767-445-3457

A NEURAL NETWORK MODEL OF MEMORY AND HIGHER COGNITIVE FUNCTIONS

ABSTRACT

I first describe a neural network model of a small region of the brain that produces associative memory. The model depends, unconventionally, on disinhibition of inhibitory links between excitatory neurons rather than long-term potentiation of excitatory projections. The model may be shown to have advantages over traditional models both in terms information storage capacity and biological plausibility. The simple learning and recall algorithms are independent of network architecture, and require no thresholds, finely graded synaptic strengths, or implausible stochastic smoothness. Several identical copies of this local network are then connected by means of massive, weak, reciprocal, excitatory projections that allow one region to appropriately control the recall of information in another. The result is a simple network with properties that seem analogous to serial memory, classical and operant conditioning, secondary reinforcement, refabrication of memory, and fabrication of possible future events. The network distinguishes between real and imagined events, and can predicate its response on the absence as well as the presence of stimuli. Some of these properties are achieved in ways that seem to provide instances of self-awareness and imagination, suggesting that consciousness may emerge as an epiphenomenon in simple brains.

Key words

cognition; conditioning; memory; neural network; P-net; refabrication; R-net; sparse

A NEURAL NETWORK MODEL OF MEMORY AND HIGHER COGNITIVE FUNCTIONS

1. INTRODUCTION

Prima facie, trying to model the cerebrum at the neural network level is a fool's game. The architecture and cellular function are only partially known. Moreover, any evolved control system is not only likely to be complex, it is almost certain to be absurdly complex. Any model that represents such systems with methodological economy is unlikely to represent them accurately.

On the other hand, there are sufficient reasons for the attempt. Even models that are in fundamental respects inconsistent with the known cerebrum have provided insights into mechanisms that may contribute to actual cortical functioning. Marr's (1971) model of the hippocampus has influenced research and our beliefs concerning that structure for 3 decades.

Moreover, it is extremely difficult for people to imagine how the workings of a machine can explain the existence of a mind. The allegedly transcendent mind is a correspondingly powerful object of the sort of mystification and special pleading that has long been inimical to scientific inquiry. Writing in the *Encyclopedia of Ignorance* shortly after the introduction of Marr's model, Gregory (1977) observed that there are many "gaps" in science, but the "mind-brain gap" seems, somehow, much greater than all the others. Any intellectually satisfying progress toward filling this gap serves the interests of reason.

When I refer to the network described here as "biologically plausible," I mean the following: (1) The model contains simplifications. However, it seems likely that one could substitute, for example, more realistic neurons and architecture for the binary neurons and random projections in this model, and the network would still work. As a contrasting example, when a model employs a fully connected network, it may not seem likely that simulations will still work after substitution of a sparsely connected architecture resembling a real brain. (2) The model also includes a few extensions of the known properties of real brains. For example, the mechanism of engram formation used here may not be a simplification of the mechanism used in real brains. However, it is consistent with what is known about real brains and may be regarded as a prediction of the model. In contrast, when a model uses inhibitory synapses to create a perfectly uniform level of inhibition in all neurons in a network, this extension is not consistent with the known properties of inhibitory synapses. Using "plausible" in this sense, I find the model described here to be unusually plausible.

"Plausibility" in this sense is not the only matter of interest in models of brains. Some models retain interest even though they are "neural networks" much more in the engineering than the biological sense of the phrase. Examples would include the theory of simple proportional analogies of Jani and Levine (2000), and the model of exploratory behavior in the face of anxiety of Salum et al. (2000). Many other networks that model such things as serial memory (Taylor & Taylor, 2000**),** pharmacological effects (Hasselmo & Schnell, 1994),

or psychopathology (Greenstein-Messica & Ruppin, 1998) use large scale block designs that are analogous to structures in the brain but do not specify biologically realistic connection matrices in local regions. Yoshida et al. (2002) provide interesting insights into aspects of recall without providing a mechanism of engram formation. Chechnik et al. (2001) provide interesting insight into the possible role of allocation by single neurons of limited resources among many synapses, though their discussion is conducted in terms of densely connected networks. Models whose local architectures are more representative of the real cerebrum are likely to be restricted to special aspects of lower level problems. Examples would include work on sequence generation (Fukai, 1999) and rapid sensory perception in neocortex (Körner et al., 1999). Such models are instructive and incorporate features that will likely prove useful in some as yet unrealized synthesis.

Many attempts at representing significant aspects of the cerebrum with realistic matrix architectures have involved models of associative memory derived from Marr's (1971) model of archicortex. These models depend on strengthening of excitatory synapses in a way that is thought of as analogous to long-term potentiation (LTP), and are subsequently referred to as "LTP models." Examples include models described by Eichenbaum and Buckingham (1990), Gibson, and Robinson (1992), and Treves and Rolls (1994).

This paper details a plausible neural network that is both simple and able to account for a variety of psychological phenomena including serial memory, classical and operant conditioning, backward conditioning, secondary reinforcement, predictive fabrication, and refabrication of incomplete memories. Random excitation of the network allows it to discover sensory consequences of motor activity and to subsequently repeat a motor activity by recalling the sensory consequence. The network distinguishes between imagination and reality, and is able to predicate behaviors, not only on the presence of stimuli, but on their absence. Some of these properties are achieved in ways that seem to provide instances of self-awareness and imagination and suggest that consciousness may emerge as an epiphenomenon in relatively simple brains.

Section 2 describes a partial model of a small region of the brain that has been less fully developed in previous papers (Vogel, 1998; 2001). The small model describes a region that might amount to a few square millimeters of archicortex (or other structure) and provides a biologically plausible account of associative memory. The model does not support certain presuppositions about information storage. Memory, for example, does not depend on LTP. Rather, memory depends on the weakening of inhibitory links (disinhibition). Even though only a small fraction of cortical neurons are inhibitory, these models have greater storage capacity than realistic LTP models. These networks have previously been referred to as "R-nets" to distinguish their random architectures from those of mathematically more tractable "P-nets" whose connection matrices form projective planes.

Section 3 describes a network consisting of 2 R-nets joined together by means of large numbers of relatively weak, reciprocal excitatory projections. The resulting model of serial memory seems more satisfying than those I have described previously (Vogel, 1995, 2001).

Finally, Section 4 describes a possible role for excitatory connections and LTP in gating the flow of information from one region to another by modeling a conditioned response

in a more complex network (a “C-net”) consisting of several R-nets, each with the same architecture and information processing rules, but variously regarded as sensory, motor, or integrative on the basis of their positions in the network. A similarly constructed “reinforcement” region is also present that responds to a small subset of “sensory” inputs by initiating synaptic training and inhibiting activity in certain other local regions.

This model exhibits enough complexity to represent a miniature mind. It does not incorporate a number of conspicuous features of neocortex including its layered structure and cortical columns. These features are not needed in the present paper though work that incorporates these structures into models of such phenomena as rapid sensory perception (Körner et al. 1999) or spatial working memory (Tanaka 1999) develops important, complementary ideas about neocortex.

As with all sparsely connected neural networks, information stored in these models must be sparsely coded (i.e., a “sensory” event must be represented by the firing of a small fraction of the neurons in the network). In real brains, information at the level of primary sensory neurons is densely coded. The process of feature extraction presumably generates sparse coding in “higher” regions of real brains. All of the networks described, here, use information that has already been sparsely coded. Near relatives of these networks are able to convert dense to sparse coding, and I do not imagine that real brains make an exclusive distinction between mnemonic and encoding processes. However, a discussion of network designs that optimize encoding as opposed to mnemonic processes is not required for this paper.

Particular regions of the network described, here, are deliberately not identified with particular regions of real brains. The learning and recall algorithms used are suitable for many different architectures in the brain, and many parts of the brain might be used to achieve the functions modeled here. As implemented, the architecture looks most like archicortex. However, it seems likely that evolution has conserved mechanisms, not only between archi- and neocortex, but between brainstem and cortex. The only architectural requirement of the model is that sufficiently many pairs of excitatory neurons be linked by inhibitory neurons.

Details of the algorithms that are scattered through the main text are organized in the Appendix.

2. A MODEL OF A LOCAL REGION

2.1 Traditional Neural Networks

In models that follow the work of Marr (1971), excitatory neurons, analogous to principal cells, are randomly connected to one another as in Fig. 1. The artificial neural networks are sparsely connected with a probability of one neuron projecting onto another comparable to that of subregion CA3 of the hippocampus.

In operation, a subset of the neurons in the network is activated. I will refer to this subset as a “training set” (though it would more commonly be called a “training vector”).

Synapses between active neurons are then trained according to a "Hebbian" learning rule which specifies that the strengths of excitatory synapses are increased when both the pre- and postsynaptic neurons are active at the same time. After training on some number of sets, a subset of a training set (a "recall set") may be activated with the objective of reactivating the original training set (the "target set"). Because of previous synaptic training, members of a target set are likely to receive more excitatory activation from a recall set than non-target neurons. An activation threshold for firing may be computed stochastically such that most target neurons are above the threshold and fire while most non-target neurons are below the threshold and do not. As the number of stored training sets increases, the number of recall errors also increases. Errors may consist of active neurons that are not members of the target set ("spurious neurons") or inactive members of the target set ("missing neurons").

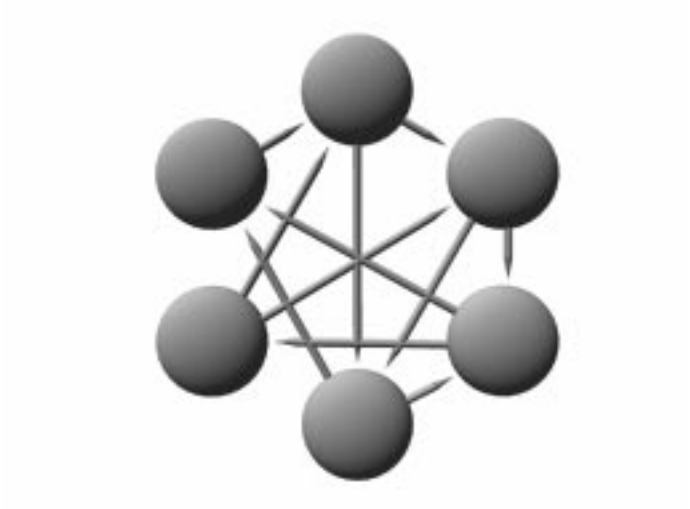

Fig. 1. A representation of a standard LTP network in which excitatory neurons project onto a small fraction of the other excitatory neurons in the network.

The required threshold depends on the size of the recall set. As the size of the recall set increases, the amount of activation of all neurons increases, and the threshold that separates target from non-target neurons increases. These LTP networks are made more robust with respect to recall set size by positing a uniform, inhibitory feedback onto all neurons. Increasing this inhibition is equivalent to increasing the threshold, and the inhibition is made proportional to the number of active neurons in a recall set (or the number of active neurons on the previous cycle of firing if the network is allowed to run through more than one cycle).

These traditional networks provide some insight into principles by which real brains may store information. However, they have several features that are biologically implausible.

First, the connectivity of the real cortex is at the lower edge of the connectivity required for successful information storage by these neural networks. Even with the assumption of perfectly uniform inhibitory feedback, the networks are brittle with respect to training and recall set sizes. The relatively sophisticated network of Bennett, Gibson, & Robinson (1994) does not effectively store training sets much larger than 330 neurons because the number of synapses trained increases with the square of the number of neurons per training set. It does not store training sets much smaller than 330 neurons because recall is not stochastically smooth.

Second, the introduction of stochasticity in the inhibitory feedback destroys the already tenuous storage capacity of the networks.

Third, the mechanism of uniform inhibitory feedback is not made explicit, and no such mechanism seems to exist in the real brain. Single active inhibitory synapses appear to be extremely effective at silencing postsynaptic neurons (Buckmaster & Schwartzkroin, 1995; Conors, Malenka, & Silva, 1988; Miles et al., 1996). No finely graduated inhibitory cloud is present.

Fourth, synaptic strengths are finely graduated, must be closely regulated, and are dependent on architectural parameters such as connectivity. This seems to make traditional networks evolutionarily implausible because any change in architecture deoptimizes synaptic strengths and thresholds.

Fifth, the brain is not randomly connected, as are the networks in these models. While non-random connection matrices exist that are stochastically as smooth as random matrices (Vogel & Boos, 1997), they are relatively rare objects that are not likely to be discovered by evolutionary means.

Sixth, the analysis of the storage capacities of large versions of these networks is suspect. The problem arises in the use of DeMoivre's Theorem to approximate a binomial distribution by means of a normal distribution (Marr, 1971). For small networks, equations for the number of errors during recall give results that closely approximate the actual number of errors found in simulations. It seems natural to assume that as network size increases the normal approximation of a binomial distribution will get better and the equations more accurate. However, it is not the network size that is the number of trials in the binomial distribution, it is the recall set size, and the recall set size typically remains constant in moving from simulations of small networks to analyses of large networks. It sometimes appears (Gibson & Robinson, 1992) that in simulations of small networks, the number of errors is slightly larger than analysis predicts. If the number of neurons in a recall set remains constant, the number of spurious neurons on the first cycle of operation may be expected to increase linearly with network size, and a small under-estimate of the number of errors in a small network becomes large in a large network and the signal is overwhelmed by noise.

2.2 A Qualitative Description of the Local Network Model (R-net)

In the local network model, direct projections between excitatory neurons are ignored. (Their function will be explored in subsequent sections.) Rather, R-nets have only random projections of excitatory neurons (analogous to principal cells) onto a relatively small number of inhibitory neurons (analogous to interneurons), and recurrent projections from the inhibitory neurons onto excitatory neurons (Fig. 2). Pairs of primary neurons are then said to be "linked" by inhibitory pathways of length 2. Of the various kinds of interneurons present in the cortex, the basket cells are most common and bear the greatest resemblance to the inhibitory neurons of the model.

In the simplest version of these networks, coactivation of two excitatory neurons causes both synapses in any link between the neurons to become "trained." The synapses are binary, being trained or not. Training of both synapses in an inhibitory link functionally "cuts" the link. Unless both synapses are trained, inhibitory links are absolutely inhibitory, excitatory neurons firing on the *n*th cycle if and only if they receive no inhibition from neurons firing on

the (*n*-1)th cycle.

This cutting may be modeled in either of two ways. The obvious mechanism is simply to weaken both synapses so that the pathway is no longer inhibitory. An interesting alternative (discussed below) is to strengthen both synapses, but to note that inhibitory $GABA_A$ synapses are excitatory when sufficient γ–aminobutyric acid (GABA) is released. When this alternative is employed, active inhibitory links are inhibitory unless (1) the projection of the inhibitory synapse onto the excitatory neuron is trained, and (2) the total excitation of the inhibitory neuron is at least as great as that produced by a single trained synapse (which may be accomplished, erroneously, by the firing of sufficiently many untrained synapses). Thus, in this simple version, excitatory neurons are binary, firing if they are not inhibited. Inhibitory neurons have activities equal to the sum of their inputs. Under the first interpretation of "cutting" a link, the firing rates of inhibitory neurons in trained links are low as are the strengths of their inhibitory projections. Under the second interpretation, the firing rates of inhibitory neurons are high as are the strengths of their projections.

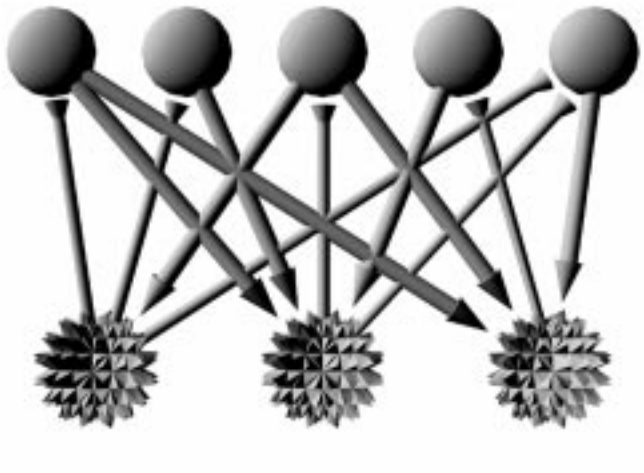

Fig. 2. A representation of an R-net. Smooth neurons are excitatory. Spiky neurons are inhibitory. The ratios of excitatory to inhibitory neurons and synapses shown are not the same as those in the model. A P-net might be similarly illustrated; however, there would be equal numbers of excitatory and inhibitory neurons, and projections would be reciprocal and arranged so that every pair of excitatory neurons would be liked through one and only one inhibitory neuron.

It is important to note that no thresholds are required under either interpretation. The only architectural requirement is that sufficiently many pairs of excitatory neurons are linked by inhibitory neurons. While the architectures described here most closely resemble archicortex, the training and recall algorithms may be applicable to many non-cortical and neocortical structures.

Employed as local regions in larger, more complex, and noisy C-nets, these local R-nets have additional properties. The simplest networks can be troubled by noise. A single spurious neuron can turn off a substantial part of a target set. To accommodate a noisy environment, neurons are allowed to accumulate inhibition which decays linearly with time. If a few neurons of a training set are active along with several spurious neurons, the spurious neurons will typically acquire more inhibition than target neurons. As the inhibition of all neurons decays, target neurons are likely to fire before non-target neurons and inhibit non-target neurons. In addition, in simulations of recall of the smallest training sets (20-neuron

sets) in R-nets, trained inhibitory links have been made slightly excitatory. This improves storage somewhat by reducing the probability of a single spurious neuron silencing a large enough fraction of target neurons to prevent recall.

When local R-nets are joined to form C-nets, neurons are also made to show burst firing, being active on every cycle for 100 cycles and then firing once every 20 cycles. In models of serial memory, this slow firing can maintain local regions in more or less stable states for prolonged periods of time, so that the interval between presentation of new members of a series can vary, though other local R-nets may intervene and control the activity of the local R-net during times when the firing rate is low.

I am unable to provide a formal analysis of the R-nets employed, here, without resorting to the same improper use of DeMoivre's Theorem described above. However, analyses of several related networks are instructive. Vogel and Boos (1997) analyzed networks (P-nets) with equal numbers of excitatory and inhibitory neurons and projection matrices that form projective planes. Such an architecture is extremely sparse and links each pair of excitatory neurons through one and only one inhibitory neuron. (As with the R-nets illustrated in Fig. 2, many pairs are linked through the same inhibitory neuron.) A lower bound was derived for the information storage capacities of P-nets for recall in one cycle of operation (i.e., after updating the network one time following activation of a recall set). Simulations demonstrated that small versions of these networks stored almost twice the information given by the lower bound (Fig. 3). In published work, these P-nets actually used LTP with thresholds. However, the analysis of P-nets that employ disinhibition is identical to that of P-nets that employ LTP. In either case, the analysis depends on the non-random nature of the connection matrix and the fact that a particular neuron fires during recall if and only if every pathway between neurons in the recall set and the particular neuron has been trained. The information storage capacity of P-nets is greatest for very small training sets, but as seen in Fig. 4, significant storage capacity is still present for larger training sets.

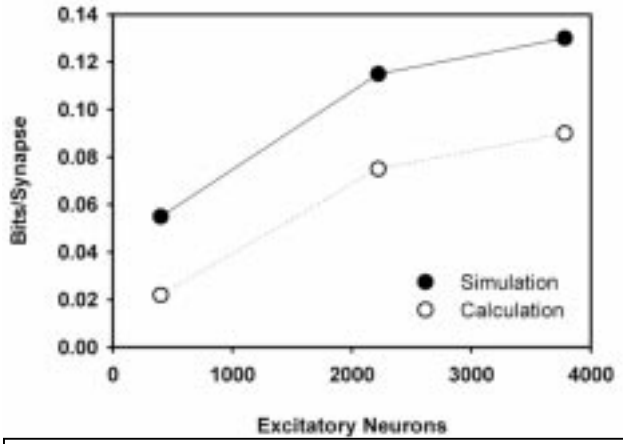

Fig. 3. Information stored per synapse as a function of P-net size with 20 neurons per training set, 10 neurons per recall set, and 1 cycle of recall. A comparison is shown of lower bound calculations and simulations. Each network has equal numbers of excitatory and inhibitory neurons. The number of synapses per neuron is approximately the square root of the number of neurons. Shannon's definition of information is used.

Subsequent calculations indicate that similar networks (R-nets) with random, but symmetrical, connection matrices will scale up from small to large networks at least as well as P-nets, allowing simulations of small R-nets to

indicate likely lower bounds for the storage capacities of large R-nets (Vogel, 1998). Finally, the argument may be extended to asymmetric R-nets (Vogel, 2001).

These arguments indicating that R-nets scale up as well as P-nets are only applicable to R-nets with at least twice as many synapses per neuron as P-nets so that nearly every pair of excitatory neurons is linked (as is the case in P-nets). The worst case networks modeled for this paper have less connectivity than this. However, the consequences of this low connectivity are easily understood, and simulation results (Subsection 2.4) make it clear that these very sparsely connected networks are suitable storage devices for mammalian sized brains.

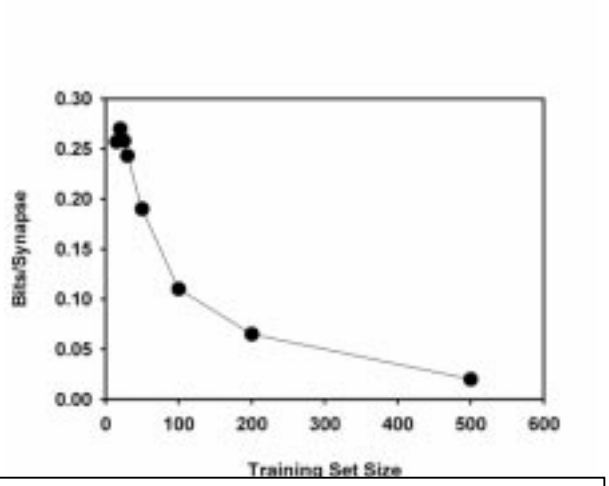

Fig. 4. Lower bound calculation of information stored per synapse as a function of training set size in a P-net with 999,001 excitatory neurons (and an equal number of inhibitory neurons) each with 1000 synapses.

The chief consequence of very sparse, random connections is that some neurons may have relatively few links to elements of a recall set, and these links may become spuriously trained by training the network on relatively few sets. The number of such neurons increases linearly with the size of the network, and very small sets are recalled poorly in large networks. In compensation, the probability of a non-target neuron having no untrained links to elements of a recall set may be seen to be the product of the probabilities for each element of the recall set. With increasing recall set size, this product gets small very rapidly, and the recall set size needed to suppress spurious neurons increases much less rapidly than the network size.

A further consequence of this poor performance when small recall sets are activated is seen when R-nets are required to recall large target sets from very small recall sets. This difficulty is largely overcome by allowing the R-net to cycle repeatedly. In most simulations reported, here, the network is allowed to pass through 100 cycles before its performance is evaluated, though an isolated R-net excited with 10 members of a 40-neuron training set will typically converge in about 20 cycles.

2.3 Biological Plausibility of the Architecture

As the histologic data is scant, it is impossible to clearly interpret the architecture of any relevant part of the brain. The architecture of the R-nets described below is based on the following, worst case interpretation of hippocampal architecture.

The projections of interneurons onto principal cells have been more extensively studied than those of principal cells onto interneurons (e.g., Buckmaster & Schwartzkroin, 1995; Gulyás et al., 1993; Sik, Tamamaki, & Freund, 1993). In a review article, Freund and

Buzsáki (1996) estimate that typical interneurons of the rat hippocampus synapse on 1000 to 3000 principal cells. Primate (Ribak, Seress, & Leranth, 1993) and especially human (Seress et al., 1993) cortex has larger numbers of inhibitory neurons than the rat, and some of these interneurons appear to have much more extensive connections. For the purposes of this discussion, the number of principal cells innervated by each interneuron is taken to be 1500. The ratio of interneurons to principal cells in human hippocampus is roughly 0.2.

Studies of the principal cell fanout onto interneurons are not common. However, Sik, Tamamaki, and Freund (1993) have described the complete axonal arbor of a single pyramidal cell of subfield CA3 of the rat hippocampus. This single example formed asymmetric (presumably excitatory) synapses on 322 parvalbumin staining neurons (presumably inhibitory interneurons). 322 is probably a considerable underestimate of the typical number of synapses on interneurons because the pyramidal cell studied had a total of only 15,295 synapses, and only synapses on parvalbumin staining cells were counted as synapses on interneurons. Perhaps as many as half the basket cells in the hippocampus do not stain with parvalbumin (DeFelipe, 1997). However, 322 is taken, here, as the number of inhibitory neurons innervated by each excitatory neuron.

Electrophysiological results of Miles (1990) indicate that about 30% of principal cell pairs have inhibitory links. If another 10% have inhibitory links that have been disinhibited, then assumptions of uniform synaptic density, random connections, and the numbers of projections described above leads to a calculation of the number of principal cells in a network of 10^6.

For this paper, all local regions (R-nets) simulated on a serial computer are constructed so that (1) the number of inhibitory neurons is 20% of the number of excitatory neurons, (2) the total number of projections onto inhibitory neurons equals the number of projections onto excitatory neurons, and (3) there are sufficient numbers of projections so that 40% of excitatory neuron pairs are connected.

It is likely that electrophysiologic studies (Miles, 1990) seriously under-estimate the connectivity of the hippocampus, and that many disinhibited or otherwise silent links exist. The hippocampus of the rat only contains about 180,000 neurons (Akdogan et al., 2002; West et al., 1991). For a subregion with 60,000 excitatory and 6,000 inhibitory neurons, the numbers of synapses described above would link 99.97% of neuron pairs. Such connectivity would greatly increase the total storage capacity of these R-nets for sets of all sizes, though it would decrease the information stored per synapse for larger set sizes (Vogel, 2001).

2.4 Biological Plausibility of the Disinhibition Algorithm

Much of what is routinely thought of as “LTP” may be attributed to disinhibition. “LTP” may be demonstrated in the Schaeffer collateral projections to subregion CA1 either by tetanizing the projections or by paired-pulse induction in which the postsynaptic neuron is depolarized during a Schaffer collateral volley. The induction of LTP by the mere tetanic stimulation of many synapses does not have any obvious computational use. However, paired-pulse induction is obviously analogous to Hebbian learning.

Long-term disinhibition in some long projection pathways between regions of the brain is well-demonstrated. Stelzer et al. (1994) find that LTP of CA1 pyramidal cells induced by tetanization of Schaeffer collaterals is accompanied by disinhibition of dendritic, but not somatic, inhibitory synapses. While LTP produced by long-term tetanization of the Schaeffer collaterals is significantly due to strengthening of excitatory synapses, the chief component of increased excitability following paired-pulse stimulation is long-term transformation of GABA synapses from inhibition to excitation (Collin et al., 1995; Nathan & Lambert, 1991). This transformation includes both $GABA_A$ and $GABA_B$ receptors (Desai et al., 1994). As pointed out by Staley et al. (1995), the firing rates (Buzsáki et al., 1992) of a majority of $GABA_A$ergic neurons in vivo appear to be sufficiently high to be excitatory. (For a mechanism of the transformation see Staley et al. (1995).)

I know of no equivalent demonstration in local connections. However, depolarization-induced suppression of inhibition (DSI) may be a related phenomenon. A 1 second depolarization of CA1 pyramidal cells (Alger et al., 1996; Morishita et al., 1997) or cerebellum Purkinje cells (Vincent et al., 1992) produces a short term (1 minute) suppression of $GABA_A$ergic inhibition. This is not a long-term effect. However, as the induction procedure does not involve both synapses in a link, it is unlike the procedure suggested by the model or employed in the studies of long projection pathways discussed above. Induction of DSI does appear to involve a retrograde messenger (Alger et al., 1996), possibly glutamate (Morishita et al., 1997), that acts on the inhibitory neuron. Such a retrograde messenger is a necessary component of any mechanism that would train both synapses in an inhibitory path of length 2, it being necessary that synapses on the inhibitory neuron be informed that the target excitatory neuron is firing. Moreover, induction of DSI is strongly facilitated by cholinergic agonists (Martin & Alger, 1999). This kind of dependence is needed to accommodate the reinforcement induced learning employed in Section 5.

The slow decay of inhibition relative to excitation in the model is consistent with the slow decay of summed inhibitory postsynaptic potentials (IPSP's) elicited by spike trains (Thomson et al., 1996).

2.5 Simulations of Simple R-nets.

Storage capacities of R-nets (Figs. 5 – 8) were evaluated by activating randomly selected sets of neurons and training synapses according to the rules given above. After some number of sets were trained, recall sets consisting of half the elements of a training set were activated, and the network was allowed 100 cycles in which to converge

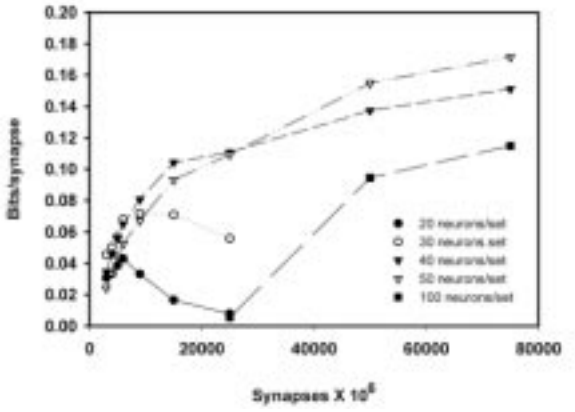

Fig. 5. The number of training sets stored as a function of the number of R-net excitatory neurons. The number of inhibitory neurons is 20% of the number of excitatory neurons, and the numbers of synapses are adjusted so that the probability of a pair of excitatory neurons being linked is 40% for all R-nets.

on the target set. For Figs. 5 through 8, the number of training sets was increased until the mean number of errors per target set (spurious and missing neurons) was 10% of the size of the target set. This criterion was chosen because it is consistent with reliable recall in the studies of serial recall that follow. As is typical of sparsely connected and coded networks, the relationship between the number of training sets and the number of errors is steeply concave upward (illustrated for a C-net in Fig. 10) so that there is not a great difference between the numbers of sets that produce 1% and 100% error rates.

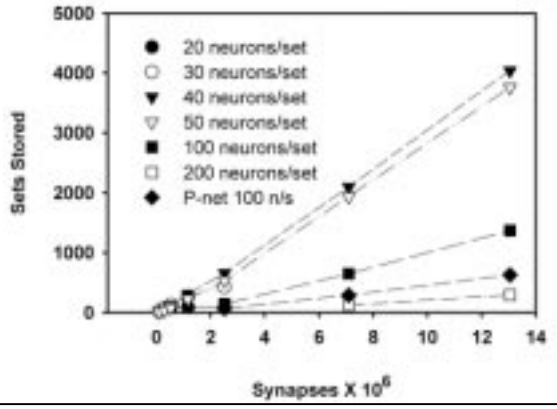

Fig. 6. The number of training sets stored as a function of the total number of R-net synapses. A lower bound calculation of the storage capacity for 100 neuron training sets in P-nets with comparably many synapses (and approximately the same total number of neurons) is also represented.

The R-nets employed in Figs. 5 through 8 range in size from 3000 excitatory neurons (each with 18 projections onto 600 inhibitory neurons each with 85 reciprocal projections) to 75,000 excitatory neurons (each with 87 projections onto 15,000 inhibitory neurons each with 435 reciprocal projections). As described previously, large networks cannot store small training sets. The 50,000 excitatory-neuron R-net was unable to store 20- or 30-neuron training sets with only 10% errors. However, the probability of a neuron being spurious depends on the product of the probabilities of each member of a recall set not having an inhibitory link to the neuron. Accordingly, as seen most clearly in Figs. 6 and 7, small increases in the sizes of training and recall sets result in large increases in the sizes of the networks that will effectively store the sets, and large networks are very broadly tuned. It may be observed that the number of 40-neuron training sets stored is almost a linear function of the number of synapses in the R-net over the entire range of R-nets studied.

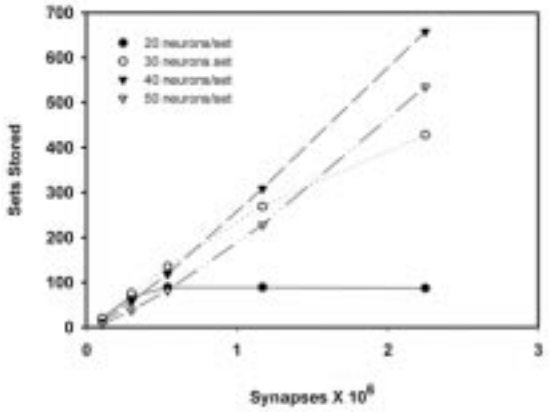

Fig. 7. An expansion of the origin of Fig. 6 meant to make clear the rapid fall off of storage for 20- and 30-neuron training sets compared to 40-neuron training sets.

Recall errors made by these R-nets may be thought of as coming from two sources. When the training sets are relatively small, errors are chiefly due to the firing of spurious neurons that have stochastically few links to the recall set. When training sets are relatively

large, the errors are chiefly due to over-training and are like those of more densely connected networks in which the relationship between storage capacities and numbers of synapses is similarly linear (Vogel & Boos, 1997; Vogel, 1998, 2001). It seems reasonable, on the grounds that the probability of a neuron being spurious is a product of probabilities, to approximate the relationship between optimal training set size and network size as a power function. Acknowledging the risks of extrapolation, such a function extrapolated to the 10^6-neuron network described previously gives an optimal training set size of roughly 100 neurons. It may be observed (Fig. 6) that the storage capacities of the simulated R-nets for relatively large training sets considerably exceeds the lower-bound calculation for P-nets with equally many synapses. If it is further presumed that the number of 100-neuron sets stored is at least as large as the lower bound estimate for a P-net with equally many synapses, the 10^6-neuron network would store more than 130,000 100-neuron sets. Whether or not this presumption is exactly correct, it seems likely that large R-nets will store large numbers of training sets of some few hundred neurons each.

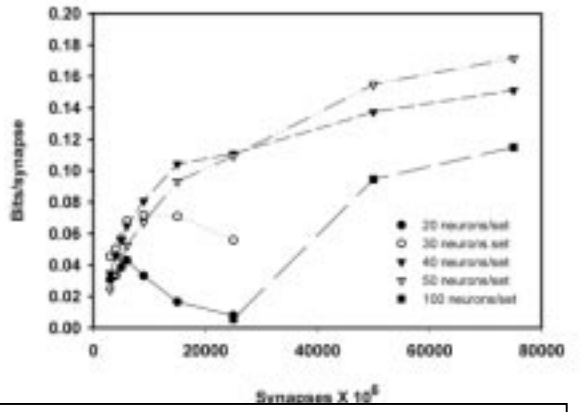

Fig. 8. Information per synapse as a function of R-net size.

R-nets with about twice as many synapses per neuron as are used for this paper (as well as P-nets) are able to store large numbers of small training sets (Vogel, 2001). However, it is worth noting that large versions of the very sparsely connected R-nets used here are not only broadly tuned, they have greater storage capacity per synapse for larger training sets than more densely connected R-nets.

The R-net with 50,000 excitatory and 10,000 inhibitory neurons is comparable in size to a subregion of the rat hippocampus (Akdogan et al., 2002), though the model is probably less densely connected than the real rat hippocampus. While a more densely connected rat hippocampus might store more than the 2000 50-neuron training sets of the simulated model, the capacity of the model seems adequate for a subregion of the rat hippocampus.

3. MULTI-REGION NETWORKS

3.1 General Considerations for a 2-Region Network.

For convenience, I refer to any network with more than one region as having sensory and motor regions. “Sensory” merely designates the R-net in which neurons are activated first. In all of the networks simulated for this section, each local R-net consists of 4000 excitatory neurons, each with 20 projections onto inhibitory neurons, and 800 inhibitory neurons, each with 100 projections onto excitatory neurons.

As in the real brain, two or more local regions may be connected by projections of the excitatory neurons of one region onto both excitatory and inhibitory neurons of another. In all of the simulations discussed below, unless otherwise specified, projections between local regions are reciprocal, and each excitatory neuron of one region has 400 projections onto excitatory neurons and 5 projections onto inhibitory neurons of the other region. The inhibitory links thus formed are governed by the rules already given for local networks. Projections onto excitatory neurons undergo LTP whenever pre- and postsynaptic neurons are active at the same time. LTP is not necessary but produces a small decrease in the number of cycles needed for elements of a training set in one region to induce the firing of elements of the training set in another region. Given that inhibitory synapses have a strength of 1, direct excitatory projections have a strength of 0.1 (untrained) or 0.4 (trained). Excitatory neurons fire if and only if the sum of their excitatory inputs is greater than or equal to the sum of their inhibitory inputs.

The large number of weak excitatory projections onto excitatory neurons produces the following result.

Consider a pair of local regions with trained sets that span both regions. If neither region has an active training set, both regions will fire neurons in sequences that appear random. These random sets of neurons are used, below, to form representations in one region of activity in another. When a training set of more than about 20 neurons is active in a region, it suppresses the random firing of neurons that occurs in the absence of an active training set.

If recall sets of different training sets are activated in each of two connected regions, each region will converge on the target set of the recall set activated in it. (In the absence of direct excitatory projections, one region wins, and both regions converge on the same set.)

If a recall set is activated in only one region, and the other region is firing at random, both regions will converge on the target set. It is this process that is made marginally faster by LTP.

3.2 Formation of Training Sets

Suppose that some activities in the motor region result in corresponding activities in the sensory region as, for example, motor activity produces corresponding proprioceptive responses. Suppose, further, that some mechanism exists (as discussed in Section 4) that causes training to occur either as a single “reinforcement training event” initiated by some sensory activity, or as the result of repetition of correlated motor-sensory activity. Then random activity in the motor region that produces sensory feedback will result in the training of sets with elements in both regions. Later activation of a recall set in the sensory region will result in activation of the corresponding motor target neurons (i.e., recall of a sensory event reproduces the motor output that originally caused it).

In the networks discussed, here, regions with no active training set spontaneously fire about 6 neurons per cycle. The particular neurons vary in a pattern that appears random to the casual observer even if the network has been trained, previously, on many training sets. A 6-neuron training set is not sufficiently large to reliably suppress random activity. However,

training events may be made to produce larger training sets by at least four means.

First, neurons that have fired most recently may be regarded as currently firing. If the network trains neurons that have fired over the last 10 to 30 cycles, training events produce training sets of roughly 20 to 30 neurons per region.

Second, if training sessions involving the same motor neurons are repeated, randomly firing neurons will be added to the training set until the set is sufficiently large to inhibit random firing.

Third, because real brains exhibit behaviors that are not learned, there are instances in which it is appropriate to construct a network with some pathways that are already trained.

Fourth, it is possible for a training event to increase the number of randomly firing neurons.

Formation of trained sets has been simulated by all of the above mechanisms. However, in the simulations described in Sections 4 and 5, sets are trained by the first mechanism. A variety of mechanisms may be employed to signal a training event. For these studies, training occurred when specific sensory sets were activated that may be thought of as representing reward or punishment.

It is unnecessary to regard every set in the motor regions as corresponding to a behavioral output. Training may occur in response to sensory events that are not the consequence of motor output. In these cases the motor sets trained are representations in motor regions of events in sensory regions, but do not necessarily correspond to motor actions.

3.3 Serial Memory

Two small modifications of the above training and recall rules will allow a 2-region network to learn sets in series. Consider a series of training sets each of which spans sensory and motor regions. Suppose that, in addition to training links between currently active neurons, links from previously active motor neurons to currently active sensory neurons are also trained. In addition, suppose that either (1) motor neurons show burst firing for a longer period of time than sensory neurons, or (2) there is a periodic event that causes inhibition of currently active sensory neurons. Training and recall of a series proceeds as follows.

Suppose set 1 is activated and trained in both regions. After it has ceased bursting, slow firing of the set maintains the network in a stable state, there being relatively few random firings between cycles on which set 1 fires. Now suppose that, by some mechanism described below, set 2 is activated. When it is trained, motor elements of set 2 are trained to sensory elements of set 1 as well as sensory elements of set 2.

At a later time, set 1 is activated in the sensory region and becomes active in the motor region. Activity of set 1 in the sensory region suppresses set 2 in the sensory region until set 1 stops bursting. Set 1 remains active in the motor region and is trained to set 2 in the sensory region. So, set 2 becomes active in the sensory region. As sensory set 2 is not trained to motor set 1, sensory set 2 turns off motor set 1 and permits firing of motor set 2.

This two region serial network cannot reliably recall two series that share a common training set. Similarly, it has difficulty recalling a series that repeats the same training set. These difficulties are overcome in the next section.

4. A MORE COMPLEX MULTI-REGION NETWORK (A C-NET)

4.1 Architecture and Rules

There are many interesting ways in which regions with similar or identical properties may be brought together to perform functions similar to those of the network described in this section. Consider the specific architecture shown in Fig. 9 in which each R-net is a copy of the 4000-excitatory-neuron network previously described.

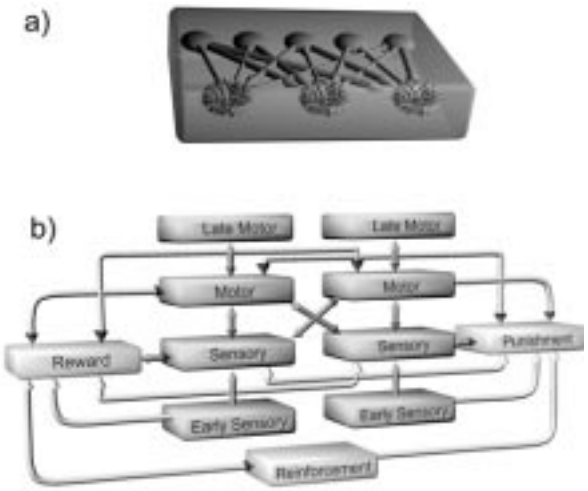

Fig. 9. (a) Each box represents a local R-net. No inhibitory projections extend outside a box. (b) A C-net. Arrows indicate the directions of excitatory projections between R-nets

This network includes two "channels." Each channel is a sequence of four local regions designated "early sensory," "sensory," "motor," and "late motor." During simulations, only neurons in the early sensory region are forced to fire by outside influences. As the early sensory regions are meant to reflect only the "sensory" environment of the network, without top-down influences, they do not receive feedback from the regions to which they project. (In a real brain such top-down influences might be used, for example, to influence attention without greatly perturbing the veracity of the information attended to. However, such considerations are beyond the scope of this paper.)

There are weak, reciprocal connections between sensory and motor regions of separate channels. (In the modeled networks, each excitatory neuron has 2 projections onto inhibitory and 160 projections onto excitatory neurons of any R-net outside its channel to which it projects). This architecture may be thought of as a blocky version of a real brain in which neurons project most densely to a small region and less densely to surrounding regions. In this architecture, information in one channel can serve as context for the other so that a channel may recall different series with common elements correctly.

The network has two "integrative" regions, designated the "reward" and "punishment" regions, each of which has reciprocal connections to the motor and sensory regions of both channels. Additionally, both integrative regions receive projections from the early sensory regions. (There are no connections to the late motor regions.) Because these integrative

regions connect to a large number of local regions, the number of projections to and from these regions is reduced to 20% of the usual number.

Finally, there is a “reinforcement” sensory region that connects only to the integrative regions. In it are two pretrained sets designated the “reward” and “punishment” sets. By a mechanism not made explicit in the model, activation of the reward set causes training to occur in all regions of the network except the punishment integrative region. Activation of the punishment set causes training to occur in all regions of the network except the reward integrative region. In addition, activation of the punishment set inhibits all activity in the late motor region. The reader will please be careful to distinguish between the reward and punishment sets in the reinforcement region and the reward and punishment integrative regions.

In most of the simulations described below, serial memory has been encoded by training recently, as well as currently, firing motor neurons to currently firing sensory neurons, and allowing motor neurons to burst for more cycles than sensory neurons. In other simulations, serial memory has been encoded by training recently firing integrative neurons to currently firing sensory and motor neurons. However, in some simulations described in Section 5, using the integrative regions in this way requires that they be linked to one another through inhibitory but not excitatory pathways so that both regions cannot fire at the same time.

For most of the simulations described below, neurons that were not inhibited were allowed to fire on every cycle of operation for the first 100 cycles (bursting) provided they continued to receive no inhibition. They then fired every 20 cycles, again provided nothing occurred to inhibit them. Neurons in the motor regions were allowed to burst for 130 cycles. When a set was activated in the reinforcement region, training sets were formed according to the first method of Subsection 3.2 with all neurons that had been active within 25 cycles being treated as currently firing. Additionally, projections were trained from motor neurons that had fired within the last 100 cycles onto currently firing sensory neurons. For sets that are not part of a series, this additional training is deleterious to storage. However, the network is sufficiently robust so that the consequences are small.

4.2 Quantitative simulations

The parameter space of the network has not been optimized. The effort to do so would not be rewarded by proportional results. Rather, these results emphasize the robustness of the network. In particular, all local R-nets have the same architectural and synaptic parameters with the exceptions that (1) there are no reciprocal projections to the early sensory regions, and (2) the number of projections between pairs of regions is sometimes reduced to keep the number of projections to and from integrative regions comparable to those of other regions. This results, for example, in some regions firing more neurons at random and developing larger training sets than others. The quantitative performance of the network might be improved, for example, by adjusting the strengths or numbers of excitatory projections to make the mean training set size the same in all regions.

The storage capacity of the network was evaluated by creating training sets encompassing all regions of the network. For some studies randomly selected sets of 15 to 30 neurons per region were activated and trained. For others, the network was allowed to run without an active training set, and training of random sets of variable size was occasionally triggered by activation of a reinforcement set.

After storage of sufficiently many sets, recall was evaluated by activating the elements of a set in the early sensory regions and allowing 100 cycles for the remainder of the network to converge on the target set. The percentage of errors indicated in Fig. 10 represents the ratio of errors to the total number of target neurons in all local regions except the reinforcement region which contains only the reward and punishment sets. The number of sets stored is substantially greater than in simple R-nets reflecting the presence of the relatively small number of additional projections between regions.

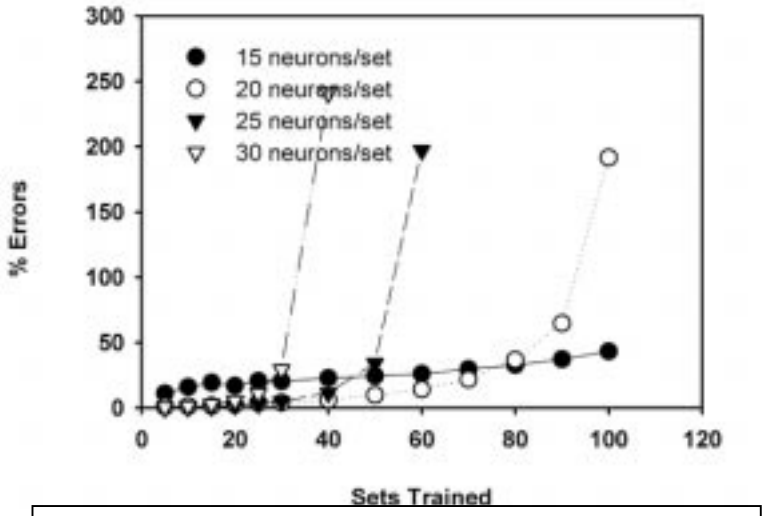

Fig. 10. Percent errors (spurious and missing neurons compared to target set size) as a function of the number of sets stored in the C-net described in Section 4.

When the network is allowed to generate training sets by means of its own more or less random firing, sets in the early sensory region ranged in size from 26 to 39 neurons (mean 30). In other local regions, set size ranged from 14 to 27 neurons (mean 23). The number of these “random” sets stored is not limited by errors, which never exceeded 1.9% with 30 sets stored. Rather, in each of 10 trials, there was a sharp transition that occurred over the range from 30 to 35 sets stored. When a small number of randomly selected neurons is activated in a C-net in which up to 30 sets have been trained, the network will continue to cycle for thousands of cycles without converging on one of the training sets. However, by the time 35 sets have been trained, the network begins to fall into some training set within 100 cycles, making it impossible to add additional training sets by means of the protocol employed.

4.3 Serial Memory

Serial memory was evaluated in the following way. First, some number of training sets was created by the first method described in Subsection 3.2 (Activity in the C-net was randomized and the network was allowed to run for 300 cycles at which time a set in the reinforcement region was activated causing the formation of a training set). Training sets were

then stitched together into a series in the following way.

A training set was activated in the early sensory regions and the entire C-net converged on the set over the 100 cycles of bursting of each neuron. Sometime during the next 1000 cycles, another training set was activated in the early sensory regions; the C-net again converged on the set over 100 cycles of bursting; and when a reinforcement set was activated, motor neurons from the first set became trained to sensory neurons of the second set. Again, at some time in the next 1000 cycles, a third set was activated in the early sensory region, and so forth.

To recall the series, training set 1 was activated in the early sensory regions. During the next 100 cycles the C-net converged on set 1. When neurons in the sensory regions stopped bursting, motor neurons continued firing for 30 cycles and, as they were trained to sensory neurons of set 2, caused set 2 to begin firing in the sensory region. These set 2 sensory neurons are not trained to set 1 motor neurons and, so, turn off set 1 and allow firing of motor neurons of set 2. When neurons in the sensory region again stop bursting, the network converges on set 3, and so forth.

Sets in a series were alternately trained to the reward and punishment sets in the reinforcement region, and recall of a series was considered successful if , after activation of the first set in the early sensory regions, the late motor elements of each set in the series were either recalled or inhibited in the proper order with no more than 10% errors on recall.

In a series of 10 trials, the network stored an average of 19 sets (range 15 – 21). Failure almost always resulted from a failure of the two channels to recall synchronously, resulting in simultaneous activation of both integrative regions.

An individual channel can learn different series that contain common sets. Suppose we wish to train two series, and are somehow more concerned with the order of sets in channel 1 than with the context in channel 2, though the context may also be a series. Further, suppose that set *x*, is to be the third element of series 1 and the tenth element of series 2. When we activate set *x* in channel 1 during the stitching together of series 1, we activate set *w* in channel 2. When we activate set *x* in channel 1 during the stitching together of series 2, we activate set *z* in channel 2. On recall, set *x* appears as the third element of series 1, along with set *w*, or as the tenth element of series 2, along with set *z*, and sets *w* and *z* have sufficient effect on the next set to appear in channel 1 to maintain the correct series.

The difficulty in this procedure lies in training the eleventh element of series 2. After set *x* is activated, the network tends to run on to the fourth element of series 1. The eleventh set of series 2 must be activated in the early sensory regions before the network has gone far toward activating the fourth element of series 1. (It may be entered at any time the network is still firing the tenth set of series 2.) A variety of mechanisms for overcoming this limitation may be imagined in marginally more complex C-nets, and a simple mechanism that does not involve reinforcement learning is briefly discussed in Section 6.

Two different series that have never been coactivated may be activated in separate channels. Each channel will then recall the series activated in it, though the response of the reinforcement region may be unpredictable when elements of the two series are trained to

different reinforcement sets. (Most commonly, in such cases, it fails to converge before sets are activated in the late motor regions.)

5. MORE COMPLEX BEHAVIORS OF A C-NET

5.1 Example Selection

The behavioral repertoire of this C-net may now be illustrated by giving examples of network behaviors that seem analogous to mammalian behavior. These examples do not exhaust the network's repertoire. However, they are, in some sense, paradigmatic.

All of the examples given here have been demonstrated by simulation. All may be demonstrated with C-nets having two channels. Most have been demonstrated in C-nets with up to 4 channels (but reduced R-net size and no late motor regions in order to accommodate serial computer memory constraints). The additional channels have been added in a modular fashion with no change in architectural or synaptic parameters.

5.2 Conditioned Learning

Suppose a C-net has been trained on some number of sets, not stitched into a series. Now, suppose a recall set of set 1 is activated in the early sensory region of channel 1. Suppose, further, that when set 1 becomes active in the late motor region, the punishment set is activated in the reinforcement region, and set 1 is trained to the punishment set. While set 1 remains firing, but not bursting, in the sensory and motor regions, set 2 is activated in the early sensory region of some channel, and when set 2 becomes active in the late motor region, the reward set is activated.

Subsequently, when set 1 is activated in the early sensory region, channel 1 converges on set 1. However, before set 1 becomes active in the late motor regions, the late motor regions are inhibited by recall of the punishment set. As soon as set 1 stops bursting in the sensory regions, set 2 is recalled and the late motor region converges on set 2. In short, the C-net now responds to the punished stimulus with the rewarded response. This example contains elements of classical, operational, and backward conditioning, and of secondary reward.

Questions naturally arise concerning the behavior of the network if, for example, set 1 was rewarded when the set was initially formed but punished when stitched into a series. It is, of course, easy to imagine networks in which the initial training of sets 1 and 2 does not include the integrative regions. However, no such special pleading is needed or even desirable.

It is a general property of the C-net that the neurons of a set in one channel may be trained to the reward integrative region on one occasion and the punishment integrative region on another. Both reward and punishment sets in the reinforcement region cannot be activated at the same time because of their reciprocal inhibition. If the neurons of set in one channel have been associated with both reward and punishment, then when they appear in a context that was rewarded, the elements of the set in the reward integrative region will fire before those in the punishment integrative region, turning on the reward set in the

reinforcement region and suppressing the punishment set. Conversely, if they appear in a context that has been punished, the punishment integrative region will fire first, activating the punishment set. If a set appears in a completely novel context, the usual outcome is that the reinforcement set will be activated that has been previously activated in the largest number of contexts. If past reinforcement has been about evenly mixed, activation of either reinforcement set may be delayed. In such a case, one or two more consistently rewarded or punished training sessions will result in more rapid reinforcement convergence.

5.3 Prediction and Refabrication

Predictive fabrication (planning) and refabrication of memory depend on the same principle. If set *y* is trained to set *x* during event 1, and set *z* is trained to set *x* during event 2, then if sets *y* and *z* are active during event 3, set *x* will be activated in preference to any other set to which *y* or *z* may be independently trained. The associations between *x*, *y*, and *z* may be either spatial (i.e., they may appear at the same time in separate channels of a 3-channel network), or temporal (i.e., set *x* may have followed set *y* in one channel and set *z* in the second channel of a 2-channel network). This principle will fill in possible future events, or memories that are only partially encoded, in a way that is likely to be reasonable, as is suggested by excessively anthropomorphic examples such as the following.

Suppose, on some occasion, a C-net forms an association between a representation of boating, set *b* in some channel, and a representation of docks, set *d* in some other channel. On some further occasion, it forms an association between Toucari Bay, set *T* in some channel, and docks, set *d*. If some series of recollections now causes recall of sets *b* and *T*, set *d* will also be recalled. So, the recall of boating at Toucari Bay will cause the dock to be recalled even if the network has never formed an association between boating-at-Toucari-Bay and the dock. If this network is imbedded in some larger C-net with additional local R-nets, the larger C-net may either plan to use the dock to go boating, a reasonable plan, or assert that on some previous occasion it used the dock to board the boat, a more reasonable refabrication than asserting that it walked on water to board the boat.

5.4 The Null Stimulus and Reality

The direct projections from the early sensory regions to the integrative regions allow the network to make “reality” part of the context of decision making. The most difficult examples to train are those in which the C-net must learn to recall elements of a set in the sensory and motor regions, but not in the late motor regions unless the set is also active in the early sensory region. To again elaborate an example, the difficult task is to learn to recall the mouse without pouncing unless the real mouse appears. This requires that the network be trained in several contexts, the majority of which result in punishment, the only context that leads to reward being that in which the set is active in the early sensory region.

It is, of course, possible to reverse the sets being rewarded and punished so that activation of the set in the sensory and motor regions always results in activation in the late motor region in the absence of the activity of a particular set in the early sensory region. To

resort, again, to elaborate analogy, multiple training sessions are needed to train the network to always run to the kitchen in the absence of the cat while only a single session is needed to train the network to not run to the kitchen in the presence of the cat. Teaching the network to respond, in this way, to the null stimulus requires more training than teaching it respond to the present stimulus. Of course, even a mammalian brain may find the distinction between "always run to the kitchen in the absence of the cat" and "don't run to the kitchen in the presence of the cat" a subtle one.

6. DISCUSSION

6.1 Memory and Conditioned Responses

Extremely sparse, powerful, local, inhibitory projections combined with relatively massive, weak, distant, excitatory projections are a common theme in the brain. It may be that these models reveal something of the function of this theme. The model based on this theme appears to support mammalian size memory capacity, is not subject to severe evolutionary constraints, and permits the modular construction of behavioral solutions to significant problems faced by living organisms. In terms of realism, there are many improvements to be made in these models, even without trying to associate particular regions of a model with particular regions of the brain. For example, it would seem more realistic to encode reward and punishment in separate, reciprocally inhibited, local regions, and to use the number of active neurons, rather than specific training sets to define reward or punishment. Placing specific sets in the same R-net is simply a convenient way of producing reciprocal inhibition (so that both sets do not fire at the same time) without elaborating new structural motifs.

The model is clearly an incomplete description of any particular region of the brain. It does not, for example, make use of such obvious organizational features as cerebral columns. However, I know of no reason why this model could not be integrated with models that use columnar structure to account for other aspects of cerebral function (e.g., Körner et al., 1999; Tanaka, 1999) .

The chief limitation of any model that requires "training events" seems to lie in the absolute need of reinforcement for learning to occur. As described here, series learning requires that each set in the series be rewarded or punished as the series is stitched together. There are two reasonable solutions to this problem.

If "training events" are retained in the model, it is not difficult to imagine that coincidence firing induces a state in a link that may lead to training if reward or punishment occurs in the near future. Such an arrangement would permit a series to be learned in response to a single reward. The problem would lie in determining the start of the series. The start might, for example, be defined by the onset of novelty. In spite of appearances, I am more interested in demonstrating the behavioral complexity of simple networks than in building airy castles, and I have not included a novelty detector in these models. I note, however, that information about novelty is present in the network in terms of the number of neurons active in the integrative region, the number being large when familiar events occur and small when novel events occur.

Alternatively, I have made preliminary progress on development of an algorithm that trains synapses on every cycle, but which requires many cycles of correlated firing for the training to become more or less permanent. It is necessary for such an algorithm to include detraining when firing is uncorrelated in order to prevent rapid over-training of the network. Such an algorithm allows repeatedly correlated early sensory neurons to generate trained sets without invoking reinforcement learning. It may be anticipated that such an algorithm will also permit training of series without the necessity of first training single sets and then stitching them together. Detraining may also help with the problem described above of defining the start of a series (the first set of the series being uncorrelated on most trials with any activity preceding the series on a particular trial).

More elaborate versions of this network may reasonably be expected to produce more complex psychological results. Recall may be initiated by an integrative rather than an early sensory region, and the integrative regions may stand in relationship to still higher local networks as the sensory and motor regions stand in relationship to the integrative regions. Top-down control might make the network less relentlessly driven by sensory events. For example, in the present network, during the transition in the sensory and motor regions from one set of a series to another, activation of a set in the early sensory regions is likely to interrupt read-out of the series. However, if an appropriate set associated with the series is already active in the integrative region during the transition, the firing of training sets in the early sensory regions will not interrupt read-out of the series.

6.2 Behavioral Logic

The simulation described in Subsection 5.4 is roughly analogous to teaching the network to run to the mouse hole and wait for the real mouse before pouncing. It is not especially disturbing that this is more difficult to learn than a statement such as “pounce if you think of a mouse.”

If, in the simulation of Subsection 5.4, the states that are rewarded or punished are reversed, the learned program is more nearly analogous to the statement, “Run to the mouse hole. If there is no cat, run to the kitchen”. That this requires more training than “If there is a cat, do not run to the kitchen” is also not surprising.

In general, it appears to be possible for sufficient training to enable this C-net to follow almost any if...then... rule with almost any combination of conjunctions (with the exception of XOR), negations, or existential imperatives in the subclauses.

6.3 Testable Hypotheses

The model suggests a few tests. Most conspicuous is a test of the disinhibition training algorithm. This might be accomplished with a dual patch-clamp apparatus, perhaps in the presence of substances that affect LTP or DSI such as acetylcholine (Martin & Alger, 1999) and carbonic anhydrase (Sun & Alkon, 2001).

The mechanism for serial memory requires that recently active projections of motor neurons onto sensory neurons be trained while only currently active projections of sensory

neurons onto motor neurons are trained. Asymmetries in synaptic training have been demonstrated. For example, Xie & Lewis (1995) have shown that, in the presence of naloxone, perforant path tetanization produces potentiation of inhibitory links onto CA1 pyramidal cells but not of the feedback links onto the entorhinal cortex. Similar tests of asymmetry and order dependence might be conducted in the same preparations as have been used to demonstrate disinhibition in the Schaeffer collaterals.

6.4 The Mystical Brain

The first and most obvious problem in discussing such matters as memory and consciousness lies in having some notion of what it is one intends to discuss. Requests to colleagues for characterizations of consciousness lead to long lists of inconclusive adjectives. Discussions of the subject come to cluster around two notions, "self-awareness" and, less commonly, "imagination" - whatever these are.

"Self-awareness" seems to indicate something self-referential, and physiologists rather commonly speak in some way of the brain looking at itself. For example, Crick (1979) writes, "There is more than a suspicion that such phenomena result from the computation pathways acting in some way on themselves.…" Allusions to imagination are less frequent, but Gregory (1977) seems to refer to something like imagination when he remarks that "Perceptual reality - consciousness - may be a feature of the simulated world or the brain on which we act predictively."

Consider the following tentative definition of consciousness: "Consciousness" is the ability of an information processing system to (1) use information about how the system is processing information adaptively (self-awareness), and (2) provide parts of the system that are capable of processing information about events outside the system with information about events that might occur (imagination).

Compare this definition with the behavior of the network during the simple conditioned learning described in Subsection 5.2. During recall, set 1 is activated progressively in the sensory and motor regions. When it is active in both regions the representation of this activity in the integrative region becomes active. This activity in the integrative region contains information about how the network is processing information. Specifically it contains the information that set 1 has become active in the sensory region and information has been processed so that set 1 is active in the motor region, and the corresponding response is about to be made in the late motor region. This information is acted on adaptively in a process that involves recall in both the sensory and reinforcement regions of training sets that represent events that might occur.

A mind-brain dualist reading this paper might well say, "Yes, there are correlates of the mind in the brain that appear to be discoverable, but when I look at these networks I do not see in them what I experience as consciousness."

A reductionist might reply, "Examining these networks is not the same as being them. Humans cannot experience what it is to be a network by examining it. Indeed, such experiences are probably among the many thoughts of which humans *a priori* are not capable.

You cannot demand of reductionism that looking at a network will make you feel what it is to be the network."

At another time, a reductionist might be able to claim that the network has all the known properties readily attributable to consciousness by dualists or reductionists. The problem for the reductionist lies in words like "known." There may always be, there may have to be, properties of consciousness that remain undescribed. So reductionists cannot really understand whether they have "really" understood consciousness. On the account given here, the issue lying between the reductionist and the dualist will never be resolved. Put as an antinomy in this form, it would be a waste of time to pursue the definitive resolution.

Aside from consciousness, it could be argued that any aspect of the "psychological complexity" of this simple network is the result of naming computational features of the network with complex names. As the properties of networks (like the behaviors of other individuals) are observed more or less empirically while such entities as our own memories or emotions are observed introspectively, the antinomy just described pervades every attempt to integrate any aspect of the psychology of the mind into the rest of science. Because looking at the "memory" exhibited by these networks does not feel like my memory, there may always be unknown properties of memory that prevent any reductionist understanding of memory.

This discussion adds little that is new to a long standing debate. However, the model provides an instantiation of a biologically plausible network with properties analogous to those of memory and consciousness. Granting every objection to the view that the network described here harbors a nanospark of consciousness, if there are correlates of mind in the activities of the brain, then finding properties of the network that behave like correlates of memory and consciousness still argues for the point of view that minds are not mystical objects, and some magnitude of consciousness may arise as an epiphenomenon in small brains organized to perform simple conditioning tasks.

APPENDIX

Architecture

All R-nets included in the C-net studied contained 4000 excitatory neurons each with 20 random projections onto and from 800 inhibitory neurons each with 100 random projections onto and from the excitatory neurons. Within a channel, each excitatory neuron in an R-net had 5 projections to inhibitory neurons and 400 projections to excitatory neurons of any other R-net to which it projected. Between R-nets of different channels, each excitatory neuron had 2 projections onto inhibitory neurons and 160 projections onto excitatory neurons. Between R-nets of channels and integrative regions, each excitatory neuron had 1 projection onto an inhibitory neuron an 80 projections onto excitatory neurons. Projections between integrative regions and the reinforcement region were like those within channels.

Algorithm

Activation of an R-net consisted of activation of certain excitatory neurons. Inhibitory neurons were then updated synchronously according the following rule.

$$a_i = \Sigma\, w_{ie} a_e$$

where a_i is the activation of the *i*th inhibitory neuron, a_e is the activation of the *e*th excitatory neuron with possible values 0 or 1, and $w_{i,e}$ is the strength of the projection of the *e*th excitatory neuron onto the *i*th inhibitory neuron with possible values of 1 (untrained) or 10 (trained). The summation is indexed over all excitatory neurons in all R-nets that project onto the *i*th inhibitory neuron.

Excitatory neurons are then synchronously updated according to the following rule.

$a_e = 1$ if $\Sigma\, w_{e,j} a_j + \Sigma\, I_i \geq 0$

$a_e = 0$ elsewise

where $w_{e,j}$ is the strength of the projection of the *j*th excitatory neuron onto the *e*th excitatory neuron with possible values of 0.1 (untrained) or 0.4 (trained), and I_i is a function of the *i*th inhibitory neuron given by

$I_i = -a_i$ if $a_i < 10$

$I_i = -1$ if $a_i > 10$ and the projection of the *i*th inhibitory neuron is untrained.

$I_i = t$ if $a_i > 10$ and the projection of the *i*th inhibitory neuron is trained.

t has a value of 0 in all cases except for small sensory and motor R-nets of 4000 excitatory neurons where performance is marginally improved by letting $t = 0.2$. No such improvement is noted when training set size is as large as 40 neurons.

All excitatory neurons except motor neurons in a C-net burst (fire on every cycle) for 100 cycles and then fire on every 20th cycle (if they are not inhibited) for 100 cycles after which they may again burst.. Motor neurons burst for 130 cycles. During training events, neurons that have fired within the last 25 cycles are trained according to the rules given as though they were currently firing. In addition, projections from motor neurons that have fired within the last 100 cycles onto currently firing sensory neurons are trained.

REFERENCES

Akdogan, I., Unal, N., Adiguzel, E., 2002. Estimation of the number of neurons in the

hippocampus of rats with penicillin induced epilepsy. Image anal. stereol. 21, 117-120.

Alger, B.E., Pitler, T.A., Wagner, J.J., Martin, L.A., Morishita, S.A., Kirov, S.A., Lenz, R.A., 1996. Retrograde signaling in depolarization-induced suppression of inhibition in rat hippocampal CA1 cells. J. physiol. 496, 197-209.

Bennett, M.R., Gibson, W.G., Robinson, J., 1994. Dynamics of the pyramidal neuron autoassociative network in the hippocampus. Philos. trans. R. Soc. Lond., B 343, 167-187.

Buckmaster, P.S., Schwartzkroin, P.A., 1995. Interneurons and inhibition in the dentate gyrus of the rat in vivo. J. neurosci. 15, 774-789.

Buzsáki, G., Horvath, Z., Urioste, R., Hetke, J., Wise, K., 1992. High-frequency network oscillation in the hippocampus. Science 256, 1025-1030.

Chechik, G., Meilijson, I., Ruppin, E., 2001. Effective neuronal learning with ineffective Hebbian learning rules. Neural comput. 13, 817-40.

Collin C., Devane, W.A., Dahl D., Lee C-J., Axelrod J., Alkon, D.L., 1995. Long-term synaptic transformation of hippocampal CA1 gammaminobutyric acid synapses and the effect of anandamide. Proc. Natl. Acad. Sci. U. S. A. 92, 10167-10171.

Conors, B.W., Malenka, R.C., Silva, L.R., 1988. Two inhibitory postsynaptic potentials, and $GABA_A$ and $GABA_B$ receptor-mediated responses in neocortex of rat. J. physiol. (Lond., Print) 406, 443-468.

Crick, H.F.C., 1979. Thinking about the brain. Sci. Am. mon. Sept., 130-137.

DeFelipe, J., 1997. Types of neurons, synaptic connections and chemical characteristics of cells immunoreactive for calbindin-D28K, parvalbumin and calretinin in the neocortex. J. chem. neuroanat. 14, 1-19.

Desai, M.A., McBain, C.J., Kauer, J.A., Conn, P.J., 1994. Metabotropic glutamate receptor-induced disinhibition is mediated by reduced transmission at excitatory synapses onto interneurons and inhibitory synapses onto pyramidal cells. Neurosci. lett. 181, 78-82.

Eichenbaum, H., Buckingham, J., 1990. Studies on hippocampal processing: experiment, theory and model. In: M. Gabriel & J. Moore (Eds.), Learning and Computational Neuroscience: Foundations of Adaptive Networks. MIT Press, Cambridge, MA, pp. 171-231.

Freund, T.F., Buzsáki, G., 1996. Interneurons of the Hippocampus. Hippocampus 6, 347-470.

Fukai, T., 1999. Sequence generation in arbitrary temporal patterns from theta-nested gamma oscillations: a model of the basal ganglia - thalamo-cortical loops. Neural netw. 12, 975-987.

Gibson, W.G., Robinson, J., 1992. Statistical analysis of the dynamics of a sparse associative memory. Neural netw. 5, 645-661.

Greenstein-Messica, A., Ruppin, E., 1998. Synaptic runaway in associative networks and the pathogenesis of schizophrenia. Neural comput. 10, 451-465.

Gregory, R.L., 1977. Consciousness. In: R. Duncan & M. Weston-Smith (Eds.) The Encyclopedia of Ignorance, Pergamon Press Inc.

Gulyás, A.I., Miles, R., Hájos, N., Freund, T.F., 1993. Precision and variability in postsynaptic target selection of inhibitory cells in the hippocampal CA3 region. Eur. j. neurosci. 5, 1729-1751.

Hasselmo, M.E., Schnell, E., 1994. Laminar Selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: computational modeling and brain slice physiology. J. neurosci. 14, 3898-3914.

Jani, N.G., Levine, D.S., 2000. A neural network theory of proportional analogy-making. Neural netw. 13, 149-183.

Körner, E., Gewaltig, M.-O., Körner, U., Richter, A., Rodemann, T., 1999. A model of computation in neocortical architecture. Neural netw. 12, 989-1005.

Marr, D., 1971. Simple memory: a theory for archicortex. Philos. trans. R. Soc. Lond., B 262, 23-81.

Martin, L.A., Alger, B.E., 1999. Muscarinic facilitation of the occurrence of depolarization-induced suppression of inhibition in rat hippocampus. Neuroscience 92, 61-71.

Miles, R., 1990. Synaptic excitation of inhibitory cells by single CA3 hippocampal pyramidal cells of the guinea-pig in vitro. J. physiol. (Lond., Print) 428, 61-77.

Miles, R., Katalin, T., Gulyás, A.I., Hajos, N., Freund, T.F., 1996. Differences between somatic and dendritic inhibition in the hippocampus. Neuron (Camb. Mass.) 16, 815-823.

Morishita, W., Kirov, S.A., Alger, B.E., 1998. Evidence for metabotropic glutamate receptor activation in the induction of depolarization-induced suppression of inhibition in hippocampal CA1. J. neurosci. 18, 4870-4882.

Nathan, T., Lambert, J.D.C., 1991. Depression of the fast IPSP underlies paired-pulse facilitation in area CA1 of the rat hippocampus. J. neurophysiol. 66, 1704-1715.

Ribak, C.E., Seress, L., Leranth, C., 1993. Electron microscopic immunocytochemical study of the distribution of parvalbumin-containing neurons and axon terminals in the primate dentate gyrus and Ammon's Horn. J. comp. neurol. 327, 298-321.

Salum, C., Morato, A.C., Roque-da-Silva, A.C., 2000. Anxiety-like behaviour in rats: a computational model. Neural netw. 13, 21-29.

Seress, L., Gulyás, A.I., Ferrer, I., Tunon, T., Soriano, E., Freund, T., 1993. Distribution, morphological features, and synaptic connections of parvalbumin- and calbindin D_{28k}-Immunoreactive neurons in the human hippocampal formation. J. comp. neurol. 337, 208-230.

Sik, A., Tamamaki, N., Freund, T.F., 1993. Complete axon arborization of a single CA3 pyramidal cell in the rat hippocampus, and its relationship with postsynaptic parvalbumin-containing interneurons. Eur. j. neurosci. 5, 1719-1728.

Staley, K.J., Brandi, L.S., Proctor, P.W., 1995. Ionic mechanisms of neuronal excitation by inhibitory $GABA_A$ receptors. Science 269, 977-981.

Stelzer, A., Simon, G., Kovacs, G., Rai, R., 1994. Synaptic disinhibition during maintenance of long-term potentiation in the CA1 hippocampal subfield. Proc. Natl. Acad. Sci. U. S. A. 91, 3058-3062.

Sun, M.-K. Alkon, D.L., 2001. Pharmacological enhancement of synaptic efficacy, spatial learning, and memory through carbonic anhydrase activation in rats. J. pharmacol. exp. ther. 297, 961-967.

Tanaka, S., 1999. Architecture and dynamics of the primate prefrontal cortical circuit for spatial working memory. Neural netw. 12, 1007-1020.

Taylor, N.R. Taylor, J.G., 2000. Hard-wired models of working memory and temporal sequence storage and generation. Neural netw. 13, 201-224.

Thomson, A.M., West, D.C., Hahn, J., Deuchars, J., 1996. Single axon IPSP's elicited in pyramidal cells by three classes of interneurones in slices of rat neocortex. J. Physiol. 496, 81-102.

Treves, A., Rolls, E.T., 1994. Computational analysis of the role of the hippocampus in memory. Hippocampus 4, 374-91.

Vincent, P., Armstrong, C.M., Marty, A., 1992. Inhibitory synaptic currents in rat cerebellar Purkinje cells: modulation by postsynaptic depolarization. J. Physiol. 456, 453-471.

Vogel, D., 1995. Burst neurons and lateral inhibition produce serial memory in a projective network. World Congress on Neural Networks '95 I, 462-465.

Vogel, D., 1998. Auto-associative memory produced by disinhibition in a sparsely connected network. Neural netw. 11, 897-908.

Vogel, D., 2001. A biologically plausible model of associative memory which uses disinhibition rather than long term potentiation. Brain cogn. 45, 212-228.

Vogel, D., Boos, W., 1997. Sparsely connected, Hebbian networks with strikingly large storage capacities. Neural netw. 10, 671-682.

West, M.J., Slomianka, L., Gundersen, H.J., 1991. Unbiased stereological estimation of the total number of neurons in the subdivisions of the rat hippocampus using the optical fractionator. Anat. rec. 231, 482-97.

Xie, C.W., Lewis, D.V., 1995. Endogenous opioids regulate long-term potentiation of synaptic inhibition in the dentate gyrus of rat hippocampus. J. Neurosci. 15, 3788-3794.

Yoshida, M., Hayashi, H., Tateno, K., Ishizuka, S., 2002. Stochastic resonance in the hippocampal CA3-CA1 model: a possible memory recall mechanism. Neural netw. 15, 1171-1183.