HIPPOCAMPUS 6:654–665 (1996)

Considerations Arising From a Complementary Learning Systems Perspective on Hippocampus and Neocortex

James L. McClelland and Nigel H. Goddard

Department of Psychology, Carnegie Mellon University, and Center for the Neural Basis of Cognition, Pittsburgh, Pennsylvania

ABSTRACT: We discuss a framework for the organization of learning systems in the mammalian brain, in which the hippocampus and related areas form a memory system complementary to learning mechanisms in neocortex and other areas. The hippocampal system stores new episodes and "replays" them to the neocortical system, interleaved with ongoing experience, allowing generalization as cortical memories form. The data to account for include: 1) neurophysiological findings concerning representations in hippocampal areas, 2) behavioral evidence demonstrating a spatial role for hippocampus, 3) and effects of surgical and pharmacological manipulations on neuronal firing in hippocampal regions in behaving animals. We hypothesize that the hippocampal memory system consists of three major modules: 1) an invertible encoder subsystem supported by the pathways between neocortex and entorhinal cortex, which provides a stable, compressed, invertible encoding in entorhinal cortex (EC) of cortical activity patterns, 2) a memory separation, storage, and retrieval subsystem, supported by pathways between EC, dentate gyrus and area CA3, including the CA3 recurrent collaterals, which facilitates encoding and storage in CA3 of individual EC patterns, and retrieval of those CA3 encodings, in a manner that minimizes interference, and 3) a memory decoding subsystem, supported by the Shaffer collaterals from area CA1 to area CA3 and the bi-directional pathways between EC and CA3, which provides the means by which a retrieved CA3 coding of an EC pattern can reinstate that pattern on EC. This model has shown that 1) there is a trade-off between the need for information-preserving, structure-extracting encoding of cortical traces and the need for effective storage and recall of arbitrary traces, 2) long-term depression of synaptic strength in the pathways subject to long-term potentiation is crucial in preserving information, 3) area CA1 must be able to exploit correlations in EC patterns in the direct perforant path synapses. © 1997 Wiley-Liss, Inc.

KEY WORDS: consolidation, amnesia, memory, interleaved learning

INTRODUCTION

Lesions of the hippocampal system produce a profound deficit in some forms of new learning and a temporally graded, retrograde amnesia for material experienced in a period of time preceding the lesion (Winocur, 1990; Squire, 1992; Kim and Fanselow, 1992; Zola-Morgan and Squire, 1990). These findings, and a computational analysis of neural learning systems, have led us to propose that the brain makes use of complementary learning systems (McClelland et al., 1995). One of these, called the *neocortical learning system*, is specialized for the gradual extraction of structure from ensembles of events and experiences, leading to the acquisition of connection weights among neurons that support generalization. Chomsky (1957) and Marr (1971) have stressed that experiences hardly ever repeat exactly. However, they do share with other experiences common structure, which we believe the neocortical learning system exploits so that appropriate responses to novel inputs can be made.

In McClelland et al. (1995) we argue that for the cortical system to extract structure, it must employ a strategy we call *interleaved learning*, in which exposure to and learning about the contents of one episode or experience is interleaved with ongoing exposure to and learning about other experiences. We stress that, as discussed in McClelland et al. (1995), the extraction of shared structure depends on interleaved learning in all known connectionist learning algorithms, biologically plausible or otherwise, including, for example, simple Hebbian learning, reinforcement learning, and back-propagation. According to this theory, temporally graded retrograde amnesia represents the gradual acquisition of memories by the neocortical system through this interleaved learning process. The function of the second system, the *hippocampal memory system*, is to store new memories and retrieve them while they remain in storage so that they can be "played back" to the neocortical system for interleaved learning with other memories and ongoing experience. We should note that all memories are subject to gradual decay from both the hippocampal system and from the neocortical system, although those that are the most strongly encoded and remain of behavioral relevance throughout the life span may be refreshed so that they persist, and perhaps can remain in storage in both systems throughout life. However, it appears that in older adults at least, most memories from late childhood and adolescence are no longer hippocampus dependent, since older hippocampal amnesics (ages 47–65) appear approximately equivalent to normal subjects on material from the first two decades of life (MacKinnon and Squire, 1989). On the other hand, the

Accepted for publication August 19, 1996.
Address correspondence and reprint requests to James L. McClelland, Center for the Neural Basis of Cognition, Room 115, Mellon Institute, 4400 Fifth Avenue, Pittsburgh, PA 15213.

same patients show profound retrograde amnesia for a decade or more preceding their lesions. According to the present article, this reflects hippocampal storage of traces of experiences, gradual loss of hippocampal traces over very extended periods, and the presence of mechanisms that allow the reinstatement of memories while they remain in storage so that they can contribute to memory performance and consolidation into the neocortical system.

GOALS OF THE MODEL

Our recent work has focused on elaborating the details of this theory of hippocampal function in a computational model. Our specific goals in analyzing and simulating the model are twofold. First we aim to examine whether the proposed mechanisms are computationally sufficient for the memory tasks that our theory ascribes to the hippocampal system: storing traces of experiences, retaining them for extended periods, and reinstating them while they remain in storage. Second, we aim to evaluate whether the computational properties that the hippocampus provides to support these functions are sufficient to account for the apparent spatial role of the hippocampus in learning and memory in rodents.

Regarding the first goal, our claim that the hippocampus has the capacity to be a locus of information storage contrasts with the views of a number of theorists (e.g., Alvarez and Squire, 1994; Murre, 1996; Teyler and Discenna, 1986; Damasio, 1989). These authors treat the hippocampal system not as the locus of memory storage, but as an area required for binding together elements of memories in disparate regions of the neocortex. The distinction between binding and actual information storage is a subtle one, and our views bear many similarities to the ideas of these other investigators (see McClelland et al., 1995, for full discussion). Here we simply note that in our theory "memories" stored in hippocampus are not complete copies of cortical patterns of activation but reduced descriptions that exploit redundancies in the cortical patterns. One of the motivations for the "binding" idea is the fact that the hippocampus is so small relative to the neocortical association areas that project to it, especially in humans and monkeys, suggesting that it could not have the capacity to represent and store the contents of neocortical patterns. This same motivation applies equally to our view of the hippocampus as storing reduced descriptions, and a goal of our computational modeling work is to specify exactly how this is achieved. Analyses and simulations of the model will determine whether the proposed mechanisms are sufficient, given the known physiological properties of the actual hippocampal system.

Regarding the second goal, our view is that the apparent spatial role of hippocampus in rodents is a manifestation of a more general capacity to associate arbitrary conjunctions of fragments of experiences. In the animal literature, O'Keefe and Nadel (1978) and many other authors take the view that the role of the hippocampus proper is restricted to spatial memory tasks, but the issue remains highly controversial (see Jarrard, 1993; Rudy and Sutherland, 1995, for discussion). Hippocampal damage in humans produces profound deficits in a number of nonspatial forms of learning and memory, suggesting a more general role in the formation of new memories. We propose that the apparently special importance of the hippocampus for spatial learning in rodents reflects this more general function rather than a specialization for space.

DATA TO BE ACCOUNTED FOR BY THE MODEL

Although our general theory of the role of the hippocampus in learning and memory is motivated by evidence arising primarily from effects of lesions of memory functions in humans and nonhumans, our attempt to model the hippocampus itself is grounded in studies of the responses of hippocampal neurons during behavior. Most of the evidence comes from rodents, although we would be very anxious to see data from monkey or other species as well. The main data of interest are as follows:

1. Representations used in hippocampus proper and related areas [entorhinal cortex (EC)]. Hippocampal neurons have highly selective and context-sensitive response profiles, even in the face of the apparently highly nonselective and diffuse response profiles of entorhinal neurons (Barnes et al., 1990; Quirk et al., 1992). Among the findings of interest is the fact that the activity patterns in hippocampal areas are highly sensitive to changes in, e.g., shape of the perimeter of the spatial arena, while the pattern of activity in entorhinal cortex appears to undergo topographic transformation with the change in the spatial arena (Quirk et al., 1992). Hippocampal areas have sparse activity patterns, ranging from 0.4% in dentate gyrus (DG) to 2.5% in CA1 and CA3, while EC activity (7%) is less sparse and far less selective (Barnes et al., 1990). In our view the sparsity and specificity of the hippocampal representations arises from the pattern separation process.
2. Spatiotemporal firing of hippocampal neurons as a function of spatial, task, and context-related inputs, including the data taken as supporting intrinsic spatial maps (McNaughton et al., 1996). A great deal of evidence has been amassed indicating that hippocampal neurons in rats tend to fire at specific locations in space. Neurons in DG, CA3, and CA1 may fire in a highly specific manner in spatial environments, exhibiting "place fields" (O'Keefe and Dostrovsky, 1971; O'Keefe and Conway, 1978; Jung and McNaughton, 1993). Because place fields persist in the dark, and because they sometimes persist in the correct location in space, even when visual cues have been altered, McNaughton et al. (1996) take the view that hippocampal neurons are fundamentally tied to an animal's sense of its location in a two-dimensional reference frame and that the hippocampus may contain preformed two-dimensional frames based on pre-existing, spatially organized connectivity among hippocampal neurons. However, in keeping with our view that the hippocampus is a domain-general computational structure rather than one specifically preorganized for

spatial information processing, we are exploring the possibility that no intrinsic spatial structure need be built into the hippocampus itself to account for the spatial firing of neurons in the hippocampus.

3. Data not yet available on the effect on neuronal firing in various hippocampal regions of surgical and pharmacological manipulations. This is a longer-term goal. An example of the kind of data we expect to explain is that reported by Mizumori et al. (1989) on preserved spatial coding in CA1 during suppression of CA3 firing. This data, relevant to the possible role of EC cells in activating neurons in CA1, comes from a recording study carried out in animals during temporary inactivation of the medial septum. The medial septum transmits modulatory cholinergic and GABAergic input throughout the hippocampus, and was temporarily inactivated by injection with Tetracain. The central data that the model must account for are that this manipulation severely disrupts spatial tuning and reduces overall firing rates in CA3, while leaving CA1 output almost unaffected. One possible interpretation of this effect is that CA1 cells can be driven by their EC afferents; as we shall see this is a requirement of the model that arises from the computational considerations outlined below.

EVALUATION

Evaluation of the model is in terms of its ability to 1) account for the data described above and 2) demonstrate computational sufficiency, through analyses and simulations. Computational sufficiency will be the principal focus of the rest of this article, in which specifics of the computational considerations can be found. In future work, we plan to continue to develop the computational analysis, even as we begin to turn our attention to modeling the physiological data in more detail.

THE COMPUTATIONAL MODEL

During an experience, the pattern of activation over neocortical associations areas (AAs) gives rise to a pattern of activation over the EC, the input–output interface of the hippocampal system. We assume that the forward pathways from the association areas to EC produce a pattern of activation on EC that maximizes preservation of information about the neocortical pattern, and that the return pathways from EC to the neocortical association areas invert the compressed EC representation, producing an approximation of the pattern that gave rise to it on the AAs. The EC pattern is called a *compressed invertible code*. The pattern on EC provides input to the hippocampal memory system, where it is recoded in areas DG and CA3 into a form suitable for storage in the synapses onto and among neurons in dentate gyrus and CA3. The pattern induces activity-dependent associative plastic changes in these synapses, which serve as the physical substrate of the memory.

Once a memory trace has been stored in the hippocampus, a probe of the memory might arise, in the form of a partial recurrence of the neocortical representation that was present initially during learning. For example, if one had previously met a student, associating his name with his face, one might wish to retrieve the name upon next seeing the student's face. The representation of the probe in the neocortical areas induces a representation on the EC, which in turn provides an input to the hippocampus. Successful recall requires reinstatement of an approximation of the pattern previously assigned to the episode in area CA3, using the pattern completion properties of the associative changes to synapses mentioned above. Return pathways from the hippocampus to the EC, via area CA1, reinstate the EC pattern, which in turn reinstates in the AAs an approximation of the activity present there during the original episode. Retrieval is successful if this reinstated pattern is sufficient for generating an appropriate overt response; in general it is likely that the recalled pattern reflects considerable loss of detail and increased noise relative to the original.

Consolidation of the trace of an episode is accomplished by gradual plastic changes in cortex which extract the shared structure in the traces of many interleaved learning experiences, as reflected in the statistics of occurrence and reoccurrence of each trace. The hippocampal system allows the interleaving of past experiences with ongoing experiences. The trace of a prior episode can be reinstated as a result of a "probe" that reactivates the stored trace. Alternatively, reinstatement may occur during active rehearsal, reminiscence, and other inactive states possibly including sleep (Marr, 1971; Buzsaki, 1989; Wilson and McNaughton, 1994a,b). The association between an individual's name and his face is a simple example of the shared structure extracted in cortex, if indeed the trace or traces containing this association reccur.

Computational Issues Facing the Hippocampal Memory System

The concept of cortico-hippocampal function in memory storage and retrieval described above indicates several key computational goals for the hippocampal memory system. A key goal that has been recognized by other researchers (McNaughton and Morris, 1987; Treves and Rolls, 1991) is *capacity*. The patterns of activity within the hippocampal system must be chosen to maximize the capacity of the system and minimize the interference between memories stored at different times. Additional goals arise within our theory, due to the focus on a long-term involvement of the hippocampal system in reinstatement and consolidation of memories. A second goal is *invertibility and information preservation*. The projections to the EC from the neocortical association areas must form patterns on EC that maximize the retention of information about the neocortical pattern representing an experience. The return connections must optimize the reinstatement of the original pattern in the neocor-

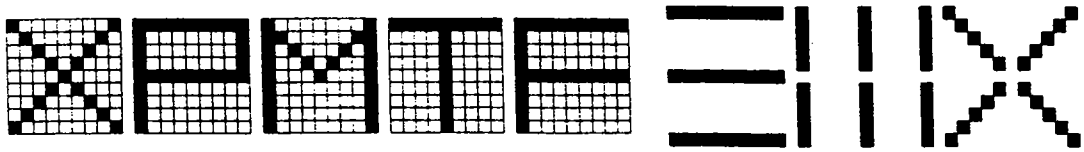

FIGURE 1. Five letters on a 5 × 9 × 9 pixel array, and a set of strokes that can provide a more efficient encoding.

tex, given the information retained in the representation in EC. A third goal is *stability*. Because of the temporally extended role of the hippocampal system in learning and memory, the encodings of probes produced by the cortical input to the hippocampal system must remain relatively stable over time. Return connections from EC to neocortical association areas must likewise remain relatively stable, so that patterns reinstated in the EC can give rise to the appropriate corresponding patterns of activation in the neocortex.

A central insight which motivates much of our present research and, we think, helps explain some of the complexity of the hippocampal memory system is that the first goal (storage capacity) appears to be in fundamental conflict with the second and third (invertibility and stability). Regarding capacity, it is a property of associative memory that interference between memories (and thus capacity) is directly related to the overlap of the patterns of activation that represent them (Willshaw, 1971, 1981; McClelland, 1986): Overlapping patterns interfere in their synaptic modifications. Overlap can be reduced by making the patterns relatively sparse (few units on in each pattern) and by recoding them in ways that reduce their similarity, a process which we call *pattern separation*. Regarding invertibility and information preservation, research on compressed, invertible encodings using connectionist "encoder" networks (Ackley et al., 1985; Cottrell et al., 1987; Brousse, 1993) has established that the most information-efficient encodings—the ones that preserve the largest amount of information about the input with the smallest number of units—are 1) dense (i.e., many cells active in each pattern), 2) retain the similarity structure of the space of encoded patterns, and 3) are structure dependent, exploiting redundancies present in the ensemble of input patterns. Thus the goals of maximizing storage capacity, invertibility and stability are in fundamental conflict.

A full understanding of the issues involved in resolving this conflict requires introducing a key distinction: componential versus conjunctive representation. To explore this distinction, we introduce an example domain consisting of pixel arrays representing character strings. Figure 1 shows an example of a five-letter string in a 9 × 9 pixel font, together with a set of strokes that can provide a more efficient encoding. A direct representation of this input, in which a single unit is dedicated to each pixel, requires 405 units. The five letters activate 112 (28%) of these units. Now consider two possible recordings, illustrated in Figure 2.

Random, conjunctive code

Each unit represents a randomly chosen conjunction of the original elements. One version of such a code, illustrated on the left of Figure 2, assigns one unit to represent every possible triple of pixels in the input—in this case there are over 10 million such triples. The letter string shown would activate about 230,000 of these units, but this represents only a small fraction of the total number (about 2%). A random subset of the 10 million triples would be sufficient to retain all of the information content of the initial input, since each pixel is actually represented in over 80,000 triples. Such a code is not only sparser than the original (2% vs. 28% of units active), but it also reduces overlap of similar inputs. In general, if S represents the sparsity of a pattern, the sparsity of a conjunctive recoding of the pattern using conjunctions of order R will be approximately S^R. Similarly, if Ω represents the overlap of two patterns, the overlap of an order-R conjunctive recoding of the patterns will be approximately Ω^R (Marr, 1969, 1971). O'Reilly and McClelland (1994) suggested that the hippocampus uses a generalization of this scheme, in which encoding units each receive excitatory connections from a randomly chosen fraction of the input units (perhaps 5%, or about 20 in this case), and a small proportion α of encoding units receiving the largest number of active inputs become active (the selection of the "winning" units is assumed to occur via lateral inhibition, see below). This form of encoding, which we call sparse, random, conjunctive coding, produces sparsification directly through the choice of (α); it produces separation effects comparable to values of R in the range of about 3–5, with smaller values of α corresponding to larger values of R.

Componential code

In a second coding scheme, illustrated in the middle of Figure 2, each unit represents a component or feature of the input, where a component is a recurring pattern of elements, or subpattern thereof. In our example, each letter can be formed from a set of 13 "strokes" represented schematically on the right of Figure 1, so that each set of 81 pixel units can be replaced by a set of 13 feature units. The representation is far more efficient, since fewer units are required; it is dense, however, in that a relatively large fraction (about 34%) of the encoding units will tend to be active to represent a given letter string. It is similarity preserving, in that similar input patterns retain (much of) their similarity to each other after they have been recoded. Componential codes with these characteristics are exactly the types of codes that are formed when connectionist networks are optimized to re-represent a set of inputs originally represented over a large set of units on a smaller set of units, although the components extracted from real-world datasets are not usually as conceptually transparent as are the components in our illustration. Note that this code is structure dependent: It succeeds only because the input patterns really do

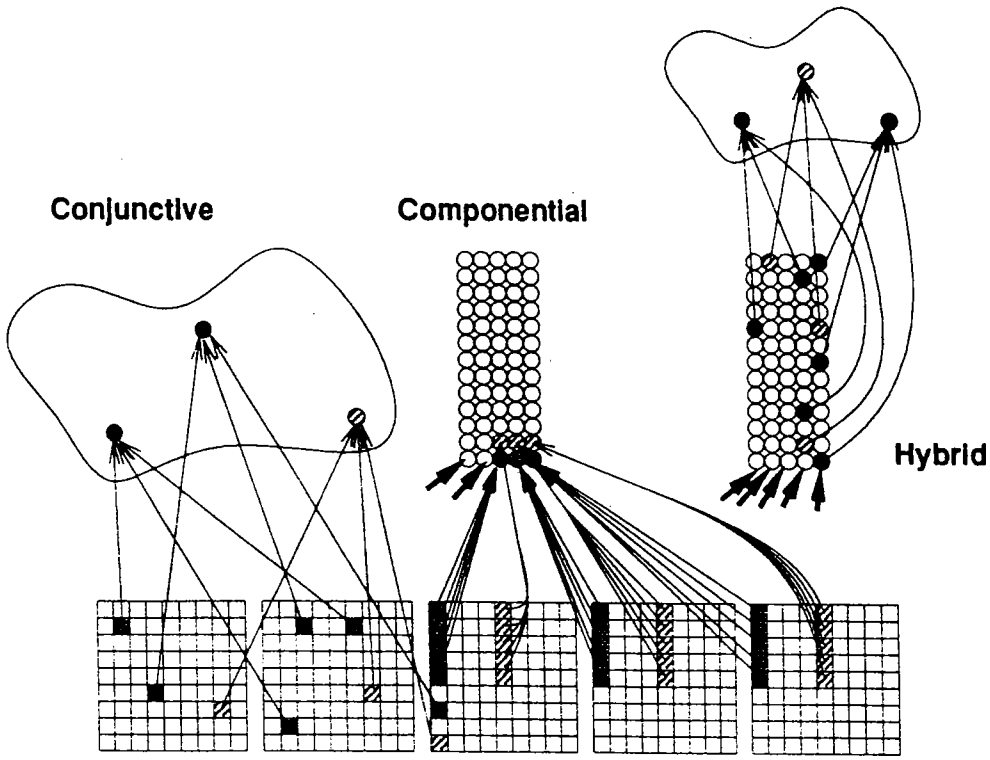

FIGURE 2. Two coding schemes and their hybrid combination.

have structure—that is, they really do consist of the represented components. If random pixel patterns were presented, the componential code would not be information-preserving. The code exploits the redundancy of the input patterns—if only letters made of the indicated strokes are ever used, then whenever certain pixels are active, certain other pixels must also be active.

A central thesis of our view of memory is that the experiences that we learn about have at least an approximately componential structure; in general we think of them as *arbitrary combinations of recurring components*. The letter string shown in Figure 1 exemplifies this. The particular sequence of letters is arbitrary but the strokes that form the letters (and indeed, the letters themselves) are recurring elements. (Again, we hasten to add that the statistical components of real experiences may not be so directly interpretable). The proposed scheme for memory, then, is to first encode the information in componential form and then subject it to conjunctive coding, as illustrated on the right of Figure 2. Specifically, we suggest that

- The pathways from the neocortical association areas to the entorhinal cortex form a compressed, componential representation of cortical contents.
- The pathways into the hippocampus then form representations capturing the particular arbitrary conjunctions of these components that occur in the given experience. Plastic changes in the connections onto and among the units representing the elements of the conjunctions then bind the elements of the pattern together.

Continuing with our example, the pixel representation corresponds approximately[1] to the representation in neocortical association areas; the stroke representation corresponds approximately to the representation in EC; and a conjunctive encoding applied to that corresponds approximately to the representation in area CA3. The EC representation is clearly condensed relative to the cortical representation, and a hippocampal conjunctive representation based on triples of strokes can now be formed with approximately the same sparsity and separation characteristics as before, using units that now represent conjunctions of these units. The example makes a key point for our theory: Sparse, random conjunctive codes are crucial for maximizing the storage capacity of the hippocampal system and minimizing interference among stored patterns, but these would be more efficient if they were applied to structure-dependent, componential codes optimized for the preservation of the information content of cortical patterns on the small number of neurons provided by the entorhinal cortex.

In the spatial domain this sparse conjunctive representation naturally gives rise to place fields: A place, in our view, is such a conjunction, though not necessarily one of simply sensory cues: the animal's sense of its relative location in space, which may be partially dependent on temporal integration of self-motion information, may contribute importantly to the sense of place. We would note as well that so-called place cells appear to represent

[1]In fact neocortical representations are themselves likely to be partially componential, to the extent that environmental structure has been extracted already.

conjunctions of location with other aspects of the animal's situation, including the task (Markus et al., 1995) and/or the relative location with respect to various features of the environment (start box, goal location, etc.) (Gothard et al., 1996). In our view, place cells might perhaps better be named "situation cells," in that their response is not so much invariant with respect to space but with respect to a particular situation relevant to the animal's interactions with the environment.

Implementation of These Ideas in the Brain

In our view, the complex architecture of the hippocampal system constitutes an elegant computational mechanism that addresses these considerations, maximizing retrieval fidelity and memory capacity given the limited resources that are allocated to this function in the brain. Based on these considerations, and on known aspects of the anatomy and physiology of the hippocampal system, we divide the hippocampal system (and its interconnections with neocortical association areas) into three subsystems, shown in Figure 3:

1. Structure-preserving invertible encoder subsystem. These are the pathways between neocortex and entorhinal cortex, including those via parahippocampal and perirhinal cortex. The goal of this system is to provide a stable, compressed, invertible encoding—in entorhinal cortex—of the cortical pattern produced by an experience.
2. Memory separation, storage, and retrieval subsystem. This subsystem consists of the pathways from entorhinal cortex to CA3, both direct and via the dentate gyrus (DG), and the recurrent collaterals within CA3. The role of this subsystem is to recode to-be-learned patterns in sparse, random conjunctive representations in areas DG and CA3, store these patterns in plastic changes in synapses within this subsystem, and, later, allow for the retrieval or reinstatement of stored patterns from partial inputs.
3. Memory decoding subsystem. This subsystem consists of the Shaffer collateral connections from CA3 to CA1, and the bi-directional pathways between EC and CA1. The role of this subsystem is to provide a means by which a retrieved CA3 coding of an EC pattern can reinstate that pattern on EC, whence the outgoing projections to neocortex reinstate the corresponding neocortical trace.

The first two of these three components have already been motivated, but the role of the third, the memory decoding subsystem, remains to be discussed. Given just the first two components, the hippocampal system is left with representations, in CA3, that are essentially random in their relation to the entorhinal input representations. Because of the random relationship, the association between these patterns must be learned; but because it must be learned, it provides a potential locus for interference among memory traces. As we discuss in detail later, we think the memory decoding subsystem is structured in a way that allows for the learning of these associations with a minimum of interference between memories.

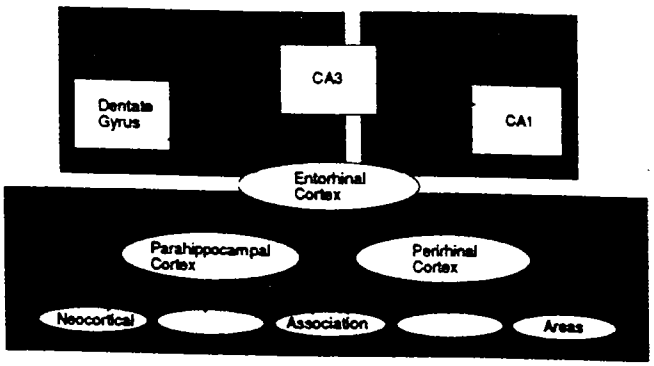

FIGURE 3. EC and CA3 form interfaces between the three cortico-hippocampal subsystems.

Memory Separation, Storage, and Retrieval Subsystem

Our analysis of the memory separation, storage, and retrieval subsystem considers how the direct and indirect (via DG) pathways from EC to CA3 assign sparse, separate patterns in CA3 based on overlapping input in EC, and how long-term potentiation (LTP) in the synapses from the EC input neurons to granule cells in DG and pyramidal cells in CA3, and among neurons in CA3, might then provide a basis for the storage of such patterns, in the sense that these changes would allow this system to retrieve an approximation of the pattern on CA3 assigned at the time of learning from a partial EC input. The results of the analyses will allow us to establish estimates of the capacity of the system and to understand how the particular characteristics of the various elements of the system contribute to memory separation, storage, and retrieval. They will also allow us to address several aspects of the firing of the principal neurons in areas CA1, CA3, and DG during spatial behavior in rats, with a view toward understanding these properties as reflecting the memory separation, storage, and retrieval process in action.

In our initial modeling work in this area (O'Reilly and McClelland, 1994), we examined how separation, storage, and retrieval might occur through the use of the sparse, random conjunctive coding solution, which we assume is exploited both in the direct EC–CA3 pathway and in the EC–DG pathway; the role of recurrent collaterals in CA3 was not considered. The basic idea was first suggested by McNaughton and Morris (1987), drawing on ideas in Marr (1971); this form of encoding has also been studied (under different names) by Torioka (1979) and Gibson et al. (1991). We used an analytically tractable model that allowed us to consider the effects of these variables in a system containing any number of neurons and synapses, overcoming the practical limitations of small-scale computer simulations of the sort used by many other investigators (Gluck and Myers, 1993; Hasselmo and Schnell, 1994) to examine cases approximating the numbers of neurons and synapses thought to be found in the relevant brain areas in the rat (Amaral et al., 1990).

In many theories of memory storage in the hippocampus, memories are thought to be stored primarily in the synapses among the CA3 pyramidal neurons, since these appear to form a

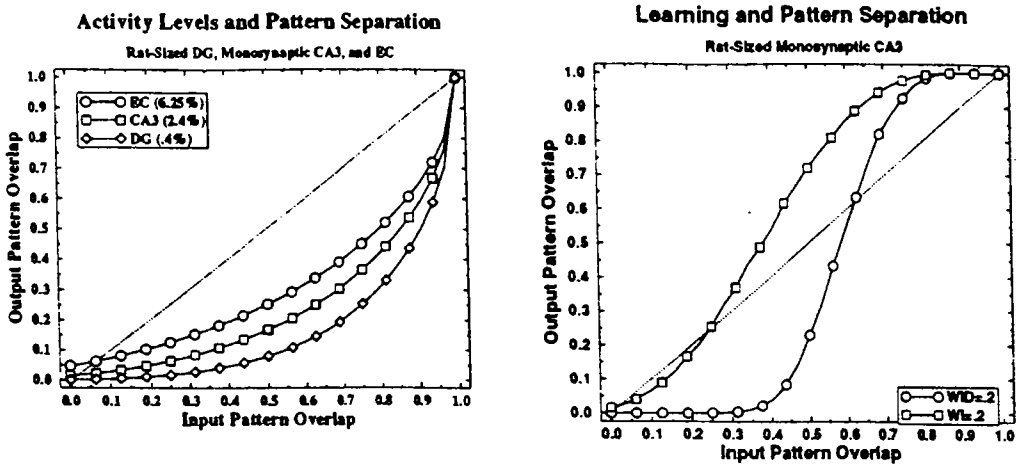

FIGURE 4. Left: Effect of output layer activity levels typical of EC, DG, and CA3 on pattern separation. Pattern separation is enhanced for sparser activity levels. Right: Effect of LTP alone versus LTP with LTD. LTP alone reduces pattern separation; LTP with LTD produces a cutoff below which separation is enhanced and above which it is diminished.

(sparsely connected) attractor neural network. However, associative enhancement of synapses in a feedforward pathway can also provide a basis for memory storage and retrieval (McNaughton and Morris, 1987). Because of this property, and because it appears that there is a high degree of *N*-methyl-D-aspartate (NMDA)-mediated plasticity in both the EC–DG and EC–CA3 projections, we treat plasticity in these pathways as part of the basis of memory storage and retrieval in the hippocampal memory system.

In the model, patterns of activation on EC were treated as binary; i.e., units were assumed to be either on or off in a given input pattern. Units at the receiving end were assumed to receive connections from a random subset of the EC input neurons, with connections assigned initial weights of 1.0. A sparse code was formed at the receiving layer by assigning the k neurons receiving the largest input activations of 1, and assigning all other neurons activations of 0 (this form of output representation is called *k-winner-takes-all* or kWTA). These assumptions are obviously idealizations. The kWTA representation is thought to approximate the effect of inhibitory circuitry within DG and CA3, which is thought to regulate the degree of activity in the system though a combination of feedforward and recurrent inhibition, as suggested by McNaughton and Morris (1987) and others. The idealizations allow the use of known properties of statistical distributions such as the hypergeometric to study the treatment of similar input patterns.

We studied how the parameters of these processes (sparsity of the input and output; fraction of input units projecting to each receiving neuron) influence the degree of separation of similar patterns, by examining the effect of these parameters on the function relating similarity of two input patterns to the similarity of the corresponding output patterns. Surprisingly, fan-in (number of afferents) had little effect on separation; the main operative variable is sparsity of the output representation, as illustrated on the left of Figure 4. We examined the combined effects of the direct EC–CA3 pathway together with the indirect pathway via DG, considering especially the case approximating data reported in Barnes et al. (1990), in which the DG patterns are very sparse (.004), and CA3 patterns are less sparse (.025) and in which each DG neuron synapses with only about 15 CA3 neurons but has a relatively large effect on the receiver—up to 50 times the effect of each direct EC–CA3 connection (Brown and Johnston, 1983). In this case, the resulting pattern separation in CA3 is far greater than it would be in the absence of the DG input, supporting the idea that one function of DG is to enhance the separation of patterns on CA3 (Treves and Rolls, 1992).

We examined the effects of associative LTP on pattern completion: After the presentation of an EC pattern and the formation of corresponding patterns on CA3 and DG, synapses from neurons active in the input representation to neurons activated in CA3 and DG were enhanced. The results, illustrated on the right of Figure 4, indicated that enhancement of synapses to learn one pattern interferes with maintaining separation of the learned pattern from other, similar input patterns: A new input pattern that overlaps with the learned pattern now tends to activate the output neurons assigned to the learned pattern. The line in the figure labeled WI shows how output pattern overlap is increased for given input pattern overlap with a moderate learning rate in the synapses subject to LTP. The incorporation of long-term depression (LTD) of synapses from inactive presynaptic neurons to active postsynaptic neurons greatly reduced this tendency, suggesting that some form of LTD is crucial for keeping memories separate. The line in the figure labeled WID shows that with this LTD incorporated, output pattern overlap is kept very low until input pattern overlap reaches about 50%.

We also found that allowing the EC–DG–CA3 pathway to play an important role during retrieval tended to interfere with pattern completion in CA3, compared with the case in which this pathway is left out and the direct EC–CA3 pathway dominates during retrieval. The reason for this is that the extreme sparsity in DG reduces completion there relative to CA3; the DG–CA3 pathway tends to pass this on to CA3. Therefore, it would be

advantageous for DG to be involved in storage but not in retrieval; experimental data would be useful here. An approximation of this could be implemented biologically: If neurons tended to have fixed instead of floating thresholds, relatively few would become active with "partial" input patterns. Our analysis showed this tends to lead to results almost as good as we find when the DG input is completely eliminated during retrieval. In summary, our analysis suggested how several properties of the hippocampal system contribute to memory separation, storage, and retrieval and indicated the computational impact of possible properties such as LTD and differential involvement of DG in storage vs. retrieval.

Memory Decoding Subsystem

In our overall model the task of the memory decoding subsystem, which comprises the Shaffer collaterals, area CA1, and the bidirectional pathway between EC and CA1, is to provide a means by which a retrieved CA3 coding of an EC pattern can reinstate that pattern on EC. The subsystem described above leaves a pattern in CA3 that is essentially random with respect to its EC progenitor, requiring that the decoding subsystem must learn the association between the EC pattern and its CA3 encoding. The CA3 code is sparse and conjunctive, which optimizes associative memory capacity by minimizing overlap between patterns. However, the EC code is dense and componential, characteristics which severely reduce associative memory capacity. Direct association of the EC pattern with its CA3 encoding would compromise the overall capacity of the system. Our proposal concerning the role of CA1 is that it provides a specialized, stable, invertible encoding of EC patterns, optimized for association with the CA3 pattern by being sparse and conjunctive. Our proposal gives an important role to the projection from EC_{III} to CA1: At the time the cortical trace of an event is encoded and stored in area CA3, a second sparse representation of the cortical trace arises in CA1, via activity in the EC_{III} to CA1 projection, and is associated with the CA3 pattern via Hebbian synaptic modification in the Shaffer collaterals. When the CA3 pattern is subsequently reinstated, these collaterals reactivate the CA1 pattern, which in turn reinstates the EC pattern via the path from CA1 to EC_V. The CA1 representation is assumed to be stable post-development so that it does not vary significantly during consolidation. The use of this intermediate sparse encoding in CA1 should increase the capacity of the system as a whole since it sparsifies and separates the pattern on the output side of the encoder network.

We have begun to analyze the operation of the memory decoding subsystem (Goddard et al., 1995), producing preliminary results that future work will build on. The characteristics required of the CA1 representation in our theory are invertibility and sparse conjunctivity. Our work to date on this subsystem has been an investigation of how the bidirectional EC–CA1 pathway could implement the encoding and decoding of EC patterns in a sparse, conjunctive form in CA1. A first set of studies examined whether patterns created using random forward projections from the model's EC to CA1 could then be effectively inverted if each CA1 unit returns connections to the same EC units from which it receives connections in the forward direction. The adequacy of this system depends on the "expansion factor" of the CA1 representation compared to the representation in EC, defined by the ratio of the number of active units in the CA1 representation divided by the number active in the EC representation. Using values for the fraction of neurons active from Barnes et al. (1990) and values for numbers of pyramidal neurons in each field from (Amaral et al., 1990), we found that invertibility, as measured by the correlation between the original and recovered EC patterns, was marginal: Using optimistic estimates a correlation of 0.92 was found, but with more pessimistic estimates a much lower correlation of only 0.65 was obtained. Our next step has been to work with the lower figure and consider ways in which the performance of the system might be improved that are consistent with the physiology.

The work just described used random binary patterns in EC. The assumption of random binary patterns in EC is both too simple (EC activity patterns show the graded and probabilistic character discussed in the preceding section) and too demanding (due to the highly overlapping receptive fields of EC neurons, there is redundancy in the input patterns which should ease inversion). Furthermore, the assumption of an exact reciprocal connection structure between EC and CA1 seems implausible. We modified the model to address both sorts of difficulty, using an idealized sensory system to generate EC patterns that are graded and exhibit correlation of EC cells' activations.

To exploit these correlations while simultaneously relaxing the assumption of 1-to-1 reciprocity, we first retained the random forward projections but considered return projections specifically chosen to target EC neurons whose activity was most highly correlated with the activity of the neurons in CA1 (see below for a consideration of the possible developmental basis of these projections). With the same expansion factor that previously produced invertibility values of 0.65, we found very substantial improvements (to $r = 0.903$, see Fig. 5a). We then investigated exploiting some degree of correlation in the forward projection. With correlations exploited in both directions, this increased invertibility further (to $r = 0.967$, see Fig. 5b) but at the price of reducing pattern separation in CA1: As the EC neurons that project to a particular CA1 neuron become more correlated, the CA1 representation becomes less conjunctive, reducing separation. There is thus a tradeoff between maximizing invertibility and maintaining separation. The findings suggest that the tradeoff is relatively benign in this case: With moderate exploitation of correlation in the forward projection, one produces invertibility that is almost as good as that obtained with maximal exploitation of correlation ($r = 0.949$), while maintaining good pattern separation (Fig. 5b,c).

Future directions

A long-term goal of our project is a fuller implementation of a model of the hippocampal memory system, including the entorhinal cortex, parahippocampal and perirhinal cortex, and the neocortical association areas. As our understanding of the compu-

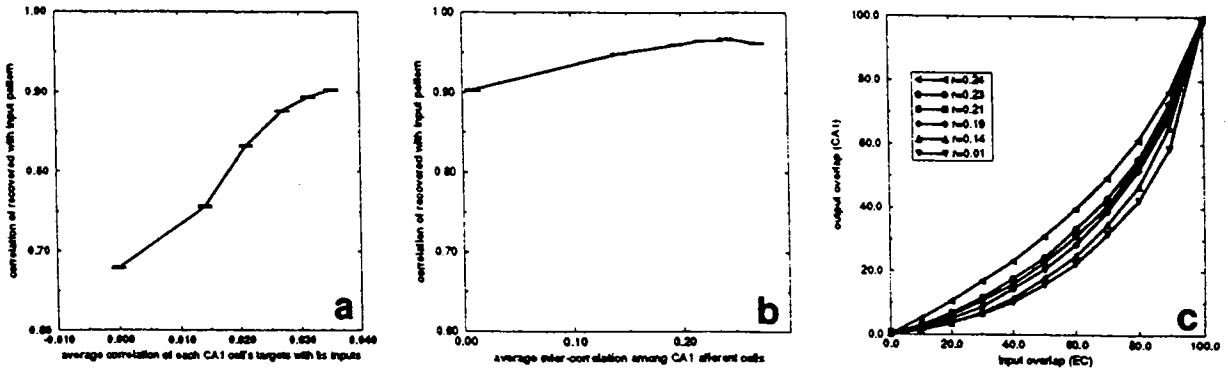

FIGURE 5. **a: Effect of correlations in the return projection from CA1 to EC on correlation of the original pattern of activation on EC with the pattern reconstructed by the EC-CA1-EC invertible encoding system. b: Further improvement arising from exploiting correlation in the forward pathway. c: Effects of exploiting correlation r in the forward pathway on pattern separation. Each curve shows how output overlap varies with input overlap for a particular value of r; a lower curve with a steeper rise toward the right signifies better separation. Note that in b, moderate correlation of inputs (r = 0.14) leads to near-optimal reconstruction, while in c, separation is only slightly worse for this case than for random forward projections (next-to-lowest and lowest curves, respectively).**

tational properties of the various components of the hippocampal system increases, and as we experience concomitant growth in the power of the available simulations, a time will come when there are specific issues that can only be addressed in a fuller model, with sufficient conceptual and practical resources. At this point we will be able to explore the sufficiency of the system as a whole and to examine the interactions between the hippocampal and neocortical systems in memory consolidation.

In the future we will also turn our attention to simulating the physiological data in more detail. While there is certainly a great deal more to be learned about the physiology, there are already several important findings that we will want our models to address. One domain of data that we particularly want to address is the set of findings related to the spatial and temporal modulation of hippocampal activity, and in particular the fascinating recent data suggesting a precession of place-cell activity within the theta cycle that anticipates the precession of place-cell activity as a function of the animal's locomotion in space (O'Keefe and Reece, 1993; Skaggs et al., 1996). These findings are among those that have been taken in support of the idea that the hippocampus contains intrinsic spatial reference frames, but we believe the data can be accounted for under the assumption that (a) the hippocampus receives spatial as well as other forms of information and (b) learns predictive associative relationships among representations of places in space based on general associative learning mechanisms.

DISCUSSION

New Knowledge About Hippocampal Function

1. There is a trade-off between the need for information-preserving, structure-extracting encoding of cortical traces and the need for effective storage and recall of arbitrary traces. Further, we can see how the areas of the hippocampus, entorhinal cortex and cortical association areas, and their interconnections could be mediating a solution to this tradeoff.
2. Long-term depression of synaptic strengths in the pathways subject to LTP (DG, CA1, and CA3) is crucial for maintaining pattern separation in face of the tendency to complete patterns which is enhanced by LTP.
3. The need to reduce or eliminate the influence of DG on CA3 during recall, suggested by Treves and Rolls (1992), is confirmed in this work, as is the idea that DG acts during storage to increase pattern separation in CA3.
4. Area CA1 must be able to exploit correlations in EC patterns in the direct perforant path synapses. This is required to optimize invertibility of the CA1 code.

Need for Explicit Modeling

The points mentioned above concerning our new knowledge are all due to the use of an explicit computational model rather than simpler verbal-qualitative reasoning. The analyses and simulations have established that our assertions about separability and invertibility arising from conjunctive and componential coding schemes are computationally valid. In the future, we will be able to show exactly how capacity varies with various parameters of the model. We hope to establish that capacity is sufficient for the putative memory role assigned to the hippocampal system, although we realize that the question of what is sufficient capacity is subject to considerable uncertainty. A more general accomplishment due to the use of the explicit model is that its analysis inspired the conceptualization of the hippocampal memory system as one that mediates the conflict between information preservation and arbitrary pattern storage.

Relationship to Other Models

These ideas overlap in many points with other proposals about the role of the hippocampus in learning and memory (e.g.,

McNaughton and Morris, 1987; Teyler and Discenna, 1986; Levy, 1989; Schmajuk and DiCarlo, 1992; Gluck and Myers, 1993; Alvarez and Squire, 1994; Treves and Rolls, 1994; Eichenbaum et al., 1994) and may be seen as a modernization and extension of the key insights offered in the early computational theory of hippocampal function presented by David Marr (1971). This and many of the subsequent mechanistic theories (McNaughton and Morris, 1987; Schmajuk and DiCarlo, 1992; Gluck and Myers, 1993; Treves and Rolls, 1994) stress the role of associative synaptic modification in learning and are consistent with the presence of associative forms of synaptic plasticity in many of the pathways into and within the hippocampal system (Bliss and Gardner-Medwin, 1973; Bliss and Lømo, 1973; McNaughton et al., 1978; Levy and Steward, 1979; Barrionuevo and Brown, 1983). The following specific comparisons and contrasts can be noted:

1. Hasselmo (e.g., Hasselmo and Schnell, 1994; Hasselmo et al., 1996) has a generally similar theory, but he does not examine capacity, invertibility, or larger-scale issues, relying on small-scale simulations to verify aspects of the theory not subject to analysis. On the other hand, his work does examine how the hippocampus might rely on the cholinergic system to switch between storage and recall modes.
2. A large number of investigators have adopted variants of the view that the hippocampus performs primarily a binding function, linking elements of memories across disparate brain regions. Murre's (this volume) theory and others in the binding tradition (Teyler and Discenna, 1986; Damasio, 1989; Alvarez and Squire, 1994) make assumptions very similar to those in our theory, although they assume that the computational role of the hippocampus is to link disparate elements of the same memory to each other across brain regions. As in our theory the hippocampus plays only a temporary role, serving to link disparate parts of a memory to each other until direct connections can be established. The main difference is that our theory claims that cortical connections learn slowly to allow discovery of structure and avoid catastrophic interference, while no such claim is made in the other theories. In contrast, Murre gives a very different motivation for slow learning of cortical connections: The reason is only that they are very sparse, because dense interconnectivity between cortical neurons in different areas would take up too much space in the brain to be feasible. Some of Murre's specific ideas, such as how one could have anterograde amnesia without retrograde amnesia, are quite compatible with our views, and the modulatory mechanisms he describes, although not emphasized in our theory, are consistent with it.
3. We are in debt to the work of McNaughton and Morris (1987), which examined the effects on pattern separation and completion of various coding schemes, and especially for the idea that significant information storage occurs in the feedforward pathways as well as the recurrent CA3 collaterals. We continue to hold the view that hippocampal representation of space is a byproduct of the use of a conjunctive coding, rather than a reflection of a specific commitment of the hippocampal system for spatial representation through prewired, intrinsic spatial maps in the hippocampus. In this regard our work diverges from some recent proposals arising in McNaughton's laboratory (Samsonovich and McNaughton, 1996), in which such intrinsic maps are postulated to account for a variety of phenomena related to hippocampal place fields. Our approach is in agreement with a number of other theorists writing in this volume (Gluck and Myers, 1996; Levy, 1996; Buhusi and Schmajuk, 1996; Murre, 1996; Hasselmo et al., 1996) and elsewhere (Treves and Rolls, 1994), who also treat the hippocampal system primarily as a memory system, and who attribute the apparent specialization of the hippocampal system for spatial tasks as reflecting its computational characteristics, rather than as a specialization for spatial information processing per se.

Experimental Directions

1. Our model requires that the direct perforant path from EC to CA1 can play a determining role in selecting which CA1 cells fire. This requirement arises from the functional role that we assign area CA1 and thus constitutes a prediction of our model. However, the prediction appears to be very difficult to test. Anatomically there is an extensive projection from EC_{III} to CA1. Nevertheless, despite a significant amount of experimental work, and a special issue of *Hippocampus* devoted to the pathway, (Vol 5, Number 2, pp. 101–146), it remains unclear to what extent it actually governs the firing of CA1 neurons in vivo.
2. Our analyses indicate that LTD in CA3 and CA1 is crucial for maintaining the ability of these areas to perform pattern separation. The data on LTD in vivo are sparse and in behaving animals are nonexistent: More of both types would be welcome.
3. Our model confirms the idea put forward by Treves and Rolls (1992) that DG should be more involved in CA3 activity during learning than during recall. Experimental data testing this hypothesis would be useful.
4. More data on firing patterns under behavioral conditions in all areas of hippocampus would be useful, but especially in EC where few studies have been made (Quirk et al., 1992). The representation in EC is of critical importance to the rest of the hippocampus since it is the primary input/output interface for the neocortex.

REFERENCES

Ackley DH, Hinton GE, Sejnowski TJ (1985) A learning algorithm for Boltzmann machines. Cognitive Sci 9:147–169.

Alvarez P, Squire LR (1994) Memory consolidation in the medial temporal lobe: a simple neural network model. Proc Nat Acad Sci USA 91:7041–7045.

Amaral DG, Ishizuka N, Claiborne B (1990) Neurons, numbers and the hippocampal network. Prog Brain Res 83:1–11.

Barnes CA, McNaughton BL, Mizumori SJY, Leonard BW, Lin L-H (1990) Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. Prog Brain Res 83:287–300.