From Space to Episodes: Modeling Memory Formation in the Hippocampal-neocortical System

by

Szabolcs Káli

M.Sc., Physics
Eötvös Loránd University, Budapest, Hungary, 1996

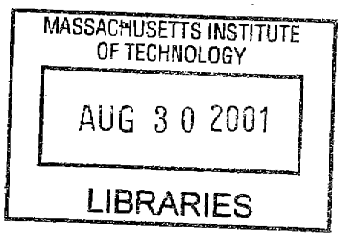
Submitted to the Department of Brain and Cognitive Sciences in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Computational Neuroscience

at the

Massachusetts Institute of Technology

September 2001



Signature of Author: _____

Department of Brain and Cognitive Sciences
August 6, 2001

Certified by: _____

Peter Dayan
Reader, Gatsby Computational Neuroscience Unit, University College London
Thesis Supervisor

Accepted by: _____

Earl K. Miller
Associate Professor of Neuroscience
Chairman, Department Graduate Committee

From Space to Episodes: Modeling Memory Formation in the Hippocampal-neocortical System

by

Szabolcs Káli

Submitted to the Department of Brain and Cognitive Sciences on August 6, 2001 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computational Neuroscience

ABSTRACT

This thesis describes the use of mathematical, statistical, and computational methods to analyze, in two paradigmatic areas, what the hippocampus and associated structures do, and how they do it.

The first model explores the formation of place fields in the hippocampus. This model is constrained by hippocampal anatomy and physiology and data on the effects of environmental manipulations on the place cell representation. It is based on an attractor network model of area CA3 in which recurrent interactions create place cell representations from location- and direction-specific activity in the entorhinal cortex, all under neuromodulatory influence. In unfamiliar environments, mossy fiber inputs impose activity patterns on CA3, and recurrent collaterals and perforant path inputs are subject to graded Hebbian plasticity. Attractors are thus sculpted in CA3, and are associated with entorhinal activity patterns. In familiar environments, place fields are controlled by the way that perforant path inputs select amongst the attractors.

Depending on training experience, the model generates place fields that are either directional or non-directional, and whose changes when the environment undergoes simple geometric transformations are in accordance with experimental data. Representations of multiple environments can be stored and recalled with little interference, and have the appropriate degrees of similarity in visually similar environments.

The second model provides a serious test of the consolidation theory of hippocampal-cortical interactions. The neocortical component of the model is a hierarchical network structure, whose primary goal is to extract statistical structure from its set of inputs through unsupervised learning. This interacts with a hippocampal component, which is capable of fast learning, cue-based recall, and off-line replay of stored patterns.

The model demonstrates the feasibility of hippocampally-dependent memory consolidation in a more general and realistic setting than earlier models. It reproduces basic characteristics of retrograde amnesia, together with some related phenomena such as repetition priming. The model clarifies the relationship between memory for general (semantic) and specific (episodic) information, suggesting that part of their underlying substrate may be shared. The model highlights some problematic aspects of consolidation theory, which need to be addressed by further experimental and theoretical studies.

Thesis Supervisor: Peter Dayan

Acknowledgments

First, I would like to thank my advisor, Peter Dayan, for supporting and guiding me throughout my graduate career in more ways than I can possibly enumerate. His high standards have always encouraged me not to settle for anything less than the best I was capable of. I am also very grateful to all (past and present) members of my thesis committee, Sue Corkin, Mike Hasselmo, Anthony Wagner, and Matt Wilson, for advice and stimulating discussion.

I have been lucky enough to be a member of not one, but two great research communities during these five years. I wish to thank all my colleagues both in the Department of Brain and Cognitive Sciences at MIT and in the Gatsby Computational Neuroscience Unit in London for their friendship and for broadening my scientific horizon. Special thanks to the people providing administrative and computational support at both places, who made sure that everything was running smoothly all the time.

I would like to acknowledge those organizations whose generous financial support made it possible for me pursue my graduate studies. In particular, I am thankful to the Department of Brain and Cognitive Sciences and the National Science Foundation for supporting me at MIT, and to the Gatsby Foundation for making possible my three-year stay at the Gatsby Computational Neuroscience Unit. The supplementary grant I received from the Soros Foundation is also gratefully acknowledged.

Last but not least, I would like to express my deep gratitude to those people who made these five years away from home a lot easier than it could have been; first, my family, my parents, Atti, and my grandparents, for supporting me in this endeavor even though, in some ways, it must have been more difficult for them than it was for me; Eszter, for creating a home away from home wherever we went; and lots of friends in Boston, London, and many other places around the world, for making my PhD years so much fun.

Contents

Chapter 1

Introduction

This thesis is concerned with the computational functions of the hippocampus and its associated structures. The hippocampus has long been a target for experimental studies, because of its relative anatomical simplicity and seductive regularity, its focal position as one of the most intensely multi-modal areas of the brain, the substantial evidence for its involvement in cognitively important functions, and a set of easily characterized disputes about what role it actually plays. For the very same reasons, and partly to help resolve the disputes, the hippocampus has also been the target of extensive theoretical study.

To date, most characterizations of the function of the hippocampus fall into three broad categories: memory, spatial processing and conjunctive coding. Based originally on neuropsychological data from humans, memorial theories posit that the hippocampus is responsible for storing certain classes of memories in such a way that they can be recalled based on partial or altered cues. The role of the hippocampus in memory may also involve a consolidation process, whereby reactivation of the memory traces stored in the hippocampus results in a gradual reorganization of neocortical memory representations. As a result of this consolidation, memories whose successful recall initially depends on the hippocampus may eventually become independent of the integrity of this structure, and also become integrated into the existing neocortical knowledge base.

Spatial processing theories were originally based on behavioral and selective lesion data from a variety of animal species and, most spectacularly, neurophysiological data from rats on the existence in the hippocampus of neurons (called place cells) whose activity signals the animal's position in space. These theories hold that the hippocampus plays a central

role in manipulating information about space, potentially implementing or supporting key operations such as the construction and use of internal maps of the layout and contents of the external world.

Finally, conjunctive coding theories, based originally on data from animal conditioning, suggest a particular role for the hippocampus in creating, and perhaps teaching the neocortex about, unitary codes to represent complex combinations of potentially multi-modal stimuli.

The approach taken in this thesis is that there is a kernel of truth in all these theories. Thus, we have sought to explore the computational relationships among them, and to integrate them through a consideration of the structure and physiology of the hippocampal formation. Anticipating the main findings of the thesis, our results can be briefly summarized as follows. First, we have found that the action of the autoassociative recurrent network of area CA3, which lies at the core of memory operations, combined with the assumed pattern separation capability of the dentate gyrus, which can support the creation of conjunctive representations, naturally leads to the emergence of place-cell-like activity patterns when applied to the sorts of sensory inputs available during spatial behavior. Second, application of this same basic set of processes, which also includes a neuromodulatory switch between storage and retrieval modes within the hippocampus, to a more general class of inputs, combined with an analysis of the basic anatomical and representational properties of neocortex, allows the construction of an explicit model of memory consolidation.

The work described in this thesis builds on much of the earlier experimental and theoretical research, and therefore the next chapter is intended to provide a selective review of the work that served as the basis for our investigations. Chapter 3 describes our model of the place cells in area CA3 of the rodent hippocampus, except for some early work on a simple, analytically solvable version of the model, which has been separated into an appendix in order to preserve the coherence of the main body of the work. Then, Chapter 4 considers the interactions between the hippocampus and neocortex in the domain of declarative memory, and introduces a network model of memory consolidation. Finally, in Chapter 5, I draw some general conclusions based on both lines of research, and present an integrated view of the function of the hippocampus.

Chapter 2

Experimental background

I begin the review by introducing the main classes of tasks in which the hippocampus appears to be important. Next, the internal structure and external connections of the hippocampus will be described. I will then summarize the physiological correlates of (spatial and nonspatial) behavior both on the single unit and the population level, as well as the basic facts about plasticity and neuromodulation in the hippocampal system. Then, Section 2.4 deals with the behavioral consequences of selective lesions and other manipulations both in experimental animals and in human subjects. This section also includes a brief description of common experimental techniques, as well as a selective review of human neuropsychology, concentrating on studies of amnesic patients, but also touching on other topics, including an overview of relevant functional neuroimaging data.

2.1 Characterizing the function of the hippocampus

In this section I introduce the main classes of tasks in which normal performance has been suggested to depend on the normal functioning of the hippocampal system, and provide a brief analysis of the key issues involved in understanding the hippocampal contribution to solving each class of problems. As already mentioned in the Introduction, the three main classes of hippocampally-dependent tasks, also corresponding to the three most influential types of theories of hippocampal function, may be described under the headings "memory", "space", and "conjunctive coding".

2.1.1 Memory

It has been known since the 1950's that damage to the hippocampus and other areas of the temporal lobe can lead to severe impairment of certain memory functions in humans. The patient H.M., who underwent bilateral temporal lobectomy to treat his epilepsy (Scoville and Milner, 1957), is probably the best-known example of this (for a review, see Corkin, 1984), although his lesions have proved to be quite complex, sparing some of the hippocampus but affecting (at least partially) many of the surrounding areas (Corkin et al., 1997). More recently, amnesic patients with more circumscribed lesions (in the medial temporal lobes or the midline diencephalon) have also been identified (see e.g. Rempel-Clower et al., 1996; Buffalo et al., 1998; Aggleton and Saunders, 1997).

Amnesia can be defined in terms of a set of positive and negative findings (Mayes and Downes, 1997). First, amnesics show impaired recall and recognition of personally experienced episodes and general facts that they have been exposed to since the onset of their amnesia – this is referred to as anterograde amnesia (AA). Second, amnesics also show impaired recall and recognition of personal episodes and general facts that they had encountered before they suffered brain damage – this is termed retrograde amnesia (RA). The kinds of memories that are affected in amnesia are called "declarative" or "explicit" memories, since they mostly involve material that can be declared verbally (and subjects are are also aware of the contents of the memory when it is recalled). Third, those types of memories that do not involve long-term retention of facts or events are intact in amnesics. In particular, their short-term or working memory is unimpaired, as are types of memories usually referred to as "implicit", such as skill memories (including the ability to acquire new skills) and lasting perceptual fluency effects such as repetition priming. Finally, measures of general intelligence are also normal in amnesics.

A lot of work has been done on developing animal models of amnesia. Although it has been difficult to identify tasks which might tap into the memories corresponding to human episodic or fact memory in other animal species (indeed, even the existence of such equivalents is heavily debated), several memory tasks have been described in various species with some degree of hippocampal dependence.

In computational terms, these memory tasks require the subject to act based on information about past experiences retained over a relatively long time period (typically beyond the reach of activity-based working memory). The behavioral expression of such long-term memory involves a sequence of several, potentially dissociable processes. First of all, in-

formation has to be encoded in a form that is independent of ongoing neuronal activity, presumably as a lasting change in the efficacy of specific synapses. Second, this information needs to be stored throughout the time between learning and recall; this probably involves some form(s) of maintenance, partly to reduce interference from subsequently stored memories. Finally, for memory retrieval, the information stored in the form of synaptic efficacies must be translated back into neuronal activity patterns, also making sure that the reinstantiated pattern is one that is appropriate for the current retrieval context.

Starting from the classic study by Marr (1971), several models have investigated how these functions might be implemented in the hippocampus; however, different models have stressed different aspects of the problem. In particular, the presence of an extensive recurrent network in area CA3 naturally led to the suggestion that these recurrent connections are the substrate where hippocampal memories are stored, and several models have investigated how area CA3 may effect the storage and retrieval of arbitrary static patterns (Marr, 1971; McNaughton and Morris, 1987; Hasselmo et al., 1995) or temporal sequences of patterns (Levy, 1996; August and Levy, 1999). Other studies have concentrated on assigning a role to different areas and pathways within the hippocampus (Treves and Rolls, 1992, 1994; O'Reilly and McClelland, 1994; Hasselmo et al., 1996; Lisman, 1999).

There is substantial evidence that, in addition to encoding, storage, and retrieval, long-term declarative memory in the hippocampal-neocortical system may also involve a fourth basic process known as consolidation. The basic observation that indicates the existence of such a process is that recent memories are more likely to be lost than old memories after circumscribed lesions within the medial temporal lobes. This has generally been interpreted as evidence that the role of these MTL structures in the storage and/or retrieval of declarative memories is only temporary. In particular, several investigators have advocated the general idea that, in the course of a relatively long time period (from several days in rats up to decades in humans), memories are reorganized (or *consolidated*) so that memories whose successful recall initially depends on the hippocampus gradually become independent of this structure (Squire, 1992; Alvarez and Squire, 1994; McClelland et al., 1995; Murre, 1997). However, other possible interpretations of the data have also been proposed (Nadel and Moscovitch, 1997; Nadel et al., 2000).

The existence of these multiple processes involved in long-term memory, at least some of which are temporally dissociable, means that the timing of a hippocampal manipulation relative to the time of learning and the time of recall may qualitatively affect the behavioral consequence of the manipulation. The distinction between retrograde effects (when learn-

ing occurs before the onset of the manipulation) and anterograde effects (when training begins after the onset of the manipulation) has already been mentioned, but at least some techniques also allow other types of dissociations.

It has proved difficult, especially in animals, to clearly delimit the kinds of memory tasks in which hippocampal lesions cause deficits. When characterizing a particular task, an important distinction that should be made is between the nature of the information that needs to be retained between training and testing, and the way in which this information is utilized during testing. Although these task characteristics are not perfectly independent (since different kinds of information can be accessed in different ways), both of them can be crucial determinants of whether a given structure is necessary for the task, particularly if that structure is thought to be involved in retrieval. As an important example, retention of most types of declarative information in humans can be tested in either recall-based or recognition-based paradigms, which seem to be supported by at least partly separate anatomical areas.

Finally, since it is not clear whether amnesia caused by damage to different brain regions has the same characteristics, and, as a related question, whether these different regions form a single or multiple functional units, the exact location of the lesion (or other local manipulation) is another important variable. All of these characteristics, namely, the timing of the experimental manipulation, the computational nature of the task (including the information to be retained and the way it is used during testing), and the location of the manipulation (or the fact that it affects multiple regions) are critical in correctly assessing the consequences of the resulting performance for hippocampal involvement in memory, and even some additional factors, such as the species studied, may have a crucial effect.

2.1.2 Space

The discovery of "place cells" in the rat hippocampus 30 years ago (O'Keefe and Dostrovsky, 1971) marked the beginning of the second basic line of research into hippocampal function. Based on these physiological data, along with the results of early behavioral experiments on navigation, O'Keefe and Nadel (1978) proposed their "cognitive mapping" theory of hippocampal function. According to this theory, the hippocampus is crucially involved in the formation of a representation of the animal's surroundings, and the use of this representation for solving navigation tasks. Subsequent lesion studies generally confirmed this

spatial role of the hippocampus, particularly in tasks which require the animal to identify or remember specific locations based on the spatial configuration of multiple cues.

However, in order to understand the role of the hippocampus in spatial behavior, we first need to understand the computational nature of spatial tasks. A large number of detailed analyses have appeared on this topic, both classic and more recent (e.g., Tolman, 1948; O'Keefe and Nadel, 1978; Gallistel, 1990; Trullier et al., 1997; Foster, 1999; Redish, 1999). Therefore, I will only provide a brief overview here, especially since this thesis is not concerned with navigation itself (although it does address the formation of the CA3 place cell representation, which may in turn be a crucial component of the rat navigation system).

The goal of spatial learning and behavior may be defined as the ability to move around in space in a way that is most beneficial to the agent (be it an animal, human, or robot). The most trivial example is the ability to take the shortest possible path to a(n appropriately chosen) goal, but the optimal behavior can be very different in other situations, such as taking the safest route instead of the shortest, or even choosing a trajectory that allows the exploration of the largest possible area in a given time. Even taking the shortest route to a goal can present different challenges depending on the environment; for instance, finding the way back to the nest for a desert ant or a pigeon involves very different computational problems from a rat finding a goal in a network of underground tunnels or a complex artificial maze. Accordingly, a number of different strategies for solving navigation tasks have been identified which are appropriate in particular circumstances (see the reviews cited above for details).

Since solving such very different navigation problems requires different computations, and these computations can be decomposed into different sets of subtasks, it is not surprising that, in mammals, different brain areas appear to be necessary for different types of navigation tasks. In particular, the basal ganglia (and more specifically the caudate nucleus) seem to be essential for tasks involving specific stimulus-response associations (such as approaching a visible cue independent of its spatial location). On the other hand, the hippocampus appears to play a crucial role when locations (including the location of the animal itself and the location of the goal) have to be inferred from the constellation of multiple cues. This observation supports the cognitive mapping theory of O'Keefe and Nadel (1978), according to which the hippocampus is the seat of the "cognitive map" which represents the spatial relationships between objects in the world (and the animal itself) and allows the animal to plan routes in the environment that satisfy particular constraints.

However, it should be noted that navigation to a goal based on the configuration of multiple cues involves several separable computations with various degrees of interdependence. First, a way of representing one's location and bearing with respect to a reference location and direction is required. Second, one needs to be able to initialize and update this representation of his location and direction based on various sources of positional information, including local and distant external (visual, auditory, tactile, and olfactory) cues, as well as self-motion information (in which case updates can be based on the process known as path integration). Third, the ability to represent the spatial location of objects in the world (such as goals, routes, obstacles) is also required. Finally, one needs to be able to combine information about his own location and direction with information about the locations of relevant objects to plan a route (or at least decide locally in which direction to move).

It is important to note that the above processes do not have to be as separate as the above description might have suggested; any particular brain area may participate in several of these computational subtasks, and, conversely, both physiological and lesion studies suggest that any of these computations probably involves a distributed network of brain regions. In addition, the representation of the spatial relationships between relevant objects and one's own body does not have to be exactly map-like as long as the representations allow the computation of a locally optimal action. An example of how globally optimal solutions to navigation problems can be obtained using local computations is the reinforcement learning approach to navigation (see Foster et al., 2000).

Thus, the main difficulty in understanding the role of the hippocampus in spatial behavior is the complexity of even idealized navigation tasks. In addition, real tasks can often be solved using multiple strategies which involve different computations, and it can be very difficult to tell what the computations are that actually allow a particular animal to solve a particular task. Therefore, it is hard to tell from the pattern of impairments on various tasks following lesions or other manipulations of specific brain regions what the contributions of the affected regions to the particular computations might have been, and it becomes essential to analyze the possible strategies for solving the task, as well as the underlying sources of information and how these need to be processed. Accordingly, several computational studies have examined how spatial representations in areas known to be important for spatial processing (and particularly in the hippocampus) may develop (Zipser, 1985; Sharp, 1991; Touretzky and Redish, 1996; Zhang, 1996; Burgess et al., 1997; Samsonovich and McNaughton, 1997; Battaglia and Treves, 1998; Brunel and Trullier, 1998; Arleo and Gerstner, 2000; Doboli et al., 2000; Hartley et al., 2000), and how these representations might be used

in navigation (Blum and Abbott, 1996; Burgess and O'Keefe, 1996; Muller et al., 1996b; Redish and Touretzky, 1997; Arleo and Gerstner, 2000; Foster et al., 2000). In addition, due to the large number of areas and even larger number of interconnections throughout the brain that might contribute to the solution of navigation tasks, one needs to take special care with the interpretation of the results of even supposedly local manipulations (such as controlled lesions).

2.1.3 Conjunctive coding

In addition to memorial and spatial theories, there exists a third group of characterizations of hippocampal function. These theories are based mainly on conditioning experiments, which in turn were motivated by the realization discussed above, namely, that navigation and (to some extent) memory tasks are computationally too complex to allow any direct conclusions about the functions of specific brain regions. In addition, some results from those experiments indicated that the internal representations used by animals with hippocampal lesions might be fundamentally different from those used by normals. This suggested that the medial temporal region (and more specifically the hippocampus itself) might be involved in determining the kinds of representations formed and/or the way they can be used in solving particular problems.

In support of this notion, it was noted a long time ago that the hippocampus is required for the normal acquisition of tasks with a certain kind and degree of representational complexity. As an example, take the negative patterning problem, which requires the animal to learn that even though both stimulus A and stimulus B individually predict reward, the simultaneous appearance of both A and B predicts no reward. This is logically an exclusive OR problem, usually implemented as an operant conditioning task. It is also the prototype for non-linear discrimination problems, since it cannot be solved by linearly combining associative strengths assigned to individual stimuli. Rudy and Sutherland (1989) reported that rats with lesions to the hippocampus were unable to learn the negative patterning problem, along with a group of other tasks that are procedurally diverse but invariably require non-linear discrimination. It is tempting to assume on theoretical grounds that the crucial element that makes hippocampal involvement critical is the nonlinearity of the task, and this assumption was indeed at the core of Rudy and Sutherland's (1989) configural association theory. Here the key observation is that nonlinear discrimination problems may actually be solved using a linear decision-making procedure, provided not only individual stimuli, but

also combinations of stimuli (such as the conjunction of A and B) are allowed to acquire associative strengths which are independent of those of individual stimuli. Therefore, if the hippocampus were critical for establishing representations of stimulus conjunctions, we might expect hippocampal lesions to cause deficits in learning nonlinear discriminations.

Despite its considerable aesthetic appeal, this simple theory turned out to be wrong, as it was not supported by subsequent experiments. However, the general idea, namely, that an important function of the hippocampus may be to develop (or participate in the development of) unitary representations of relevant stimulus conjunctions, is a central assumption in several recent theories of hippocampal function. For example, the model of O'Reilly and Rudy (2001), which has its roots in an updated version of configural association theory (Rudy and Sutherland, 1995) and in McClelland et al.'s (1995) computational account of the interactions between hippocampus and neocortex, predicts that the hippocampus is required for tasks involving fast, incidental learning of stimulus conjunctions. A more specific, and prevalent idea is that the dentate gyrus, the first stage of hippocampal processing, is responsible for establishing unique representations for arbitrary configurations of stimuli (although its input area, entorhinal cortex, may already be able to represent stimulus conjunctions, which would subsequently be "orthogonalized", or made more unique, by the dentate gyrus). In contrast, some other theories such as (Gluck and Myers, 1993) assign a more general role to the hippocampus in working out optimal representations in associative learning tasks.

2.2 Anatomy

The hippocampal formation is located inside the medial temporal lobes of the cerebral cortex. It can be divided into several subregions based on cell types and patterns of connectivity. The major subdivisions are the dentate gyrus, areas CA3 and CA1, and the subicular complex (which can be further divided into the subiculum, the presubiculum, and the parasubiculum). For simplicity, all these structures will together be referred to as the "hippocampus". These areas are heavily interconnected, and also receive inputs from and send outputs to a number of neocortical areas and many subcortical structures (see Figure 2.1).

The predominant source of input to the hippocampus comes from entorhinal cortex (EC), which, in turn, receives multimodal sensory information from a large number of cortical

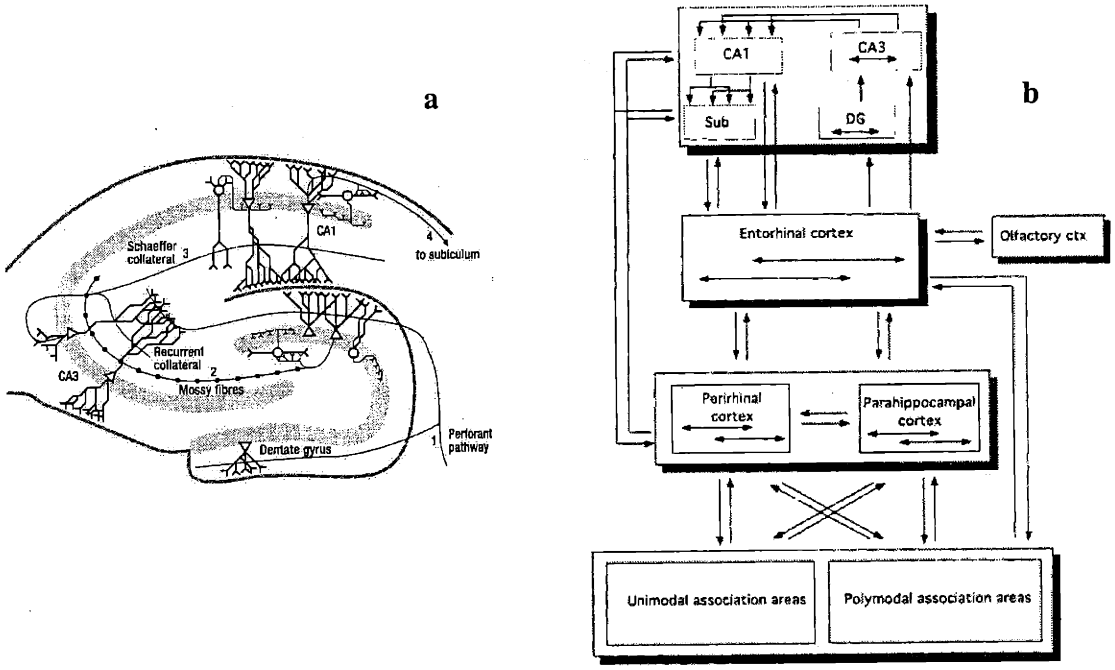


Figure 2.1: Anatomy of the hippocampal system. **(a)** Schematic drawing of a transverse slice through the hippocampus, showing the main areas, cell types, and connections (adapted from McLeod et al. (1998, p. 282)). **(b)** Schematic diagram of the major pathways within the medial temporal lobe. Two-headed horizontal arrows within areas indicate local associational projections. (Adapted from Fig. 3B of Lavenex and Amaral (2000).)

areas, either directly or through the perirhinal and parahippocampal cortices (Lavenex and Amaral, 2000). Entorhinal cortex is a special form of six-layered neocortex that can be divided into three superficial and three deep layers which are separated from each other. The superficial layers give rise to two separate sets of connections to the hippocampus, which are collectively referred to as the perforant path (Amaral, 1993); axons originating in layer II of EC terminate in the dentate gyrus and area CA3, while layer III of EC projects to CA1 and the subiculum. The mossy fibers form a set of connections from the dentate gyrus to CA3, and the pyramidal cells in CA3 project to CA1 via the Schaffer collaterals (Amaral and Witter, 1989). CA1 in turn projects back to deep layers of entorhinal cortex both directly and through cells in the subiculum. Therefore, the hippocampal circuitry contains a major loop from entorhinal cortex, through all major subfields of the hippocampus, finally back to entorhinal cortex (the deep layers of EC project to the superficial layers to make the loop complete). There are also strong cortical backprojections from entorhinal cortex to the perihinal and parahippocampal cortices, and from there to many of the same association

areas that provide the inputs to the medial temporal lobe (Lavenex and Amaral, 2000).

The individual subregions of the hippocampus (particularly the dentate gyrus, CA3, and the subicular complex) all have interesting and complex internal circuitry (Amaral and Witter, 1989), but most of this complexity is beyond the scope of this review. However, since much of the modeling work that we have done on hippocampal place cells assumes area CA3 as its anatomical substrate, I will describe the architecture of CA3 in somewhat more detail. The main distinctive feature of CA3 is its extensive set of recurrent connections, that is, the large number of connections between cells within this area. In the rat, there are about 300,000 pyramidal cells in CA3 in each hippocampus, and every pyramidal cell in CA3 contacts about 2% of all other CA3 neurons on both sides of the brain, which means that each of them receives about 12,000 synaptic inputs from the others (Amaral et al., 1990; Ishizuka et al., 1990; Li et al., 1994). An average CA3 pyramidal cell also receives about 4000 direct perforant path inputs from entorhinal cortex. In contrast, there are only 50 or so mossy fiber inputs to each CA3 neuron (Acsády et al., 1998), and even though these synapses are believed to have considerably higher efficacy than the previous two types of synapses (Yamamoto, 1982; McNaughton and Morris, 1987), CA3 neurons should primarily be driven by their recurrent and perforant path inputs unless these synapses are considerably suppressed. The proportions of different connections are quite similar in man, but cell numbers are approximately 20 times larger (Seress, 1988).

In addition to its main output through entorhinal cortex, the hippocampus (and CA1 and the subiculum in particular) also sends connections to other cortical areas (including frontal and cingulate cortices) (Irle and Markowitsch, 1982; Thierry et al., 2000; Wyss and Van Groen, 1992) and the diencephalon (including the mammillary bodies and anterior thalamic nuclei). In fact, it has been suggested that the hippocampus, the medial septum, the anterior thalamus, and the mammillary bodies, perhaps along with some other regions (such as retrosplenial cortex and entorhinal cortex), may form a single functional circuit (Aggleton and Saunders, 1997). The evidence for this hypothesis comes partly from the large number of anatomical connections that exist between these areas, and partly from the similarities of the functional (behavioral) deficits that result from lesions to these structures.

The hippocampus also receives extensive innervation from a number of subcortical neuromodulatory centers. This includes cholinergic and GABAergic input from the medial septum and the diagonal band of Broca, a noradrenergic projection from the locus ceruleus, and serotonergic input from the raphe nuclei (Vizi and Kiss, 1998). As we will see later, these neuromodulatory inputs, and in particular the septal projection via the fornix, are

thought to play an important role in the proper functioning of the hippocampus, perhaps by enabling switching between different global dynamic states of the hippocampal network. The CA1 region of the hippocampus also feeds back to the septum, via both excitatory and inhibitory connections (Alonso and Kohler, 1982).

In most of the hippocampal areas the principal cell type is the pyramidal cell, except in the dentate gyrus, where the main cell type is the granule cell (Amaral et al., 1990). All areas also have substantial and diverse populations of inhibitory interneurons (Freund and Buzsáki, 1996), which play important roles in regulating the activity of the excitatory neural population. Interneurons also appear to be preferentially targeted by some neuromodulatory inputs, including the serotonergic and the septal GABAergic projection (Vizi and Kiss, 1998).

2.3 Physiology of the hippocampal region

2.3.1 Behavioral correlates of neural activity

Besides controlled lesion studies (which will be considered in later sections), probably the best source of information regarding the function of specific brain regions is the measurement of single cell activity in controlled conditions *in vivo*, in either anesthetized or awake and behaving animals. The usual method is to monitor the spiking activity of a single cell or a collection of cells, and look for correlations between the neural activity and the sensory stimulation and/or behavioral response of the animal. The aspect of the stimulus or behavior that is most correlated with the activity of the cell is then said to be 'encoded' by that neuron.

A word of caution is probably appropriate here. First of all, experiments generally vary conditions only along a very restricted, heuristically chosen set of dimensions in the stimulus space. Therefore, it is usually quite possible that the neuron actually cares about a thus far unexplored variable that just happens to be correlated with one included in the analysis. Secondly, even if we could identify the variables that are truly critical in determining the activity of the neuron, we would not really be justified in concluding that the region containing this neuron is necessarily critical for all computations in the brain that involve this quantity, although it would be probably safe to assume that the region is involved in some such computation. Nevertheless, these recordings are obviously useful tools in exploring brain function, especially employed in conjunction with other techniques.

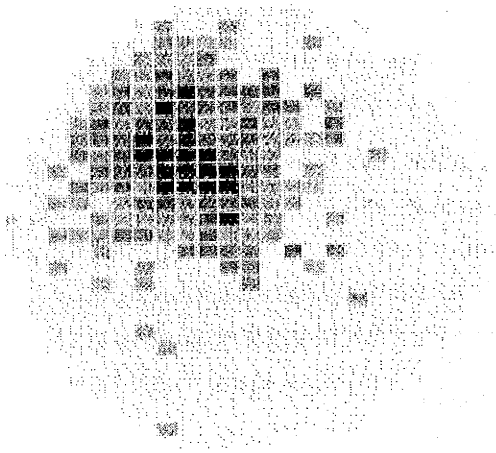


Figure 2.2: Hippocampal place field. The figure shows the firing rate of a CA1 pyramidal cell inside a circular arena. The grayscale indicates the number of spikes fired when the animal was inside a particular square bin during the recording session. Dark shading corresponds to high rates and light shading to low rates. (Adapted from Muller et al. (1991a).)

Spatial representations in the hippocampus and beyond

One of the experimental findings which had the most influence on thinking about the hippocampus was the discovery of place cells by O'Keefe and Dostrovsky (1971). They found that principal neurons in areas CA3 and CA1 of the rat hippocampus are active only when the animal is located in a well-defined local region of the environment (a place field; O'Keefe, 1976; Muller et al., 1987). An example of such a place field (recorded in a circular arena) is shown in Figure 2.2. In the three decades since the original study, the characteristics of the place cell representation have been mapped out in considerable detail (for reviews, see e.g. O'Keefe and Nadel, 1978; Muller, 1996; Redish, 1999), and neurons with place-cell-like properties have been described in several other brain regions, including the dentate gyrus (Jung and McNaughton, 1993), entorhinal cortex (Barnes et al., 1990; Mizumori et al., 1992; Quirk et al., 1992; Frank et al., 2000), subiculum (Barnes et al., 1990; Sharp and Green, 1994; Sharp, 1997), parasubiculum (Taube, 1995b), and even outside the medial temporal region, in the dorsal striatum (Wiener, 1993; Mizumori et al., 2000).

According to different estimates, about 10-50% of all pyramidal cells in CA3 and CA1 have

place fields in a typical environment (Thompson and Best, 1990; Wilson and McNaughton, 1993; Muller, 1996). Different place cells tend to have different place fields; in fact, the location of place fields appears to be random, and roughly independent of the anatomical location of the cell (O'Keefe and Conway, 1978; Muller and Kubie, 1987). The place fields collectively cover the environment, and the place cells that are active in a given environment together provide a population code for spatial location (Wilson and McNaughton, 1993). Place fields in very different environments appear to be completely unrelated; whether a given cell has a place field in the environment at all is independent for different cells, and the spatial relationships in two environments between place fields of cells that have place fields in both seem to be unrelated (O'Keefe and Conway, 1978; Muller and Kubie, 1987; Wilson and McNaughton, 1993).

The typical maximum firing rate of a place cell within its place field is about 5-40 Hz, and the firing rate outside the place field is less than 1 Hz, but the average rate tends to vary smoothly in space (Muller et al., 1987). Some place cells have multiple place fields in some environments, but it seems that, for most purposes, such cells can be thought of as multiple independent place cells. Place cells in CA3 and CA1 have very similar properties under normal circumstances (Muller et al., 1987); however, they are affected differentially by certain lesions and pharmacological manipulations. Hippocampal interneurons are also spatially selective, although their spatial selectivity is much lower than that of place cells (Kubie et al., 1990).

In open field environments, the activity of place cells tends to be independent of which way the animal faces (although place cells can also be directional in other conditions; McNaughton et al., 1983; Muller et al., 1994; Markus et al., 1995). In addition, removal of some of the available cues does not tend to affect the place cell representation. These observations (among others) indicate that hippocampal place cells cannot be appropriately regarded as higher order sensory neurons. However, salient visual cues do have a strong influence: if the single cue card in an otherwise featureless circular arena is rotated, all place fields follow this rotation (Muller and Kubie, 1987). In the absence of reliable visual cues, the firing of place cells can also be controlled by the animal's path integration system, a brain system which can update the representation of the animal's location based on self-motion information (Quirk et al., 1990; Knierim et al., 1995; Sharp et al., 1995; Jeffery and O'Keefe, 1999).

In a novel environment, the firing of place cells appears to be location-specific from the moment when the animal enters the environment (Hill, 1978). However, it may take several

minutes or even hours for the place cell representation to become reliable and stable, and it seems to require active exploration (Wilson and McNaughton, 1993). After this initial exploration period, place fields remain stable and robust for a long time (at least several days or weeks), even if the animal is absent from that particular environment for most of the intervening time (Thompson and Best, 1990).

Now let me return briefly to neurons with spatially selective activity outside the hippocampus proper. Dentate granule cells have spatial selectivity comparable to place cells, and they were also found to be strongly selective to head direction – however, data only exist for linearly restricted environments (Jung and McNaughton, 1993). Principal neurons in both superficial and deep layers of entorhinal cortex also fire in a location-specific manner, but their spatial specificity is much weaker than that of place cells (Barnes et al., 1990; Quirk et al., 1992; Frank et al., 2000). Also, their activity seems to be more tightly bound to sensory details of the environment than the activity of place cells (Quirk et al., 1992), so that different environments that share some features lead to related (but different) activity patterns in EC. This property is shared by subicular cells (Sharp, 1997), which are otherwise more similar to place cells in their spatial selectivity, although the tuning tends to be considerably wider than that of place cells (Sharp and Green, 1994).

Head direction cells are closely related to place cells both anatomically and functionally (for reviews, see Muller et al., 1996a; Taube, 1998). Instead of being sensitive to location, these neurons signal which way the animal faces relative to some internal reference direction, independent of the location of the animal. Such cells have been found in the postsubiculum (Taube et al., 1990a,b), the anterior dorsal nucleus of the thalamus (Blair and Sharp, 1995; Taube, 1995a), the lateral mammillary nuclei (Stackman and Taube, 1998), as well as several other brain regions. Unlike the relative preferred locations of place cells, the relative preferred directions of head direction cells seem to be the same in every environment (Taube et al., 1990b). Therefore, the head direction cells collectively implement an internal compass for the animal. Environmental manipulations like cue rotation tend to affect head direction cells and place cells in the same way and in close temporal synchrony, indicating that the two systems are strongly coupled (Knierim et al., 1995, 1998).

All the physiological data I have presented so far were recorded from a single species, namely the rat. Much less detailed information is available from other species. Place cells have been found in the mouse hippocampus; they were found to be noisier but otherwise similar to rat place cells (McHugh et al., 1996; Rotenberg et al., 1996). In the monkey, Rolls and colleagues have described what they call 'spatial view' cells in the hippocampus; instead

of signalling the animal's location, these neurons appear to provide an allocentric code for locations elsewhere, and particularly for the position in space that the monkey is looking at (Rolls and O'Mara, 1995; Georges-Francois et al., 1999). Many of the cells continued to respond to the animal looking towards a particular location, even when the view was obscured by a curtain or in complete darkness (Robertson et al., 1998).

Nonspatial correlates of unit activity

While proponents of the cognitive map theory suggest that the primary and perhaps the only role of the hippocampus is the processing and storage of spatial information (O'Keefe and Nadel, 1978; O'Keefe, 1999), others maintain that the hippocampus is critical for the processing of certain types of nonspatial information as well (Cohen and Eichenbaum, 1993; Eichenbaum et al., 1999). Supporting the latter view, several experiments have identified nonspatial behavioral correlates of hippocampal neuronal activity.

As an example, hippocampal neurons respond specifically to different relevant stimuli during delay and trace eyeblink conditioning in the rabbit (Berger et al., 1976, 1983; McEchron and Disterhoft, 1997). Early in training, hippocampal cells preferentially respond to intrinsically behaviorally relevant stimuli (the airpuff). Over the course of training, neurons start to respond to the conditioned stimulus (a tone), even before (and perhaps unrelated to) any behavioral expression of learning.

Hippocampal neurons are also differentially activated in various (visual, auditory, and olfactory) discrimination tasks (e.g., Wiener et al., 1989; Sakurai, 1996) and delayed recognition memory tasks (e.g., Wible et al., 1986; Otto and Eichenbaum, 1992b; Deadwyler et al., 1996). In a recent study by Wood et al. (1999), rats performed a recognition memory task at several distinct locations while the activity of a large number of hippocampal cells was recorded. Some of these cells behaved like conventional place cells; however, a large proportion of cells was selective for sensory stimuli (odor identity), behavioral events (approaching a food cup), or cognitive state (whether reward was expected), irrespective of the spatial location of the animal.

Much less work has been done in elucidating the response properties of neurons in entorhinal, perirhinal, and parahippocampal (in the monkey) or postrhinal (in the rat) cortices. The available data have been recently summarized by Suzuki and Eichenbaum (2000). Many cells are selective for different sensory stimuli. In addition, some cells maintain higher levels of

firing after the stimulus is removed, indicating some kind of memory representation (Young et al., 1997). Finally, some cells (particularly in perirhinal cortex) display stimulus-selective suppression or enhancement of activity when a previously encountered stimulus is presented again (Young et al., 1997; Brown and Xiang, 1998). This phenomenon has been suggested to support behavioral judgements of familiarity and recency.

2.3.2 Neuronal population behavior

The collective behavior of large localized populations of neurons may be characterized by the electro-encephalogram (EEG). The hippocampus exhibits some characteristic EEG patterns with strong behavioral correlates (for a review, see Buzsáki, 1996). One of these patterns is the theta rhythm, which is a periodic signal with frequency in the 4-12 Hz range in rodents and rabbits. It occurs during locomotion, sensory stimulation, and rapid-eye-movement sleep. It can be recorded in all areas of the hippocampal formation as well as in entorhinal cortex. Theta rhythm in these areas appears to be driven by a similar signal in the septal nuclei, although the hippocampal circuitry has been shown to be able to generate rhythmic firing in the theta range intrinsically under certain conditions (Leung, 1998). Most of the cells fire nonrhythmically during theta waves; however, their firing is correlated with theta. Place cells show an additional interesting effect: as the animal moves through its place field, the cell fires at earlier and earlier phases of the theta cycle (O'Keefe and Recce, 1993; Skaggs et al., 1996). The magnitude of this so-called phase precession effect can be up to 300° as the animal traverses the field, and no fully satisfactory explanation of the phenomenon has been offered. Theta rhythmicity is often accompanied by oscillations in the gamma (40-100 Hz) frequency range, whose amplitude is modulated by the theta rhythm (Bragin et al., 1995).

In the absence of theta, the hippocampal EEG is characterized by short (40-120 ms), irregular bursts of high frequency (around 200 Hz) activity which are called sharp waves (Buzsáki, 1986, 1989). They occur during eating, drinking, grooming, awake immobility and slow-wave sleep. During sharp waves, groups of pyramidal cells in CA3, CA1, subiculum, and entorhinal cortex fire in synchronous bursts. This event is initiated within the CA3 recurrent network, and spreads through complex interactions between interneurons and pyramidal cells at later stages (Buzsáki, 1996; Chrobak and Buzsáki, 1998). During sharp waves, neurons in the deep layers of entorhinal cortex discharge high frequency (140-200 Hz) volleys. This activity was recently found to be correlated with neural activity in

medial prefrontal cortex (Siapas and Wilson, 1998) and perirhinal cortex (Collins et al., 1999), both at the level of the EEG and at the level of single units.

It has been noted that sharp waves seem to provide ideal physiological conditions for the modification of synaptic connections (particularly the induction of long-term potentiation (LTP), which will be described later in this section). This and other observations have led Buzsáki (1989, 1996) to propose a two-stage model of memory trace formation. According to this theory, weak memory traces formed within the hippocampus during exploratory behavior (theta activity) are made permanent during sharp-wave activity. It also suggests that the synchronized bursts associated with sharp-wave activity may also induce LTP in target areas in neocortex, and therefore it provides a physiological framework for the idea of memory consolidation.

In support of this theory, there is now considerable evidence for the reactivation of memory traces during sleep. Pavlides and Winson (1989) found that CA1 place cells which were active during awake experience showed increased activity during subsequent sleep compared to previous sleep episodes. Evidence that full hippocampal memory representations are reinstated during sleep comes from the observations that both spatial (Wilson and McNaughton, 1994) and temporal (Skaggs and McNaughton, 1996) correlations of activity between pairs of cells during behavior are substantially preserved during subsequent slow wave sleep. Correlations were strongest during EEG ripples (Wilson and McNaughton, 1994; Kudrimoti et al., 1999), and decayed with a time constant of approximately 10 minutes. Two different awake experiences appeared to have independent effects on subsequent sleep activity patterns (Kudrimoti et al., 1999). Recently, evidence for the reactivation of hippocampal ensemble patterns during rapid eye movement (REM) sleep has also been obtained (Louie and Wilson, 2001). Finally, a similar increase attributed to awake experience was found in correlations during sleep between pairs of neurons in parietal cortex, and between neurons in hippocampus and neurons in parietal cortex (Qin et al., 1997).

2.3.3 Neural plasticity

Almost all types of excitatory synapses in the hippocampus have been shown to be capable of undergoing long-lasting activity-dependent modifications, although the relation of these *in vitro* phenomena to *in vivo* plasticity, and especially to behavioral learning and memory, remains somewhat obscure. Since all the modeling work described in the thesis employs quite generic associative learning rules, I will only review some basic findings here; for a

more detailed account, I refer the reader to numerous recent reviews of the topic (e.g., Elgersma and Silva, 1999; Malenka and Nicoll, 1999; Martin et al., 2000).

Long-term potentiation (LTP) was first observed in the perforant path connections from EC to dentate granule cells (Bliss and Lømo, 1973), and since then has been found to be inducible in the perforant path inputs to CA3 and CA1 (Breindl et al., 1994), the Schaffer collaterals from CA3 to CA1, and the recurrent connections within CA3 (Zalutsky and Nicoll, 1990; Debanne et al., 1998). Some of the characteristics of LTP, including the fact that it requires concurrent activation of the pre- and postsynaptic neuron for its induction, have made it a natural candidate for the physiological implementation of Hebbian learning, which in turn is assumed to underlie many instances of behavioral plasticity.

Many of the synapses which can support LTP can also undergo long-term depression (LTD), whereby a decrease in the efficacy of the synapse is induced either by pairing high post-synaptic activity with low pre-synaptic activity (hetero-synaptic LTD; Lynch et al., 1977; Levy and Steward, 1979), or by appropriate stimulation protocols at single synapses (homo-synaptic LTD; Dudek and Bear, 1993; Thiels et al., 1996). Both LTP and LTD induction at these synapses require the activation of N-methyl-D-aspartate (NMDA) receptors, which leads to an influx of calcium ions into the postsynaptic cell, and initiates a cascade of bio-chemical events that results in a synapse-specific change in the efficacy of synaptic trans-mission.

Recent evidence suggests a role for action potentials back-propagating through the den-dritic tree of the postsynaptic neuron in the induction of long-term plasticity (Magee and Johnston, 1997; Markram et al., 1997). In addition, the sign and magnitude of change in synaptic efficacy depends on the relative timing of pre- and post-synaptic action potentials (Markram et al., 1997; Bi and Poo, 1998; Debanne et al., 1998). In particular, if the presy-naptic spike precedes the postsynaptic spike by a short time (on the order of 10 ms), the synapse is potentiated, whereas if the presynaptic spike comes after the postsynaptic spike, the synapse is depressed.

The mossy fiber connections from the dentate gyrus to area CA3 are also modifiable by activity, but this form of plasticity has been found to have characteristics that are signifi-cantly different from those described above (Harris and Cotman, 1986; Jaffe and Johnston, 1990; Nicoll and Malenka, 1995; Manabe, 1997). In particular, mossy fiber LTP does not require the activation of NMDA receptors, it is likely to be expressed presynaptically, and its associativity and specificity are debated.

As already mentioned, the link between LTP/LTD and memory is still controversial (Shors and Matzel, 1997; Martin et al., 2000). However, there are some recent findings which are strongly suggestive of such a link. For example, Moser et al. (1998) saturated LTP in the perforant path in rats. They found that those animals in which it was impossible to induce further LTP failed to learn a spatial navigation task. Pharmacological blockade or genetic knockout of NMDA receptors in the hippocampus also causes impairment in a wide variety of tasks (for a review, see Martin et al., 2000). Finally, Xu et al. (1998) showed that exploration of a novel environment reverses recently induced LTP in the hippocampus of the rat.

2.3.4 Neuromodulation in the hippocampus

The different types of neuromodulatory inputs acting upon the hippocampus were described briefly in the Anatomy section. Data from several different kinds of experiments indicate that the levels of different neuromodulators change selectively as a function of behavioral state. Acetylcholine (Ach), norepinephrine (NE), and serotonin (5HT) are each present at higher levels during waking than during slow wave sleep, whereas REM sleep is characterized by the lowest levels of NE and 5HT and the highest level of ACh in the hippocampus (Hasselmo, 1999).

This co-variation of neuromodulator levels with behavioral state, together with evidence for different modes of information processing during different behavioral states, suggests that these neuromodulators may have a role in switching between different global processing modes in the brain. In particular, Hasselmo has suggested that cholinergic modulation (perhaps along with the septal GABAergic projection) may switch the hippocampus between an encoding mode (during active waking characterized by theta activity in the EEG) and a retrieval mode (during quiet waking and slow-wave sleep, characterized by irregular sharp wave activity) (Hasselmo et al., 1996; Hasselmo, 1999). This idea is in general agreement with Buzsáki's two-stage model described earlier in this section, and is also supported by both experimental data on the physiological and psychological effects of ACh, and computational studies which show the necessity of separating learning and recall in models of memory based on autoassociative networks.

The cholinergic input from the septum can selectively regulate the efficacies of different sets of connections through presynaptic inhibition caused by the activation of muscarinic ACh receptors. Generally speaking, high levels of cholinergic input (such that are present during

active waking) suppress intrahippocampal connections and feedback projections (including the CA3 recurrent connections, the Schaffer collaterals, and the connection from CA1 back to entorhinal cortex) while leaving external inputs to the hippocampus (like the perforant path connection to both CA3 and CA1, the lateral entorhinal input to the dentate gyrus, and, to some extent, the mossy fibers) relatively unaffected (Hasselmo et al., 1995; Hasselmo, 1999). Some similar effects of $GABA_B$ receptor activation have also been observed (Ault and Nadler, 1982). In addition, cholinergic input to the hippocampus has been shown to enhance long-term synaptic plasticity (Burgard and Sarvey, 1990; Huerta and Lisman, 1993), and leads to the suppression of inhibition (Pitler and Alger, 1992) and the direct depolarization of hippocampal pyramidal neurons (Benardo and Prince, 1982). As previously noted (Hasselmo et al., 1995, 1996; Hasselmo, 1999) and confirmed by our simulations of place cells (see Chapter 3), these effects of high levels of cholinergic modulation create exactly the right circumstances for the learning of new information in the hippocampus while minimizing interference from previously stored information. Conversely, low levels of cholinergic modulation lead to the reactivation of previously established memory traces within the hippocampus, and a large influence of hippocampal activity on neocortical representations.

2.4 The medial temporal lobe and behavior

After having reviewed how behavioral variables are reflected in the activities of neuronal populations in the hippocampus and other medial temporal lobe (MTL) structures, we now look at the flip side of the coin to see how these structures are involved in shaping behavior. This question has been addressed mostly by evaluating how different manipulations (either controlled or accidental, as in the case of most human lesions) involving these brain regions affect performance on particular tasks. Thus, I will begin this section by reviewing the manipulations that are most often used in this context. The rest of the section describes what the application of these techniques has revealed about hippocampal function, organized around the themes (memory, space, and conjunctive representations) identified earlier.

2.4.1 Common experimental techniques

Controlled lesions in animals

Controlled lesions, which result in the destruction of specific brain areas or connections, can be invaluable tools in assessing the contribution of these structures to particular behavioral functions. However, extreme care must be taken both in the selection of experimental procedures, in order to minimize damage outside the target region, and in the interpretation of results, taking into account all the effects of the manipulation (including concomitant damage to other structures that inevitably occurs with some techniques). A major problem with interpreting the literature on lesion experiments is that different studies often used different techniques (partly as newer and more selective lesion methods become available), with different kinds of collateral damage (of sometimes unknown magnitude), which makes it extremely difficult to compare or integrate results. Therefore, it becomes important to be aware of the pattern of damage associated with different techniques, and so I will review the most common ways of making lesions to medial temporal lobe structures, with special attention to their selectivity.

Surgical removal. Many early studies used ablation techniques to remove the hippocampus. However, since the hippocampus is hidden inside the cerebral hemispheres, substantial damage to various cortical areas is unavoidable. Indeed, it was later found that much of the behavioral impairment seen in these experiments is actually attributable to damage to the cortical areas of the temporal lobe (such as the entorhinal and perirhinal cortices; Squire, 1992). In addition, surgical techniques are at least as damaging to axons passing through the target area as to the neurons of the area itself, possibly disrupting additional circuits that are functionally unrelated to the target area.

Ischemia. For some reason, the CA1 region of the hippocampus appears to be especially vulnerable to conditions when blood supply to the brain is insufficient. Relatively selective lesions to area CA1 following ishemic events have been reported in several species, including humans (Zola-Morgan et al., 1986), monkeys (Zola-Morgan et al., 1992), and rats (Auer et al., 1989). The main problem with this type of lesion is that it is practically impossible to rule out that damage not detected by conventional histology might have occurred in other areas as well. In fact, there is some evidence that ischemic damage to CA1 can lead

to subsequent pathological processes that damage medial temporal cortices (Mumby et al., 1996).

Electrolytic and radio frequency lesions. These techniques involve inserting the tip of an electrode into the target structure, and damaging the structure by passing current into it. Electrolytic lesions use direct current to burn the tissue, while radio frequency lesions work with high frequency alternating current and cause coagulation. Although these techniques cause less direct collateral damage than ablation procedures, they do damage fibers of passage in and around the target structure, and can also lead to pathological activity and secondary damage in other regions.

Neurotoxic lesions. In recent years, a range of substances have been identified that, when injected into the target structure, kill neurons whose cell bodies are located near the injection site, but leave passing axons unaffected. Frequently used neurotoxins include ibotenic acid (IBO), N-methyl-D-aspartate (NMDA), kainic acid (KA), and quisqualate (QUIS). Not all neurotoxins have identical effects; in particular, Jarrard and Meldrum (1993) found that, when injected into the hippocampus, KA and QUIS did not affect all kinds of cells in the hippocampus equally, and caused substantial damage in extrahippocampal structures. In contrast, especially with a large number of injection sites and small amounts of neurotoxin per site, the damage caused by IBO and NMDA can be restricted to the hippocampus, and affect all cell types within the hippocampus. Another important difference is that while KA and QUIS lesions often lead to seizure activity and resulting additional damage, this has not been observed with IBO or NMDA lesions.

Now let us turn to some commonly used lesion procedures whose interpretation requires special care. First, as already mentioned, attempts to lesion the hippocampus can cause variable amounts of damage to entorhinal and perirhinal cortices when conventional lesion methods are used, and this additional damage has to be taken account in the interpretation of memory deficits. Second, hippocampal lesions with techniques affecting fibers of passage tend to damage the alveus and the fimbria-fornix, thereby disconnecting the subiculum and entorhinal cortex from several subcortical structures (including the septum). Third, entorhinal cortex lesions remove most cortical inputs to and outputs from the hippocampus, although there may be some projections from perirhinal cortex to the hippocampus that bypass entorhinal cortex, and there are some direct projections from area CA1 and the

subiculum to various cortical regions. More importantly, entorhinal lesions leave the extensive network of subcortical connections to and from the hippocampus completely intact. On the other hand, lesions of the fimbria-fornix, which are sometimes interpreted as functionally equivalent to selective hippocampal lesions, leave the cortical inputs and outputs of the hippocampus intact, but cut off the hippocampus (including regions of the subicular complex) and entorhinal cortex from a large number of subcortical areas. In particular, such lesions damage the septal projections to all hippocampal fields as well as to entorhinal cortex, the projections from the hippocampus to the mammillary bodies, anterior thalamic nuclei, and the septum, and various other connections between entorhinal cortex, the subicular complex, and diencephalic structures. Therefore, although fimbria-fornix lesions may render the hippocampus non-functional by removing crucial subcortical inputs, they should also severely affect other circuits involving entorhinal cortex and subicular areas.

Human lesions and neuropsychology

Everything I have said about the difficulties arising from differences in loci and types of lesions and methods of assessment in animal studies of the hippocampus applies to an even larger extent to neuropsychological studies of humans with brain lesions. There are almost as many different kinds of lesions as patients; lesions arise from many different etiologies, most of them affecting several different brain areas without destroying any one of them completely. In addition, there are only few cases where the extent of damage could be confirmed by post mortem histological analysis; in all other cases, the location of the lesion was inferred indirectly from clinical data, or calculated using structural and/or functional neuroimaging. The tasks used to assess memory deficits also come in many different varieties; most of them are fairly complex and high-level, and some of them are scored subjectively.

We will be concerned mainly with lesions that result in amnesia. The major areas in the brain commonly associated with amnesia are the medial temporal lobes, the midline diencephalon, and the basal forebrain. Within the MTL, regions of interest include the CA1 and CA3 fields of the hippocampus, the dentate gyrus, and the subicular complex, as well as adjacent neocortical areas such as the entorhinal, perirhinal, and parahippocampal cortices. In the midline diencephalon, areas whose lesions can cause amnesia are now thought to be the anterior thalamic nuclei and the mammillary bodies, while in the basal forebrain, the critical areas include the medial septum, the diagonal band of Broca, and the

fornix. Whether these structures form a single or multiple functional circuits is still quite controversial.

Common causes of amnesia include Herpes simplex encephalitis (which typically affects perirhinal and parahippocampal cortices), Korsakoff's syndrome (which causes damage in the medial diencephalon as well as other areas; in many cases, it is associated with alcoholism), anoxia (where damage to memory-related areas can be relatively restricted to area CA1), brain trauma (with damage to various sites), and stroke. I should note here that I only consider organic amnesia (i.e., cases associated with physical brain damage), and I ignore cases of so-called psychogenic amnesia, in which there is no evidence for any kind of lesion (and which is often characterized by marked retrograde amnesia in the absence of significant anterograde amnesia).

Reversible functional inactivation

Conventional lesions are not very well suited for investigating the involvement of specific brain areas in particular memory processes (such as encoding, storage/maintenance, recall, and consolidation). For example, if we find that memory under certain conditions is impaired by lesions made before training to a specific structure, we have no way of telling whether that structure is involved in encoding, storage, consolidation, or retrieval. The deficit might even be completely unrelated to memory, affecting instead some other function necessary for successful performance in the task. Multiple lesion groups with lesions made at different times, as well as some other control procedures, can help clarify these issues to some extent. However, the irreversibility of conventional lesions still does not make it possible to distinguish between some of the possibilities. For instance, if we find that lesions at any time before testing impair performance, we may conclude that the lesioned structure is critical for retrieval (or performance), but we have no way of assessing its involvement in encoding, storage, or consolidation.

In recent years, reversible inactivation techniques have been employed to get around some of these limitations. These techniques involve the infusion of a local anesthetic or neurotransmitter antagonist into the target structure (other methods like cooling have also been applied), which renders that part of the brain temporarily nonfunctional, but (ideally) allows complete recovery once the effect of the manipulation wears off. Therefore, these techniques in principle make it possible to determine not only which brain structures are involved in a particular function, but also the time periods during which each of these

structures is important. However, these methods are extremely complex, and it is often very hard to ascertain that the manipulations have exactly the desired effect. In particular, just like with conventional lesions, it is difficult to make sure that all of the target structure is inactivated but no other area is affected. In addition, there is now a temporal aspect to specificity as well which combines with the spatial aspect: since the manipulation has a characteristic time course and spatial spread, and everything is concentration-dependent, it is often impossible to inactivate the whole target area for the same period of time. For a review of functional inactivation studies of memory, see Ambrogi Lorenzini et al. (1999).

Pharmacological manipulations

Pharmacological manipulations other than those intended to shut off the hippocampus or other areas completely have also been used to study the contributions of various factors to the normal functioning of the hippocampal system. These manipulations can in general be either global or local in nature; global manipulations involve systemic administration of the drug, and thus affect a large number of brain regions, while local manipulations involve infusion of the drug directly into a target structure. Accordingly, global methods do not allow the differentiation of the effects of the drug in different parts of the brain. Local methods, on the other hand, have similar advantages and suffer from similar limitations as the reversible inactivation methods discussed above. In the current context, two particular classes of pharmacological manipulations are particularly relevant: those that affect plasticity in the hippocampus and those that affect neuromodulatory influences on the hippocampus and the septal cholinergic input in particular. The basic properties of long-term potentiation/depression (LTP/D) and the physiological effects of acetylcholine have already been described, and here I will concentrate on the behavioral effects.

Blockade of plasticity. As described earlier, NMDA-receptor-dependent long-term plasticity has been proposed to underlie long-term memory storage in the brain. Consistent with this proposal, many of those learning paradigms which, on the basis of lesion studies, are thought to require the hippocampus for their acquisition, have also proved to be sensitive to NMDA receptor inactivation during training. Two points are worth emphasizing here that to some extent limit the usefulness of this approach. First, inactivation of NMDA receptors does not block all forms of learning in the brain. This is particularly relevant in the case of the hippocampus, since some pathways in the hippocampus (e.g., the mossy fibers) are

known to exhibit NMDA-receptor-independent forms of plasticity. This means that, even if we find that performance on a task is unaffected by blockade of NMDA receptors (in the target region or even in the whole brain), we cannot conclude that the task does not require synaptic plasticity at all. Second, NMDA receptors probably have functional roles other than the one they play in the induction of LTP, and a deficit after their inactivation may not have anything to do with learning.

Cholinergic modulation. Transections of the fornix, the main subcortical input and output pathway of the hippocampus, often lead to memory impairments that are equal to or even greater than those following complete lesions of the hippocampus. One of the major projections that make up the fornix is the cholinergic pathway from the medial septum and the diagonal band of Broca to various targets in the hippocampus. Therefore, it is reasonable to ask whether the learning deficits following fornix lesions are attributable to the loss of cholinergic innervation of the hippocampus. Drugs that antagonize or enhance the effects of acetylcholine on its various receptors have been applied (either systemically or locally) to study the role of acetylcholine in the hippocampus (and elsewhere) in learning.

Genetic alterations

In recent years, more and more sophisticated tools have been developed that allow the alteration of the genetic makeup of brain cells in various ways. Some of the manipulations that are most relevant to us target proteins that are known to be involved in synaptic plasticity, more specifically long-term potentiation in the hippocampus. Among the genes involved in different stages (early vs. late) of LTP that have been knocked out experimentally are those encoding the alpha subunit of calcium-calmodulin dependent protein kinase II (CaMKIIα, which is important in the early phase of LTP), protein kinase A (PKA, which is important in the late phase of LTP) and the NR1 subunit of the NMDA receptor.

Many of the early knockout studies suffered from at least one of two potentially serious problems: regional non-specificity, and, sometimes more crucially, temporal non-specificity. The problems arising from regional non-specificity are obvious and are similar to those discussed earlier for other kinds of global manipulations. On the other hand, the problems arising from the fact that a traditional knockout causes the targeted protein to be absent throughout development are mostly particular to genetic techniques. The potential side-effects range from various compensatory effects to a major degeneration of the brain

(also because many genes have multiple roles during development and/or adult life) to the lethality of the knockout.

However, some recent innovations make it possible to get around both of these limitations (at least in some cases). By combining regionally specific promoters with conditional deletion of the target gene, it has become possible to create mice with dysfunctional NMDA receptors only in the CA1 region of the hippocampus (Tsien et al., 1996). In addition, the expression of transgenes can be controlled externally using the tetracycline regulatable tTA system (Mayford et al., 1996). Most recently, these methods have been combined to yield an inducible, reversible, and CA1-specific knockout of NMDA receptor function (Shimizu et al., 2000). However, the applicability of such methods is currently somewhat limited, since ways of targeting manipulations to arbitrary brain regions are lacking. Finally, techniques where the expression of certain proteins is increased (rather than decreased) have also been employed to study learning and memory, although it is much more difficult to justify predictions about what the outcome of such manipulations should be.

Functional imaging

Functional neuroimaging techniques, and functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) in particular, have the potential of providing insights into the functioning of the normal human brain during the execution of various tasks. These techniques measure the three-dimensional spatial distribution of total metabolic activity in the brain, and are normally used to determine which brain areas are differentially activated under certain behavioral conditions (relative to some control condition). In this respect, these methods rightfully belong to the physiology section, since they measure the response of the brain to behavioral variables. However, we will still consider them in this section because they have been used to address the same issues using similar tasks as the other techniques (such as lesions) considered here. In addition, the spatial and temporal resolution of imaging methods, which are on the order of millimeters and seconds, respectively, make the results more directly comparable with results from lesion studies than with those from cellular level physiology.

Notwithstanding their limitations (due to limited spatial and temporal resolution as well as the difficulties involved in the design of the experiments and the interpretation of the results), neuroimaging techniques offer good ways of generating candidate hypotheses about the involvement of specific parts of the brain in particular tasks. In this section, I will just

briefly review some results relating to memory processing, and I refer the reader to several recent reviews (e.g., Desgranges et al., 1998; Lepage et al., 1998; Buckner et al., 1999; Schacter and Wagner, 1999; Tulving et al., 1999; Buckner et al., 2000; Cabeza and Nyberg, 2000a,b) for further details and discussion.

2.4.2 The role of the hippocampus in memory

Common memory tasks

Let me begin this review of the experimental results concerning the role of the hippocampus in memory by describing some of the tasks and procedures that have been often used in studying either humans or various animal species. Since countless variations on the basic paradigms have been used, this review is meant to be representative rather than exhaustive.

Assessment of declarative memory in human subjects. In terms of the material for which memory is tested, two main strategies have been used with human subjects. One approach is to test memory for material that was acquired under controlled conditions in the laboratory; commonly used types of material include lists of words, meaningful text (i.e., sentences or stories), pictures of objects, scenes, or abstract drawings. The obvious advantage of this approach is that the nature and amount of exposure to the study material are known exactly, and the same test can be administered to a large number of subjects, which facilitates the development of standardized measures of performance. The main disadvantage is that, at least in humans, this strategy can only be applied to anterograde memory, and cannot be extended to measure retrograde memory impairment.

Therefore, a different approach has been used to assess retrograde memory in humans. Instead of material previously studied in the laboratory, material that the subject encountered (or may have encountered) in his/her normal life is used to probe memory. This material either consists of items that the general public is normally exposed to, such as public events, names or faces of famous people, and the contents of popular television series, or of facts and events from the subject's personal life (obtained from the subject and/or family members in an interview). Problems with this approach include the fundamentally different nature of the experiences of different people, and the subjective scoring methods that need to be used in some of the tasks. In addition, since (in most cases) completely different methods have been used to assess anterograde and retrograde memory impairment in brain

damaged patients, the results do not provide much evidence about the relationship between anterograde and retrograde amnesia for any particular kind of material.

Declarative memory in non-human species. For several reasons, it has been difficult to find equivalents of these paradigms that can also be used in other animal species. First, it is not clear whether other species even have the same kind of fact and event memory as we do, although, as we will shortly see, there are reasons to believe that they might. Second, the methods that are routinely used for testing memory performance in humans cannot be easily adapted for use in animal experiments. In particular, recall-based paradigms typically require verbal responses, and often rely on verbal instructions to set up the task. Of course, some tasks could be modified to require a non-verbal behavioral response, but then it is hard to ascertain that the animal cannot rely on some learning mechanism other than explicit recall to perform the task correctly. As an example, some tasks that were originally designed to test the recall of specific past experiences in animals turned out to be solvable based on simple judgements of familiarity (i.e., recognition). Similarly, the recall context can in principle be set up non-verbally, but it is difficult to make sure that the context is effective in encouraging the animal to perform memory recall.

The existence in animals of memories like human episodic memory seems particularly difficult to prove, especially since it can only be tested in recall-based paradigms (not, for instance, recognition). Episodic memory may be defined as the ability to recall in detail specific past experiences, including information about what happened, where, and when. This definition differs in one crucial aspect from that of Tulving and Markowitsch (1998); in particular, it does not require the presence of a special kind of awareness that apparently accompanies episodic remembering in humans (consequently, this kind of memory has been termed "episodic-like" to distinguish it from the full human experience of episodic memory). The definition used here appears to be a much more pragmatic one as it does not preclude the consideration of kinds of memory in animals which might be behaviorally (and perhaps also neurally) indistinguishable from human episodic memory.

Recently, a series of experiments has been conducted with scrub jays, a food-storing bird species, which suggests that non-human animals may indeed be capable of episodic remembering (Clayton and Dickinson, 1998, 1999; Clayton et al., 2001). In these experiments, animals were allowed to cache different kinds of food (perishable vs. non-perishable) at different locations at different times, and their search preferences at the time of recovery of the food cache were evaluated for evidence of their memory of the details of the caching event.

The authors found evidence that the jays could recall many aspects of a unique caching event, including its time, location, and what was stored there (and perhaps even whether they were observed by another jay when they performed the caching). If similar evidence could be obtained from other species (such as rats) for which neurobiological techniques have been developed, one could hope to study the neural mechanisms of episodic(-like) memory in animals.

However, most of the tasks to date which have been used to study the involvement of the hippocampus and other related structures in memory in non-human animals are markedly different from those in which impairment is best documented in amnesic patients. A typical example, and probably the most extensively used memory task in this context is the delayed non-match to sample (DNMS) task. In the simplest version of the task, a stimulus is presented briefly, followed by a delay period. At the end of the delay, the animal is presented with a choice between two stimuli, one of which is the stimulus presented previously and the other one is novel, and the animal is rewarded for choosing the novel stimulus. DNMS is therefore essentially a test of recognition memory. Another type of task that has often been used to assess memory in animals is concurrent discrimination, in which the animal is presented several different pairs of stimuli (one pair at a time), and learns to associate one stimulus in each pair with reward in the course of multiple presentations of each stimulus pair. The stimuli can either be very simple or quite complex (such as whole scenes). The hippocampal dependence of many other memory tasks has been studied, including the retention of information in various conditioning paradigms.

Kind of information affected

As already mentioned, only certain kinds of memories, termed declarative or explicit memories, are affected in amnesia. Within the domain of human declarative memory, some researchers have for a long time distinguished episodic and semantic memories (Tulving, 1972). More recently, the definition of episodic memory has been refined; as discussed above, it may be defined as explicit remembrance of specific autobiographical events (or episodes), with associated detailed contextual information including time, location, and perceptual experience. Within semantic memory, distinctions can be made between personal semantics (factual knowledge about one's life, which may or may not be associated with specific episodes), knowledge of facts about the world, as well as of public events and personalities, and general semantics (including the use of objects, the meaning of words,

knowledge of concepts and categories, etc.). General semantics is in many ways dissociable from other kinds of semantic knowledge, and is not always considered to be part of declarative memory.

Episodic and semantic memory in amnesic patients. Some researchers have proposed that the medial temporal lobe memory system, which is the system damaged in amnesia, contributes in the same way to the processing of all kinds of declarative memory (although other parts of the brain such as the frontal lobes may also be involved in episodic memory) (Squire, 1992; Squire and Zola, 1998). Consequently, these theories predict that all types of declarative memories, be it personal episodes or general facts, will always be proportionally impaired in amnesia due to temporal lobe damage. Other theories suggest that episodic and semantic memory rely on at least partially separable processes, and, consequently, that it may be possible to find dissociations between these different kinds of memories in brain damaged patients (Tulving, 1995; Tulving and Markowitsch, 1998).

In a study that examined new fact learning in amnesics, subjects had to memorize new fact-like information over multiple trials in several weeks (Hamann and Squire, 1995). The main finding was that amnesics learned very slowly and achieved much lower asymptotic performance than controls. In addition, their fact learning ability appeared to be impaired to about the same extent as their episodic memory, as evidenced by the finding that both fact and episodic memory in amnesics after a one week delay were about as good as in controls after a four week delay. On the other hand, Tulving et al. (1991) reported that their severely amnesic patient K.C., who was completely unable to recall specific past experiences, could nevertheless acquire (even though at an abnormally slow rate) considerable amounts of new factual knowledge, which he then retained for at least 12 months. It is interesting to note that they used a special study procedure designed to reduce interference due to incorrect answers at early stages of learning, which was shown to significantly improve performance in amnesics (but not in controls). However, the severely amnesic patient E.P. of Hamann and Squire (1995) failed to show any new learning even using this study-only procedure (although this patient and the patient of Tulving et al. (1991) had quite different patterns of brain damage, with K.C. having extensive damage to frontal, parietal, occipital, and retrosplenial cortices in addition to the medial temporal lobe on the left side, with only minor damage on the right, while E.P. has more symmetric lesions which affect most of his medial temporal lobes and some areas of lateral temporal cortex, but only minor damage elsewhere).

The complex relationship between episodic and semantic memory is further highlighted by the study of Verfaellie and Cermak (1994). They used an experimental design in which words were presented either once or twice in a list, in a color unique to the presentation. Recall of the color in which the item had been presented was taken as evidence for episodic memory for the presentation. They demonstrated that amnesics can recall generic information without necessarily recalling any of the episodes when they acquired that information. On the other hand, they also found that amnesics do not benefit as much from multiple presentations of the same information as controls do.

Recent studies of people with early onset amnesia provided further evidence that large amounts of semantic knowledge can be acquired despite severe impairment of episodic memory (Vargha-Khadem et al., 1997; Gadian et al., 2000). The patients suffered bilateral damage to the hippocampus at birth or in early childhood, and, as a consequence, they showed severe amnesia for episodic information. On the other hand, they all succeeded in normal education, and acquired close to normal amounts of factual knowledge, as well as general semantic information including language skills. Based on these findings, Vargha-Khadem et al. (1997) suggest that the hippocampus itself is only critical for episodic memory, and semantic memory depends mostly on other temporal lobe areas.

More recently, Verfaellie et al. (2000) found some support for this theory in adult patients; they found that a patient whose lesion included the hippocampus as well as surrounding cortical areas failed to learn any new information in everyday life, while another patient with a lesion limited to the hippocampus managed to acquire familiarity with new vocabulary and famous faces.

Squire and Zola (1998) offer a different explanation of the findings of Vargha-Khadem et al. (1997) and Tulving et al. (1991) (which may also apply to the data of Verfaellie et al. (2000)). They suggest that the ability to acquire semantic knowledge in these cases is the consequence of some residual episodic learning capability, and episodic and semantic memory are in fact proportionately impaired. They also quote the results of Reed and Squire (1998), who found that even patients with lesions limited to the hippocampus acquired abnormally little semantic information (about new words, famous people, and public events) since the onset of their amnesia.

There are also reports in the literature of retrograde amnesia affecting episodic and semantic memory disproportionately. In many cases, episodic memory is more severely impaired than semantic memory (Warrington and McCarthy, 1988; Markowitsch et al., 1993; Verfaellie

et al., 1995), but the opposite pattern of impairment has also been observed (De Renzi et al., 1987; Grossi et al., 1988; Rusconi et al., 1997). However, it has been suggested that these relatively unusual dissociations result from damage outside the classical medial temporal lobe - medial thalamic memory circuits.

Evidence from functional imaging. Functional imaging studies of normal people also suggest that there are dissociations among the areas involved in processing different types of declarative memory (and at different stages of processing) (Buckner et al., 1999; Schacter and Wagner, 1999; Tulving et al., 1999; Cabeza and Nyberg, 2000a; Maguire et al., 2000). Probably the most ubiquitous finding in these studies is that performance of even the seemingly simplest memory task involves activity changes in a large number of areas scattered across the brain. The most robust activations were actually found not in the medial temporal lobe (as one might have expected from the lesion studies), but in prefrontal cortex, although more recent studies consistently showed that MTL areas (and the hippocampus and parahippocampal gyrus in particular) are also involved. The actual patterns of activation vary to a large extent between studies, but some findings have been relatively consistent under similar conditions, revealing differential involvement of brain areas depending on the kind of memory tested (e.g., semantic vs. episodic) and the type of material (e.g., verbal vs. pictorial), but also on dominant memory process (e.g., encoding vs. retrieval), task difficulty, and many other variables.

Many of the observed activations are strongly lateralized. A striking example is that prefrontal activations in semantic retrieval tasks occur almost exclusively on the left. Within left prefrontal cortex, activations have been seen in almost all areas, depending on task details (including domain of knowledge tested). Semantic tasks also often activate the left middle temporal gyrus, occipito-temporal regions, anterior cingulate cortex, and parts of the cerebellum.

The prefrontal contribution to episodic memory encoding was initially also proposed to be lateralized. Studies using verbal materials generally support this idea, showing preferential activation in left inferior prefrontal cortex. On the other hand, encoding nonverbal materials often resulted in bilateral or right-lateralized prefrontal activations. Early studies often failed to find activation during episodic encoding in medial temporal areas; however, more recently a large number of studies showed both anterior and posterior MTL activations (Schacter and Wagner, 1999), which were often lateralized. Word encoding activated the MTL mostly on the left, novel face encoding on the right, while scene and object encoding

led to bilateral MTL activation. Activations were also often found in the right cerebellum for verbal materials.

Studies of episodic retrieval have typically used free recall, cued recall, or recognition of events previously presented in the laboratory. In most cases, prefrontal activations are right-lateralized; a fairly consistently activated area is right anterior prefrontal cortex. Other prefrontal areas associated with episodic retrieval include Brodmann's areas 47 and 10 on the left (sometimes associated with retrieval effort), and bilateral areas 10, 9, and 46 (sometimes associated with retrieval success). Once more, in contrast to earlier efforts, recent studies consistently found medial temporal activations in episodic retrieval. These tended to be bilateral, irrespective of type of material, and to accompany successful retrieval. Other areas that were often activated include retrosplenial cortex, medial areas of the cuneus and precuneus, as well as lateral parietal, anterior cingulate, occipital, and cerebellar regions.

One recent study examined the brain areas (and their functional connectivity) involved in the retrieval of real-world memories acquired up to twenty years before the experiment (Maguire et al., 2000). They found activations in some of the same areas that were also implicated in the retrieval of memories formed relatively recently in the laboratory, including the left hippocampus, left parahippocampal gyrus, retrosplenial cortex, left middle temporal gyrus, and cerebellum. However, activations were also present in other areas, notably the temporal pole (bilaterally). This study also examined how the functional connectivity (as measured by a structural equation modeling approach) depended on the kind of material retrieved. The four categories they used were autobiographical events (or episodes), autobiographical facts (previously called personal semantics), public events, and general knowledge (facts). They found that the functional connectivity between the hippocampus and the parahippocampal gyrus increased significantly during the retrieval of autobiographical events both compared to the retrieval of autobiographical facts, and compared to all other conditions. Connectivity between parahippocampal cortex and the temporal pole also increased for the retrieval of autobiographical events. In contrast, retrieval of public events and general facts increased the functional connectivity between the temporal pole and lateral temporal cortex.

Memory impairments in animals. In the animal literature, the most clear-cut examples of tasks affected by hippocampal lesions lie in the domains of spatial and relational processing (although they almost invariably have memory components as well), and therefore will be considered in later sections. However, some simple memory tasks have also

been found to be affected by hippocampal and other medial temporal lobe lesions. A much-studied example is the delayed non-match to sample (DNMS) task, effectively a test of recognition memory, which for a long time has been thought to be sensitive to hippocampal lesions, although some evidence to the contrary has also been presented. A meta-analysis of studies looking at the effects of lesions to different medial temporal lobe structures in the monkey on performance in the visual DNMS task with trial-unique stimuli was recently conducted by Baxter and Murray (2001). They found that performance was at most mildly affected by lesions limited to the hippocampus. In contrast, damage to the perirhinal cortex caused severe impairment, while the greatest impairment was found with larger lesions including the entorhinal and/or parahippocampal cortices in addition to perirhinal cortex (but sparing the hippocampus itself).

Baxter and Murray (2001) did not find any effect of the length of the delay in the range of 15 seconds to 40 minutes. However, some earlier studies indicated that DNMS performance can be normal in animals with hippocampal or perirhinal lesions at very short delays (less than 10 seconds; Zola-Morgan et al., 1989b,a; Zola et al., 2000). The general pattern of findings in the rat is quite similar. Damage to rhinal cortex leads to impairment in the olfactory continuous version of the DNMS task in rats (Otto and Eichenbaum, 1992a); whether the hippocampus itself is necessary for normal performance is more controversial (Duva et al., 1997; Clark et al., 2001).

Studies which looked at sensory discrimination tasks found that animals with damage to the hippocampus and/or adjacent cortex are impaired in learning new discriminations only when either a long delay or a large number of presentations of other stimuli intervene between two presentations of the particular object pair. Animals with hippocampal lesions can also be impaired in the long-term retention of acquired discriminations (Zola-Morgan and Squire, 1990).

Dependence on type of access

Recall vs. recognition. There has been some debate over whether the impairment in explicit memory in amnesics depends on the way the task requires them to access information. One important distinction is between (free) recall and recognition of previously studied items. While recognition can be thought of as a relatively straightforward matching process between the current input and stored representations of studied items, recall involves more complex processing based on the available cues (which, in the case of free

recall, is the context), and is thought to require auto- and/or heteroassociative processing. It is important to note that, at least in a simple psychological framework, items that can be recalled should also be recognized successfully. Indeed, some models of recognition suggest that there are essentially two routes to recognition: one is based on recollection, and the other on familiarity, perhaps including recency effects (Yonelinas, 1994).

Based on these differences, we might expect that lesions could affect recall and recognition differently. Indeed, frontal lesions appear to have a larger effect on free recall than on recognition (Jetter et al., 1986). Whether there is a difference between the impairment of free recall vs. recognition after damage to the system underlying amnesia is much more controversial. Some studies found evidence that recall was disproportionately more impaired than recognition in amnesics (Hirst et al., 1986, 1988). However, another study by Haist et al. (1992) found no difference in the level of impairment in recall vs. recognition. It has been suggested that additional frontal damage in the patients in the earlier studies might explain the discrepancy; alternatively, whether materials can be semantically organized during encoding might affect the results (Mayes and Downes, 1997).

Lesion studies in animals suggest that the location and extent of the lesion within the medial temporal lobe could also be important factors. In particular, it has been suggested that while recall, and therefore (according to the dual-process model) also the recall-based component of recognition, depends on the hippocampus, familiarity-based recognition is supported by perirhinal cortex. The evidence for this view was recently reviewed by Brown and Aggleton (2001), and can be briefly summarized as follows. As mentioned in the physiology section earlier, a substantial sub-population of neurons in perirhinal cortex respond differentially to the first and subsequent presentations of stimuli, which could be used as the basis for making familiarity judgements. Such responses rarely occur in the hippocampus. This dissociation between neural activities is confirmed by larger-scale imaging studies in rats using the so-called immediate early gene c-fos; neural activity was found to be higher after processing novel stimuli than after processing familiar stimuli in perirhinal cortex, but not in the hippocampus (although the opposite pattern of differential activation was observed when the task involved spatial arrangements). As we have already seen, controlled lesions of perirhinal cortex cause a much more severe impairment in the delayed non-match to sample task than hippocampal lesions, which is consistent with the observation that the DNMS task can be solved by discriminating between novel and familiar (or recently perceived) stimuli (although in principle can also be solved through explicit recall of earlier presentations). The above-mentioned single-case studies of human patients with impaired recall and spared

recognition abilities support a possible anatomical dissociation. However, even patients with selective damage to the hippocampus were found to be impaired in not only recalled-based, but also familiarity-based recognition. On the other hand, both event-related potential and fMRI studies of normal subjects found evidence for dissociable components of recognition memory. In particular, a recent fMRI experiment (Eldridge et al., 2000) showed that, even when previously studied stimuli were successfully recognized, the hippocampus was activated only when the subject claimed to have explicitly recalled the encoding episode, but not when the response was claimed to be based on familiarity.

Priming. An additional way in which the consequences of experience can be measured in subsequent testing is by looking at perceptual fluency effects such as repetition priming. Priming involves essentially the same kinds of material that can also be processed in declarative memory. It may be defined as an increased facility for detecting or identifiying stimuli as a result of prior exposure to these (or similar) stimuli, independent of the ability to recall (or recognize) earlier experience with them. Priming is one of the learning abilities (along with others like motor skills, habits, and certain types of simple conditioning) that have traditionally been thought of as independent of the neural structures underlying amnesia (Squire, 1992). One specific suggestion has been that priming may rely on relatively high-level, but modality-specific association cortices (Ochsner et al., 1994), but at least some forms of priming may also be supported by lower-level sensory areas. A different, although not necessarily contradictory, suggestion has been that the neural changes responsible for the priming effect happen in the same connections which support the cortical storage of declarative memories (McClelland and Rumelhart, 1985; Becker et al., 1997; Stark and McClelland, 2000).

The evidence is quite strong that priming for material which was already familiar before training is completely normal in amnesics (see e.g. Squire et al., 1993). Contrary to early evidence, it now seems that amnesics can show completely normal priming for at least some kinds of novel material as well. However, amnesics do seem to be impaired in priming involving novel associations, especially when it involves binding together perceptual and conceptual information (Ochsner et al., 1994). Finally, it is worth noting that priming also doubly dissociates from recognition memory, especially since perceptual fluency effects could in principle serve as a familiarity signal that could be used for recognition. Brain-damaged patients have been described who either had no recognition memory and preserved priming (e.g., Stark and Squire, 2000), or impaired priming and preserved recognition memory (e.g.,

Wagner et al., 1998).

Temporal aspects of processing

Anterograde and retrograde amnesia. We suggested earlier that the timing of experimental manipulations can be a crucial factor in determining their amnesic effects, since long-term memory comprises several temporally dissociable basic processes (such as encoding, storage, retrieval, and consolidation) which might require different anatomical structures or neural mechanisms. The most basic distinction is between manipulations preceding learning (anterograde effects) and those following learning (retrograde effects). Most amnesic patients show both retrograde and anterograde memory impairment. However, there is very little evidence indicating that either the temporal extent or the severity of retrograde amnesia may be correlated with the severity of anterograde amnesia. In fact, Shimamura and Squire (1986) found no correlation between the severity of retrograde amnesia in remote time periods and the extent of anterograde impairment.

In addition, there are reports in the human neuropsychology literature of dissociations between retrograde and anterograde amnesia. Extensive retrograde amnesia, with at most moderate anterograde amnesia, has been described in a number of cases (for a review, see Kapur, 1993). It has been suggested that this pattern of impairment may result from damage outside the hippocampal system, perhaps involving the temporal pole or the frontal lobes. In this context, it is also worth noting that people with this kind of selective retrograde amnesia (also called focal RA) show substantial variations in their impairment (e.g., whether their RA is graded, and whether it mostly affects episodic memory or fact memory). There are also a few reported cases of marked AA in the absence of significant RA, including a patient with histologically confirmed, selective bilateral lesion to the CA1 field of the hippocampus (Zola-Morgan et al., 1986).

In experimental animals, anterograde and retrograde memory deficits resulting from lesions to the hippocampus typically parallel each other. However, there also appear to be several tasks which can be acquired without the hippocampus, but for which hippocampal lesions following training cause a severe impairment in subsequent testing (Sara, 1981; Ross et al., 1984; Jarrard and Davidson, 1991; Land et al., 2000; Sutherland et al., 2001). This suggests that whether a dissociation between anterograde and retrograde amnesia is seen in humans may also depend on the type of memory tested. On the other hand, the amnesic effects of controlled lesions of neocortical areas outside the MTL in animals have not been tested

systematically, although such experiments might provide an explanation for the diversity of amnesic impairments observed in humans.

All in all, the existence of dissociations between anterograde and retrograde impairments after lesions suggests that some brain regions are critical only for a subset of the temporally dissociable processes that are necessary for intact long-term memory performance. The results of manipulations other than lesions also confirm the conclusion that different neural mechanisms are necessary at different stages of processing, particularly encoding and retrieval. For example, blockade of NMDA-dependent synaptic plasticity during training causes great impairment in several hippocampally-dependent tasks, but the same manipulation has no effect during recall (e.g., Steele and Morris, 1999). Since many of these tasks are spatial in nature, they will be discussed in more detail in the next section. Similarly, the muscarinic cholinergic receptor antagonist scopolamine interferes with the acquisition, but not the recall phase of the trial-unique delayed non-matching to sample task in monkeys (Aigner et al., 1991). Cholinergic receptor antagonists also interfere with operant delayed matching (Dunnett, 1985) and the learning phase of several spatial tasks in rats, and similar results have been obtained using list learning paradigms (Ghoneim and Mewaldt, 1975; Frith et al., 1984) and paired associate learning (Crow and Grove-White, 1973; Caine et al., 1981) in human subjects.

Gradedness of retrograde amnesia. One of the most heavily debated issues in hippocampal research is the gradedness of retrograde amnesia, i.e.: for what kinds of information and under what circumstances are recent memories affected more than old memories in retrograde amnesia? This question is of great theoretical significance, because the gradedness of RA is widely regarded as the main signature of the consolidation of memory traces from the hippocampus and perhaps related cortical areas to other parts of neocortex, which idea in turn is the cornerstone of several influential theories of hippocampal function. However, it should be pointed out that the clearest form of graded retrograde amnesia, where performance is actually better for older memories than for more recent ones in a single patient, is *not* a necessary consequence of memory consolidation.

In humans with relatively selective lesions, retrograde amnesia is typically graded (Squire and Alvarez, 1995; Mayes and Downes, 1997). However, numerous cases with apparently ungraded memory loss for at least some types of information have also been reported (Nadel and Moscovitch, 1997). It has been suggested that the extent, severity, and degree of gradedness of RA may depend on the locus and extent of the underlying lesion (Rempel-

Clower et al., 1996; Reed and Squire, 1998; Squire et al., 2001). In particular, lesions that are mostly limited to the hippocampus (and perhaps entorhinal cortex) often lead to temporally graded RA, with larger lesions causing more extensive amnesia, such that damage limited to CA1 causes amnesia for at most a few years before the lesion, while lesions affecting more of the hippocampal fields causes RA that extends to 15-25 years (Rempel-Clower et al., 1996). Even more extensive medial temporal lobe lesions may produce retrograde amnesia covering up to 50 years, while completely ungraded RA may be caused by damage to lateral and/or anterior parts of the temporal lobe (instead of, or in addition to, areas in the medial temporal lobe). This last pattern of findings has often been interpreted as damage to permanent storage sites in neocortex.

Nadel and Moscovitch (1997); Nadel et al. (2000) proposed that whether the impairment appears to be temporally graded or ungraded may also depend on the type of memory considered. In particular, they suggested that episodic memory tends to be more impaired than semantic memory following lesions that go beyond CA3, CA1, and the dentate gyrus. They further claim that existing episodic memories (and also spatial memories) are qualitatively different from normal in people with damage to "the hippocampal system" (although they do not specify which medial temporal lobe structures in particular would be important in this distinction). According to them, all personal memories in amnesics (even very remote ones, which were previously considered to be intact) lack the rich detail that characterizes true episodic memory in normals (the memories of controls, on the other hand, became less detailed, or more semanticized, for more remote time intervals) (Nadel et al., 2000). The results for spatial memories in a patient with bilateral lesions to the hippocampus and the parahippocampal gyrus were in some ways similar (Rosenbaum et al., 2000). Although gross aspects of the patient's remote spatial memory were normal – i.e., the "cognitive map" seemed to be intact, in apparent contradiction with the general spatial theory of hippocampal function –, the amount of detail (in this case, the number of landmarks) in the representation was much lower than in normal subjects. The conclusion from these studies was that the "hippocampal system" is necessary for normal episodic and spatial memory irrespective of the age of the memories.

Due to its great theoretical significance, a large number of studies have examined the gradedness of retrograde amnesia in various animal species, including monkeys (Salmon et al., 1987; Zola-Morgan and Squire, 1990; Gaffan, 1993; Thornton et al., 1997), rats (Winocur, 1990; Kim and Fanselow, 1992; Bolhuis et al., 1994; Cho et al., 1995; Cho and Kesner, 1996; Wiig et al., 1996; Maren and Fanselow, 1997; Anagnostaras et al., 1999; Kubie et al., 1999;

Mumby et al., 1999; Sutherland et al., 2001; Winocur et al., 2001), mice (Cho et al., 1993), cats (Uretsky and McCleary, 1969), and rabbits (Kim et al., 1995). These studies used a diverse array of tasks, ranging from object discrimination to socially transmitted food preference (and also including spatial and contextual tasks that will be discussed later), and applied different lesion techniques, making it difficult to draw general conclusions. This is especially true since the studies do not present a uniform picture, some demonstrating temporally graded RA, others more compatible with a temporally uniform impairment, or even showing better performance in amnesics at more recent time points.

As an example, consider the long-term retention of concurrently acquired object discriminations, which is impaired by hippocampal lesions (Zola-Morgan and Squire, 1990). Discriminations learned at different times before the lesion are affected to a different extent. Unlike normal monkeys, which show better retention of recently learned material (normal forgetting), hippocampal animals are most severely impaired at remembering recently learned discriminations, and their performance is better (as good as that of normals) for old material. Similar results have been obtained using fornix lesions in rats (Wiig et al., 1996), while the results using lesions to perirhinal cortex in rats and monkeys may be more consistent with a long-term (ungraded) impairment of memories acquired before the lesion (Wiig et al., 1996; Thornton et al., 1997).

As discussed earlier, reversible functional inactivation techniques in principle allow a more precise determination of which sub-processes of memory really depend on a particular brain area. Thus far, only a limited number of such studies has been carried out. In a series of experiments, Ambrogi Lorenzini et al. (1999) have studied the consolidation of the rat's passive avoidance response through reversible functional inactivation of a large number of brain areas. They found that the hippocampus was necessary for encoding, retrieval, and early consolidation (about the first hour after acquisition); both entorhinal cortex and the fimbria-fornix were found to be necessary during encoding, but not retrieval or consolidation; and perirhinal cortex appeared to be important during encoding, retrieval, and the late but not the early stages of consolidation of this task. The basolateral nuclei of the amygdala were also found to be critical for early consolidation.

2.4.3 The role of the hippocampus in spatial behavior

The presence of spatially selective cells throughout the hippocampus (as described earlier) already strongly suggests that the hippocampus plays a major role in spatial behavior

in rodents. Indeed, it was mainly these physiological data that led O'Keefe and Nadel (1978) to propose their 'cognitive mapping' theory of hippocampal function. Many of the subsequent lesion studies, some of which are described in this section, generally confirmed this spatial role of the hippocampus, particularly in tasks which require the animal to identify or remember specific locations based on the spatial configuration of multiple cues.

The Morris water maze

Let us first consider the water maze task introduced by Morris (1981). In this task, the rat has to swim in a large pool of opaque water to a platform in order to escape. The platform may be visible above the surface or hidden below. Normal rats can quickly learn to swim straight to the platform from any location in the pool, even if the platform itself is hidden but its location can be inferred from the position of visual cues outside the apparatus. If the platform is then removed, the animals spend most of the time around the location where the platform used to be.

Rats with hippocampal lesions (Morris et al., 1982) or fornix transection (Whishaw et al., 1995) are severely impaired in learning the water maze task if the platform is hidden but are unimpaired if the platform is visible. They are eventually able to learn to swim to the hidden platform from a fixed starting position, but, unlike normal rats, are subsequently unable to navigate to the platform from a new starting position (Eichenbaum et al., 1990). Once trained with the platform visible, lesioned animals can still navigate to the location of the platform if it is removed or becomes hidden (Whishaw et al., 1995; Whishaw and Jarrard, 1996).

Rats with subiculum lesions appear to be more severely impaired than rats with hippocampal lesions in the hidden platform version of the task (Morris et al., 1990). While hippocampal rats search for the platform by swimming around the pool at the correct distance from the wall, indicating some knowledge of the location of the platform, subicular rats search at random as if they had never been in the environment before. Finally, combined lesions to the hippocampus and the subiculum lead to a more severe impairment than lesions to either structure alone.

The effect of hippocampal and subicular lesions on the retention of the hidden platform water maze task has also been examined (Bolhuis et al., 1994). Normal rats showed substantial forgetting after 14 weeks, but they relearned the task very rapidly. In contrast,

rats that received hippocampal or subicular lesions either 3 days or 14 weeks after training, performed at chance during subsequent testing, and were severely impaired at relearning the task. Sutherland et al. (2001) found that selective hippocampal lesions caused retrograde amnesia for both the hidden platform version of the water maze and a simple object discrimination task learned in the same environment. However, the acquisition of only the spatial task and not the object discrimination task was affected by the lesion, showing a clear dissociation between anterograde and retrograde deficits. They also report a significant trend towards worse performance with longer training-lesion intervals for both tasks, which argues against consolidation in this case.

Riedel et al. (1999) used both chronic and acute infusions of a AMPA/kainate glutamate receptor antagonist to reversibly inactivate the hippocampus at different times during and after training in the water maze task. They found that the functionality of the hippocampus was necessary during both encoding and retrieval. They also tested the effect of inactivating the hippocampus for seven days during the retention interval, beginning either one or five days after the end of training. Both of these manipulations caused severe impairment, indicating that they interfered with either the storage/maintenance or the consolidation of memory traces.

Various pharmacological manipulations have been found to interfere with acquisition of spatial tasks in the Morris water maze. For example, rats treated with cholinergic receptor antagonists are severely impaired at learning the standard hidden platform water maze (Sutherland et al., 1982; Whishaw, 1989), and similar results have been obtained using other spatial tasks such as the radial maze (Eckerman et al., 1980; Wirsching et al., 1984). Several different versions of the water maze task have also proved to be sensitive to NMDA receptor inactivation during training (Morris et al., 1986; Bannerman et al., 1995; Steele and Morris, 1999), again paralleling many other spatial tasks such as a delayed version of the radial arm maze, and delayed alternation in the T-maze (Tonkiss and Rawlins, 1991), a task which will be described shortly.

In many of these cases, however, whether an impairment can be observed may depend on various details of the task and the training procedure. For example, Bannerman et al. (1995) showed that pretraining in one water maze can eliminate the effect of NMDA receptor blockade on learning to navigate to a hidden platform in a second water maze, but only if the task used in pretraining is also spatial in nature. On the other hand, Saucier and Cain (1995) found that even non-spatial pre-training can eliminate the effect of NMDA receptor blockade in this task, and suggested that sensorimotor disturbances caused by

NMDA receptor antagonists might be responsible for the deficit observed under standard conditions. The resulting controversy still has not been resolved (see Hoh et al., 1999; Martin et al., 2000).

Steele and Morris (1999) studied a hippocampally dependent delayed matching to place task in the water maze, in which normal rats are capable of one-trial learning. They found that either chronic or acute infusions of an NMDA receptor antagonist disrupt the retention of information acquired in one trial across delays of 20 minutes or longer (but not 15 seconds). At the same time, NMDA receptor blockade had no effect on the recall of a previously learned water maze, in agreement with earlier studies implicating NMDA receptors in encoding but not in retrieval.

The radial arm maze

The radial arm maze consists of a central platform and several (typically eight) narrow corridors arranged symmetrically around it. The far end of each arm contains food at the beginning of each trial, and performance is measured by the number of times the animal enters a previously visited arm (which therefore contains no food). In the delayed version of the task, the rat is only allowed to visit some of the arms at the beginning of the trial, and the rest of the arms only become available after some delay. Normal rats make very few errors in either version of the task, even though they do not appear to use any simple search strategy (Olton and Samuelson, 1976). Rats normally identify the arms based on their spatial location relative to external cues, although they can also learn a cued version of the task, in which the arms can be distinguished based on their intrinsic features but the arms are rotated after each choice (Olton et al., 1979).

Rats with hippocampal lesions are impaired in the acquisition of the spatial, but not the cued version of the task (Jarrard, 1993). Lesions to entorhinal cortex (Rasmussen et al., 1989), as well as selective lesions of the granule cells of the dentate gyrus (McNaughton et al., 1989) have a similar effect. However, once learned, the retention of neither task was affected by hippocampal lesions. Remarkably, rats with combined lesions to perirhinal and postrhinal cortices were recently found to be unimpaired in both the radial arm maze and the Morris water maze (Bussey et al., 2000).

Other spatial tasks

Various tasks in T-, Y-, and plus-shaped mazes have also been used to assess the role of the hippocampus and other structures in spatial processing. In particular, several studies have looked at different versions of the delayed forced-alternation task in these environments. In this task, the rat starts from one arm of the maze, and goes into the only other open arm for reward. After a delay, it is then returned to the same or a different start arm, is given a choice between the arm visited before and a previously unvisited arm, and is rewarded for choosing the new arm. In different versions of the task, the "new" arm can be identified by its location in space with respect to external cues (a real spatial task), features of the arms themselves, or based on whether it can be reached with a left turn or a right turn from the start arm. In a study designed to separate out the contributions of these different strategies, Rasmussen et al. (1989) found that normal rats could learn both the spatial (allocentric) and the left-right alternation (egocentric) version of the task (although they found the spatial version much easier). Rats with lesions to entorhinal cortex could learn the egocentric version as well as or better than normals, while they were mildly impaired in the spatial task. In this context, it is quite surprising that fornix lesions appear to cause a severe impairment in the simplest version of the task in the T-maze (Shaw and Aggleton, 1993), where the orientation of the maze is fixed and alternation is between the two branches of the T, since the animal could in principle use either spatial or nonspatial strategies to solve this task.

The role of the hippocampus in learning associations between locations and different kinds of reward was tested in a Y-maze which had food at the end of one arm and water at the end of another (Hirsh et al., 1978). Normal animals could easily learn to go to the arm containing food when they were hungry, and to the arm containing water when thirsty (in fact, they learn this while exploring the maze even if they are not motivated at the time). In contrast, animals with fornix lesions could not learn both associations simultaneously; they could learn one task, but then they would unlearn it while learning the other.

Spatial memory in humans

Several studies that describe amnesic patients mention that many of these people seem to have difficulty finding their way around, especially in unfamiliar environments. However, until recently, no systematic evaluation of the navigational abilities of amnesic patients was undertaken. This is somewhat surprising given the profound spatial deficits resulting from

hippocampal lesions in animals (particularly rats), and the special role attributed to spatial processing in some theories of hippocampal function based on these animal experiments. Finally, in the last few years, several studies examined spatial memory in amnesic patients. Interestingly, patients with either left or right hippocampal lesion were found to be unimpaired in a spatial task designed to be analogous to the Morris water maze (Bohbot et al., 1998). However, patients with lesions to the right parahippocampal cortex were impaired in this task. Teng and Squire (1999) found that their patient with extensive bilateral medial temporal lobe lesions had no knowledge of the spatial layout of the environment he moved into after he became amnesic. In contrast, he had normal spatial memory for the region where he grew up, and from where he moved away a long time ago. The same preservation of remote spatial memory accompanied by severe anterograde amnesia for spatial material was found in the patient of Rosenbaum et al. (2000), who also had relatively widespread brain damage which involved the hippocampi and the parahippocampal cortices bilaterally. In sum, there is now considerable evidence from lesion studies that the medial temporal lobes are involved in spatial behavior in humans as well as in other species, although the evidence seems to implicate the parahippocampal cortex rather than the hippocampus.

The results of recent functional imaging studies generally confirm these conclusions. Navigation in virtual environments was consistently found to activate the right parahippocampal gyrus and sometimes the right hippocampus (Aguirre et al., 1996; Maguire et al., 1998b,a), along with an extensive network of parietal and occipital cortical areas. The parahippocampal cortex was also activated by passive viewing of scenes (including rooms, buildings, and landscapes), but not faces or discrete objects (Epstein and Kanwisher, 1998; Maguire et al., 2001).

2.4.4 The role of the hippocampus in the formation of complex representations

Nonlinear discrimination

As described at the beginning of this chapter, the hippocampus has been suggested to play a crucial role in the formation of unitary representations for specific combinations (or conjunctions) of individual stimuli. Further, it has been proposed that this contribution of the hippocampus to representing stimulus conjunctions would make it essential for the learning of nonlinear discrimination problems. As discussed earlier, the negative patterning

task is a prototypical nonlinear discrimination problem, and several studies have confirmed that rats with lesions to the hippocampus are unable to learn the negative patterning problem (Rudy and Sutherland, 1989; Alvarado and Rudy, 1995a; McDonald et al., 1997). On the other hand, Davidson et al. (1993) failed to replicate these results, and the reason for this discrepancy is still unknown.

Another non-linear discrimination problem that has been intensively studied is transverse patterning, which requires animals to learn to choose A over B, B over C, and C over A. This task has also been found to be affected by hippocampal lesions (Alvarado and Rudy, 1995b,a; Dusek and Eichenbaum, 1998). On the other hand, Bussey et al. (1998) reported that fornix transections actually *facilitated* the acquisition of the transverse patterning task. However, as we have already discussed, there is both anatomical and functional evidence that fornix transection is not equivalent to direct lesions to the hippocampus or its cortical inputs. In fact, a direct comparison of the effects of hippocampal lesions and fornix transection on several configural discrimination tasks has consistently shown no impairment following fornix lesions, independent of the degree of impairment following hippocampal lesions (McDonald et al., 1997). Thus, the results on the negative patterning and transverse patterning problems are mainly consistent with the idea that the hippocampus plays a crucial role in learning to solve nonlinear discrimination problems.

However, some other findings flatly contradict this original, simple formulation of configural association theory. For example, hippocampal damage causes no impairment in a (non-linear) task seemingly quite similar to negative patterning, in which stimulus A by itself predicts reward and stimulus B by itself predicts no reward, and the presence of conditional cue C reverses the reward contingencies (i.e., A in conjunction with C predicts no reward and B in conjunction with C predicts reward) (Gallagher and Holland, 1992; Alvarado and Rudy, 1995a). For the closely related biconditional problem, in which stimulus A predicts reward and stimulus B predicts no reward in context C, and reward contingencies are reversed in context D, the results are completely mixed (McDonald et al., 1997). Since all of these tasks are clearly nonlinear, and are thought to require the ability to represent stimulus conjunctions, the hippocampus cannot be critical for this operation. This is in contradiction with configural association theory and, indeed, as pointed out by O'Reilly and Rudy (2001), with many other theories that (either explicitly or implicitly) assume that the hippocampus is critically involved in the processing of stimulus conjunctions. On the other hand, the hippocampus does appear to be necessary in many of the tasks described above, and there have been several attempts at modifying the theory so that it can successfully

accommodate more of the above data (Rudy and Sutherland, 1995; O'Reilly and Rudy, 2001).

Contextual effects

Closely related to the configural learning tasks described above are some of the contextual learning effects seen in conditioning. It has been observed in several different tasks in different species that the retention of a conditioned response can be much better when tested in the context used during initial training than in a different context. This effect can be absent in animals with lesions to the hippocampus (Good and Honey, 1991; Penick and Solomon, 1991).

In addition, hippocampal damage appears to prevent conditioning to the context where shock occurs in a fear conditioning paradigm, even though conditioning to the explicit cue predicting the shock is unimpaired (Kim and Fanselow, 1992; Phillips and LeDoux, 1992). Entorhinal and fornix lesions appear to have a similar effect (Maren and Fanselow, 1997), and so does blockade of NMDA receptors during training, independent of whether the animals are tested 1 day or 28 days after training (Fanselow et al., 1994). It is worth noting that in the case when the explicit cue is omitted or does not predict the shock, conditioning to the context does occur even in hippocampal animals (Phillips and LeDoux, 1994).

In contextual fear conditioning, hippocampal damage also has a time-dependent retrograde effect (Kim and Fanselow, 1992). The hippocampus was lesioned either 1, 7, 14, or 28 days after training. Contextual fear was completely abolished in the group that received the lesion one day after training, and the deficit compared to normals gradually decreased for longer time intervals. Conditioning to the tone was unaffected in all animals. More recently, these results were confirmed using a within-subject design (Anagnostaras et al., 1999), making unlikely a performance deficit account of the findings.

Flexible use of acquired representations

It has often been observed that hippocampal animals, even when they are apparently unimpaired at learning a particular task, do not (perhaps cannot) use the acquired information to make inferences about subsequent novel situations in the same way as normals (Eichenbaum, 2000). It has been suggested that memory representations acquired in the absence of the hippocampus are somehow hyperspecific, and do not allow inferences to be made

outside the exact context of the original learning experience (Eichenbaum, 2000). However, currently available data do not seem to rule out other, alternative explanations, for instance, that there is no significant difference between the memory representations acquired by normal and hippocampal animals, but the hippocampus is required during testing in a novel situation to make efficient use of this representation. New experiments (e.g., using reversible functional inactivation techniques) will be needed to distinguish between these possibilities.

One of the clearest demonstrations of the inflexible use of information by hippocampal animals is provided by the experiments of Dusek and Eichenbaum (1997). They trained rats on a series of overlapping olfactory discriminations; i.e., the rats were rewarded for choosing odor A over odor B, B over C, C over D, and D over E. Both normal and hippocampal rats could learn this set of discriminations well and equally fast. However, when subsequently tested on the novel stimulus pairing B vs. D, normal animals consistently chose B, showing that they were able to combine information from multiple previous experiences. In contrast, both animals with lesions to the entorhinal and perirhinal cortices and animals with fornix transection failed this test of transitive inference, showing no preference for either B or D in the probe trial. Similar results have been obtained using different tasks (such as paired associate learning; Bunsey and Eichenbaum, 1996).

2.5 Synthesis and open questions

In this chapter, I have provided a selective review of experimental data about the hippocampus. To what extent do these data allow us to answer two fundamental questions about the hippocampus, in particular, what is its function, and how does it implement this function? The facts that the data were obtained using a wide variety of techniques, and thus provide information about very different levels of neural organization, and that the data concern several different species, make it very difficult to assess how much the available data collectively constrain theories of hippocampal function. These complex relationships between data from different experiments is one of the reasons why computational modeling can be useful in understanding brain function; models can be used to link different levels by making explicit the consequences of data and assumptions at one level of description for higher levels.

The fact that the data come from experiments using animals as different as rats and primates raises another obvious question, namely, whether we can expect to reveal a single, unified

theory of hippocampal function which applies to all mammalian species, or whether we need several, species-specific theories. It is certainly true that most of the available data on rodents point towards a primarily spatial, and perhaps general representational role for the hippocampus, while data on primates (and especially humans) mostly indicate a role for the hippocampus in declarative memory. On the other hand, there are some signs that the hippocampus might be involved in all of these functions in all species. Another possibility is that although the hippocampus carries out essentially the same computation in all animals, it works in a different neural environment, getting different inputs and having its output interpreted in different ways, reflecting the representational needs and capabilities of the particular species. The assumption that the computations carried out by the hippocampus might be conserved is supported by the similarity of the anatomy and basic physiology of the hippocampal system in all species studied. However, most theories of hippocampal function to date are specific to a particular domain (i.e., space, representation, or memory). Therefore, an important problem for theoretical research is to determine whether the same basic (computational and biological) mechanisms can account for the data in different domains and species. It would be particularly relevant to find out whether ideas about memory functions in primates and ideas about spatial processing in rodents can be brought into register. Our model of CA3 place cells, which is described in Chapter 3 of this thesis, is a step in this direction, showing how the place cell representation might be a result of what has traditionally been considered as memory processing.

An important observation already alluded to above is that we can only hope to understand the function of the hippocampus if we consider it in relation to the other brain structures that it interacts with. We might be able to work out *how* the hippocampus operates by looking at it in isolation, but we cannot hope to find out *why* it works that way without taking into account the context in which it operates. Therefore, we need to think both about how the other areas connected to the hippocampus (neocortical as well as subcortical) work and what their functions might be, and about the ways they interact with the hippocampus and what computations all these areas might collectively engage in. Our initial effort in this direction, which is described in Chapter 4, concentrates on the interactions of the hippocampus with neocortex, and how they might collaborate in the storage and recall of long-term memories. However, the connections of the hippocampus with various subcortical structures, and their role in spatial processing (as well as memory) also needs to be addressed.

Chapter 3

A plastic recurrent network model of CA3 place cells[1]

The hippocampus is known to be involved in spatial learning and memory in rodents. Some of the most convincing evidence for this is the presence of place cells in areas CA3 and CA1 of the hippocampus (O'Keefe and Dostrovsky, 1971; O'Keefe, 1976), and of many other types of spatially selective cells in neighboring areas (Jung and McNaughton, 1993; Quirk et al., 1992; Frank et al., 2000). Principal neurons in CA3 and CA1 are active only when the animal is located in a well-defined local region of the environment (a place field; Muller et al., 1987), and collectively provide a population code for spatial position (Wilson and McNaughton, 1993). The question we address is how this comes to be in a way that is consistent with the evidence for the involvement of the hippocampus in more general forms of memory.

A key anatomical feature of area CA3 is that its pyramidal cells receive the majority of their inputs from other CA3 pyramidal cells (Amaral and Witter, 1989; Amaral et al., 1990). The resulting recurrent network has been extensively explored as a plastic attractor model of the way that the hippocampus acts as a general memory (Marr, 1971; McNaughton and Morris, 1987; Rolls, 1996; Levy, 1996; Hasselmo et al., 1996), but has been widely ignored by models that are intended to account for various properties of place cells (Zipser, 1985; Sharp, 1991; Touretzky and Redish, 1996; Burgess et al., 1997) (but see Battaglia and Treves, 1998).

[1]**This chapter has been adapted from (Káli and Dayan, 2000), and contains an extended discussion and some minor modifications.**

The model of Samsonovich and McNaughton (1997) was the first to explore the consequences of the CA3 attractor network for the place cell representation. Their model assumes the existence of a collection of independent continuous sets of attractors realized by the CA3 recurrent network, and successfully accounts for some of the basic experimental observations about place cells. However, in a model with fixed, independent sets of attractors, it is hard to explain the recent experimental findings by Skaggs and McNaughton (1998), who found partially overlapping place cell representations in two distinct but similar-looking parts of an apparatus. Such models generally predict either identical or completely different firing patterns in this situation. In addition, Samsonovich and McNaughton's (1997) model does not address the question as to how the strengths of the CA3 recurrent connections, which are essential for the existence of appropriate attractors, become established. As is critical for models in which the hippocampus acts as a memory, there is substantial evidence for synaptic plasticity in most major hippocampal pathways, including those providing feedforward inputs to area CA3 (Zalutsky and Nicoll, 1990; Breindl et al., 1994) and the CA3 recurrent collateral connections (Zalutsky and Nicoll, 1990; Debanne et al., 1998). These activity-dependent synaptic changes provide the obvious means for setting up the appropriate connection strengths, and, in conjunction with the attractor structure, thereby allow us to relate a major aspect of spatial processing to a major aspect of memory processing.

Brunel and Trullier (1998) and we (Káli and Dayan, 1998) independently implemented models which rely on modifiable recurrent connections in CA3 to explain the differences in the directionality of place cells in different kinds of environment. Some of this early work will be described in this chapter; some further details, particularly on an analytically solvable simple version of the model, can be found in Appendix A. However, the strongest challenge for models, and particularly models based on attractor networks, comes from data on the behavior of place cells in multiple environments that are similar, or are related by simple geometric manipulations. In this paper, we present an attractor model with appropriate behavior in these cases.

3.1 Place field formation in simple environments

Our model is grounded in two assumptions. The first is that observed place cell activity patterns reflect the stable states of the CA3 attractor network, a network whose dynamics is governed by its intrinsic recurrent excitatory connections supplemented by inhibitory feedback (see Figure 3.1). Inputs to CA3, arriving via learned feedforward connections from

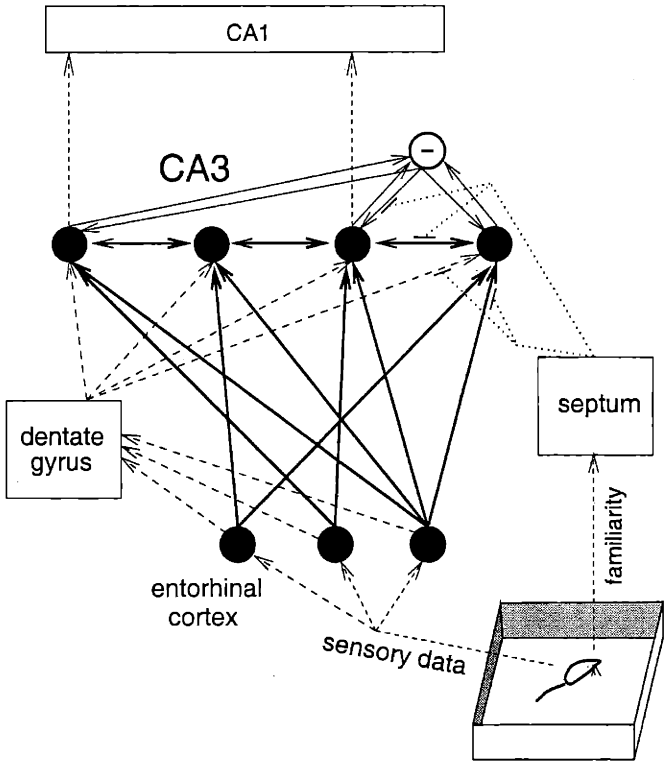


Figure 3.1: Model architecture. The inputs to the network are the activities of neurons in entorhinal cortex, which are determined by sensory features in the environment. This representation is then transformed by feedforward pathways (the direct perforant path connections to CA3 and the pathway through the dentate gyrus) and recurrent processing in area CA3, which involves lateral connections between CA3 pyramidal cells (filled circles) as well as their connections with an inhibitory neuron (open circle). The solid lines indicate neuronal connections that are modelled explicitly, and the thick ones (the CA3 recurrent connections and the perforant path inputs to CA3) are modifiable. Each type of connection is all-to-all in the model. All inputs to CA3 pyramidal cells are gated by neuromodulatory signals (dotted lines) from septal nuclei, whose activity depends on familiarity with the current environment. Note that we have not implemented neurons in CA1 or the connections between CA3 and CA1, although most of the place cell data actually come from recording in CA1.

entorhinal cortex (EC), are used to select among the stored attractors. We utilize experimental data as well as computational considerations to propose some general constraints on how the EC spatial representation may depend on sensory features of the environment, and also suggest a plausible functional form for this dependence in the simple case that all the information about location that is directly available comes from the walls of the experimental apparatus.

The second basic assumption is that the network establishes new attractors to represent novel situations. This involves an orthogonalization process which is assumed to take place in the dentate gyrus, as well as on-line modulation of synaptic plasticity and the relative efficacies of the different types of connections, controlled by familiarity with the environment, possibly via neuromodulatory signals from septal nuclei.

In this section, we provide a detailed description of the main components of our model, including the neural architecture (as shown in Figure 3.1) and dynamics, as well as the input representation. We then demonstrate the basic properties of the model by showing how place fields are generated in the simple case of a single environment surrounded by walls, using an idealized set of weights. In the next section, we tackle the issues related to learning, and introduce a familiarity-based on-line learning process for establishing an appropriate weight structure. The rest of the paper is devoted to modelling a set of more complex experimental paradigms. The values of the parameters used in the simulations are summarized in Table 1 at the end of the description of the model (on page 79).

3.1.1 CA3 neural architecture and dynamics

The main aspect of hippocampal circuitry we actually implement is the CA3 recurrent network (Figure 3.1). The model CA3 contains a collection of 1200 pyramidal cells, each connected to all the others through modifiable weights. This high degree of connectivity mimics the extensive recurrent collateral connections of CA3 pyramidal neurons (Ishizuka et al., 1990; Li et al., 1994). Due to the relatively small number of neurons in the model, the number of connections per cell is still much lower than in reality, even though the degree of connectivity is higher. This does not pose a problem, though, as long as the cells a particular neuron connects to can be considered from a functional point of view as a random sample, the number of connections per neuron is high enough, and any one connection is weak enough. In this case, neural responses are determined by averaged population effects, and

the actual number of connections only enters the calculations as a constant scaling factor for the individual weights.

Local feedback and feedforward inhibition are thought to play an important and complex role in neural dynamics in CA3. Inhibitory interneurons are spatially much less selective than pyramidal neurons, but their activity during locomotion changes periodically at the theta frequency. We ignore this temporal variation as well as the diversity of interneurons and patterns of connectivity, and include in the model a single global inhibitory neuron, which fosters competition between stored patterns and keeps global activity levels approximately constant. This cell receives input from all the excitatory neurons, and provides inhibitory feedback to each which is proportional to the product of the firing rate of the inhibitory neuron and the depolarization of its postsynaptic target. This nonlinear form of inhibition was chosen because our simulations indicated that, compared to more conventional subtractive inhibition, it leads to improved robustness in the network with respect to variations in weight magnitude (for details on networks with shunting inhibition, see Grossberg, 1988). It is also consistent with the observed effect of $GABA_A$ receptor activation. We adapt the equations introduced by Wilson and Cowan (1972) to model the dynamics of the CA3 neural population. The following set of equations describes how the membrane potential of CA3 cells in our model changes over time:

$$\begin{aligned} \tau \dot{u}_i &= -u_i + \sum_j J_{ij} g_u(u_j) - h g_v(v) u_i + I_i^{PP} + I_i^{MF} \\ \tau' \dot{v} &= -v + w \sum_j g_u(u_j), \end{aligned} \tag{3.1}$$

where u_i is the membrane potential of the i'th pyramidal cell, v is the membrane potential of the global inhibitory cell (all relative to their resting potentials), τ and τ' are the membrane time constants for pyramidal neurons and the inhibitory cell, respectively, J_{ij} is the strength of the connection from neuron j to neuron i, h is the efficacy of inhibition, w represents the strength of the excitatory connection from any one pyramidal cell onto the inhibitory cell, and I_i^{PP} and I_i^{MF} are the inputs to cell i through the perforant path and the mossy fibers, respectively. $g_u(u) = \beta[u - \mu]_+$ is the threshold linear activation function for the pyramidal cells, where $[...]_+$ makes all negative arguments zero while leaving positive numbers unaffected, μ stands for the threshold and β is the slope of the activation function above the threshold. Similarly, $g_v(v) = \gamma[v - \nu]_+$ for the inhibitory neuron. As will be described in detail later, some of the terms in these equations are assumed to be influenced by neuromodulatory control, and therefore may be absent in certain phases of processing.

The value of the inhibitory time constant τ' has no effect on the location of the fixed points of the network, although it can change their stability. In the simulations that are described later, we set $\tau' = 0$, so that v is always equal to $w \sum_j g_u(u_j)$. This simplifies the theoretical treatment of the model, and makes the simulations numerically more stable. We conducted simulations to verify that, within a wide range of the parameters, this manipulation does not affect the qualitative dynamical behavior of the model and indeed leads to the same stable patterns of activity. It is worth noting that, in this general class of models (although in a different parameter regime), setting $\tau' > 0$ can give rise to oscillations (which, of course, are consistently observed in the hippocampus during active behavior). Even in an oscillatory regime, however, the mean activities of the units can closely resemble the activities of the units at the fixed points found when $\tau' = 0$ (Li and Dayan, 1999).

3.1.2 Input representation

Instead of building a detailed model of rodent sensory processing, we consider as inputs to our model the firing rates of pyramidal neurons in superficial layers of entorhinal cortex, which provide most cortical input to the hippocampal formation. Unfortunately, there is relatively little direct experimental evidence about the nature of spatial representations in EC, and especially about how these depend on details of the environment. However, there is something of a consensus amongst modelers (e.g., Burgess et al., 1997), which we generally follow. Although entorhinal neurons are found to be spatially selective (Barnes et al., 1990; Quirk et al., 1992; Frank et al., 2000), they appear to be much noisier and more broadly tuned than place cells in the hippocampus. Quirk et al. (1992) also found them to be more "sensory bound" than hippocampal cells in that their firing fields transform in a smooth manner following substantial changes in the shape of the environment. This is very unlike the complete remapping seen in place cells under similar circumstances (Muller and Kubie, 1987). The anatomy of the inputs to EC is rather better understood (Burwell and Amaral, 1998). Many of the inputs to EC come from higher-order association areas, which contain complex representations of the sensory information available to the animal. In particular, cells may convey information about both the identity of a perceived object and its location with respect to the animal, or, to put it differently, about the location of the rat with respect to particular objects in the world. Such information about multiple objects may be combined in EC in order to form a more reliable view-based representation of the animal's location in space. Spatial information derived from path integration may also be available, and may be combined with visual information to determine EC activities.

In the model, each EC cell is assumed to respond to a subset of the available cues. Based on the suggestion that EC is involved in conjunctive coding (Myers et al., 1995), each EC cell in our model combines in a conjunctive manner the sources of spatial information to which it is sensitive. Since the animal's sensory experience depends on both its position and the direction it faces, we assume (in the absence of data either way) that the activity of entorhinal neurons is head direction as well as location dependent. A model EC cell fires maximally when all the cues it is sensitive to are in the position corresponding to the cell's preferred location and orientation, and activity diminishes as some or all of the sources of information signal a different location or orientation. We achieve this by multiplying together Gaussian tuning curves, each of which is tied to the location of a different cue and peaks at the preferred location of the cell. We assume that these individual tuning curves can have different variances.

In cases where the environment has walls, these were found to be important sources of spatial information (O'Keefe and Burgess, 1996). For simplicity, we assume that the activities of EC neurons are completely determined by the rat's position and heading relative to the walls. We also restrict ourselves to rectangular environments, and assume that all cells are sensitive to the position of all four walls (whose allocentric bearings will be referred to as 'north', 'west', 'south', and 'east'). The only difference in the cue selectivity of EC cells in our model is that they are assumed to be sensitive to spatial information derived from path integration to different degrees. However, since this last property is only expected to be manifested under special circumstances, we actually ignore this variation in most of what follows, and only consider it when we describe the results of our modelling of the experiment of Skaggs and McNaughton (1998). We assume that the tuning curve components tied to the walls of a rectangular apparatus are ridge-like functions with Gaussian dependence on the distance from the wall. The variances of these tuning functions may also depend on the location and heading of the animal – in particular, we assume that the variance is lower if the animal is closer to, or facing *away* from the wall. The latter dependence is based on the influence of a path integration input whose precision is greater when the animal is coming from somewhere nearer the wall and should have been able to maintain its location accurately using path integration.

The total activation of a model EC neuron as a function of the rat's location and heading is described by the following expression:

$$z_k = b\, z_k^N z_k^W z_k^S z_k^E e^{\rho_{EC}\cos(\phi-\phi_k^{EC})}, \tag{3.2}$$

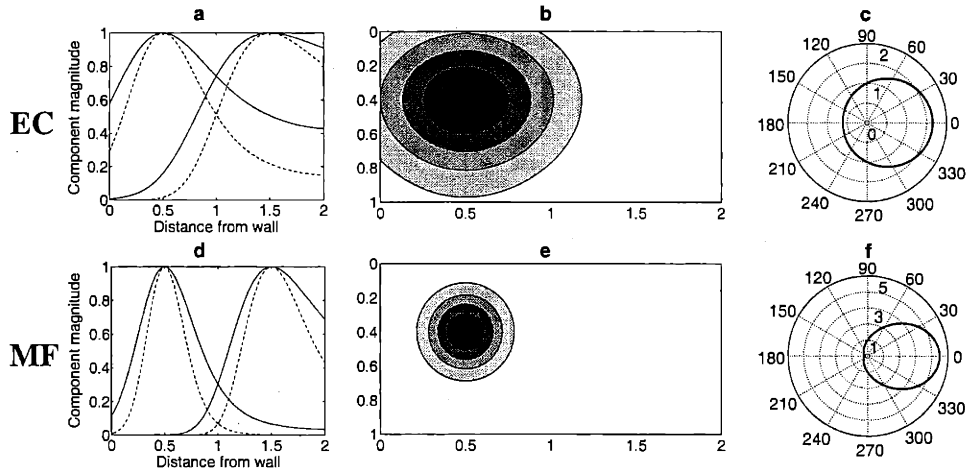


Figure 3.2: Input components and net spatial and directional tuning. **a**: The dependence of a single (essentially one-dimensional) spatial component of the tuning function of cells in entorhinal cortex (EC) on the distance of the rat from the wall to which that component is tied. Two examples are shown, with preferred distances of 0.5 and 1.5, respectively; for each preferred distance, the solid curve is for the case when the rat is facing the wall, and the dashed curve is for the opposite head direction. Note how the width of the curve changes with preferred distance and actual head direction. **b**: The net two-dimensional tuning of a sample EC neuron in a rectangular box of dimensions 2×1; the preferred location of the cell is (0.5, 0.4); the current heading of the rat at each location is the same as the preferred head direction of the cell. **c**: This polar plot shows the activity of an EC neuron as a function of the difference between its preferred direction and the actual heading of the rat. **d** to **f**: Plots similar to **a-c**, for the mossy fiber (MF) inputs to CA3; note that both the spatial and the directional tuning is much sharper here, due to the orthogonalization property of the dentate gyrus. For all contour plots in this article, darker shading indicates higher activity, and the contour lines are at 20, 40, 60, and 80% of the maximum activity of the given cell or set of cells. Activities are normalized and the absolute values are omitted in most figures since these could be set arbitrarily in the model by changing parameters essentially unconstrained by experimental data.

where k indexes the neuron, b is a constant to set the scale, and z_k^a is the component of the neuron's tuning function tied to wall a. The last term describes the dependence on head direction (which is assumed to be independent from the spatial components) as a circular Gaussian function (with sharpness parameter ρ_{EC}) of the difference between the current head direction ϕ and the cell's preferred heading ϕ_k^{EC}. Equation 3.2 bears some resemblance to the spatial tuning function used by Touretzky and Redish (1996), in that it also takes the form of a product of terms corresponding to different sources of information. However, they use this tuning function to directly model the spatial response properties of hippocampal place cells, and the parameters change with experience, while our EC representation is always the same for a given location and head direction in any particular environment.

The components of the tuning function tied to particular walls have the following functional form:

$$z_k^a = e^{-\frac{(d_a - d_k^{EC,a})^2}{2\sigma_{EC,a}^2}}, \tag{3.3}$$

where d_a is the actual distance from wall a (a can be either N, W, S or E), $d_k^{EC,a}$ is the distance from wall a of the neuron's preferred location, and $\sigma_{EC,a}$ is the width of this component, which depends on the current position and heading of the animal according to

$$\sigma_{EC,a} = \sigma_{EC}(1 + 0.35d_a^2)(1 + 0.2\cos(\phi - \phi_a)), \tag{3.4}$$

where ϕ_a is the direction of wall a (0, $\pi/2$, π, and $-\pi/2$ for N, E, S, and W, respectively), and σ_{EC} is a constant. Equation 3.3 and the positional dependence in Equation 3.4 are similar to the expressions describing the spatial tuning of "sensory" cells in the model of Burgess et al. (1997), and "boundary vector cells" in Hartley et al. (2000). The numerical values of the parameters in the above equations have been chosen suitably for environments of around the size employed in most relevant experiments.

Figure 3.2a shows two examples of the spatial and directional dependence of input components in EC, while Figure 3.2b and c display the resulting net spatial and directional tuning for a sample EC neuron.

3.1.3 Feedforward connections

There are two separate neural pathways from EC to area CA3 (see Fig. 3.1), which have quite different characteristics and likely serve different computational purposes (McNaughton and Morris, 1987; Treves and Rolls, 1992). One of these pathways is via the perforant path projection to the dentate gyrus (DG), which in turn provides a set of feedforward inputs to CA3 through the mossy fibers. Dentate granule cells are spatially selective, and, at least in linearly restricted environments, they have also been found to be sensitive to direction (Jung and McNaughton, 1993). Unlike EC neurons, dentate granule cells have sharper spatial tuning than CA3 place cells, and we assume that they are also sharply tuned for head direction. Episodic memory theories of hippocampal function suggest that an important function of the DG is that of orthogonalization, i.e. reducing the similarity between input patterns in order to facilitate their discrimination (Treves and Rolls, 1994; O'Reilly and McClelland, 1994), and, in keeping with the theme of linking memorial and spatial processing, we assume it plays a similar role for spatially based inputs. One way the DG is thought to decrease pattern overlap is to implement a sparser representation (perhaps through direct competitive interactions), and indeed, the proportion of active cells in the DG at any given time is reported to be only about 0.5% (B.L. McNaughton, cited by O'Reilly and McClelland, 1994; see also Jung and McNaughton, 1993).

A typical CA3 pyramidal cell receives on the order of 50 mossy fiber (MF) inputs, which are thought to be relatively powerful (Yamamoto, 1982; McNaughton and Morris, 1987). Combined with the sparseness of the DG representation, this means that a CA3 neuron is very unlikely to have more than one active mossy fiber input at any given time. In circumstances under which CA3 cells are driven primarily by these inputs, place cells essentially inherit the tuning characteristics of their afferent granule cells. We assume, for simplicity, that each CA3 cell has at most one active MF input in any given environment. This defines the base preferred location and direction for that neuron, which, of course, may then be altered by the recurrent connections in CA3. Multiple active MF inputs may explain why some place cells have multiple place fields even in simple environments (Muller et al., 1987); however, we ignore this complexity for the purpose of this paper. In addition, in order to make better use of the limited number of cells we can implement in our numerical simulations, all our model CA3 pyramidal cells are activated by MF inputs somewhere in any given environment, rather than the 30% or so found in practice (Wilson and McNaughton, 1993).

In its current form, the model considers both the mossy fiber connections and the perforant path connections from EC to DG as being fixed. Since our goal is to model activity in CA3, and that is completely determined by its inputs and internal dynamics, we can therefore skip modeling the dentate gyrus explicitly, and proceed by characterizing how the MF input to CA3 (which results from processing in DG) depends on the characteristics of the environment. We assume that, for any single environment, the MF input to CA3 place cells has a similar functional form to the tuning function of EC cells described in the previous section, but both the spatial and the directional tuning is assumed to be sharper as a result of sparsification and orthogonalization in DG (see Fig. 3.2d-f). This can be achieved by replacing the spatial spread parameter σ_{EC} with a smaller value, σ_{MF}, and by replacing ρ_{EC}, characterizing the sharpness of directional tuning, with a larger ρ_{MF} in Equations 3.2 and 3.4. The proposed orthogonalization property of the dentate gyrus becomes more pronounced when we look at multiple environments. We assume that, except when two environments are quite similar, the MF inputs to CA3 in two different environments are completely unrelated. We will return to the case of exceptionally similar environments in a later section.

The perforant pathway (PP) also provides a direct connection between EC and CA3, and has a large degree of divergence and convergence. Thus, CA3 cells can sample the EC representation very effectively. In the model, we implement this property using all-to-all connections between EC and CA3 neurons, although this is obviously a simplification. This pathway is also known to be capable of long-term synaptic plasticity (Breindl et al., 1994). In the model, the strengths of these connections (denoted by W_{ik} for the connection from entorhinal cell k to CA3 cell i) are initially set to zero, and they are assumed to be modifiable by associative Hebbian learning.

3.1.4 Network dynamics

Although we will shortly be interested in the spatial representation that results from on-line learning during exploration, we first test our model using an idealized set of connection strengths in order to gain some insight into its dynamical behavior. For this, we just assume that the weights result from an idealized form of Hebbian associative learning, and thus reflect the correlations between connected neurons. It has been noted (Muller et al., 1991b; Shen and McNaughton, 1996) that such an associative learning process for spatially selective neurons can lead to connections whose strength is a function of the distance

between the preferred locations of the pre- and post-synaptic neurons, exactly the sort of connections that can support a place-field-like attractor structure in CA3 (Samsonovich and McNaughton, 1997). Here we assume that the CA3 recurrent weights are determined by the correlations between the mossy fiber inputs to the cells, and the perforant path weights between EC and CA3 are given by the correlations between EC activities and MF inputs to CA3. These correlations are calculated as spatial averages (which assumes spatially homogeneous exploration) over all locations and head directions in the part of the environment where the postsynaptic cell is active, resulting in the following expressions for the recurrent weights J_{ij} and perforant path weights W_{ik}:

$$J_{ij} = \kappa \frac{\iiint\limits_{I_i^{MF}>0} I_i^{MF} I_j^{MF}\, dx\, dy\, d\phi}{\iiint\limits_{I_i^{MF}>0} dx\, dy\, d\phi} \quad \text{and} \quad W_{ik} = \kappa \frac{\iiint\limits_{I_i^{MF}>0} I_i^{MF} z_k\, dx\, dy\, d\phi}{\iiint\limits_{I_i^{MF}>0} dx\, dy\, d\phi} \tag{3.5}$$

where I_i^{MF} is the mossy fiber input to neuron i in CA3, z_k is the activity of neuron k in entorhinal cortex, and κ sets the learning rate.

Using these expressions, we can calculate the weights resulting from even exposure to a rectangular box (with one side twice as long as the other). Then, letting $I_i^{PP} = \sum_k W_{ik} z_k$ and $I_i^{MF} = 0$, where z_k is the EC activity pattern corresponding to a particular location and heading in the environment, we simulate the neural dynamics described by the full Equations 3.1 for a fixed number of iterations (using Euler's method). We find that, within a broad range of model parameters, the network always settles into a stable state by the end of the iterations. Furthermore, for most initial CA3 activity patterns, the same final state is reached for given feedforward inputs. This shows that these states are actually attractors of the neural dynamics, and that they have suitably large basins of attraction. The final state of the network was determined for different input patterns in EC, representing different positions and head directions of the animal over a grid that covered the whole environment. The firing rate map for a given cell is defined as the final activity of that cell as a function of the actual location and head direction of the animal.

Throughout the paper, two different kinds of plots are used to display the activities of neurons (and their inputs). Quantities characterizing single cells as a function of actual position and heading (such as firing rate maps) are shown in "single cell plots", which is the kind of plot traditionally used to describe the spatial activity patterns of place cells. A single cell plot may contain multiple subplots to represent different headings at any given location. The second kind of plot we use is the "population plot", which describes the behavior of *all*

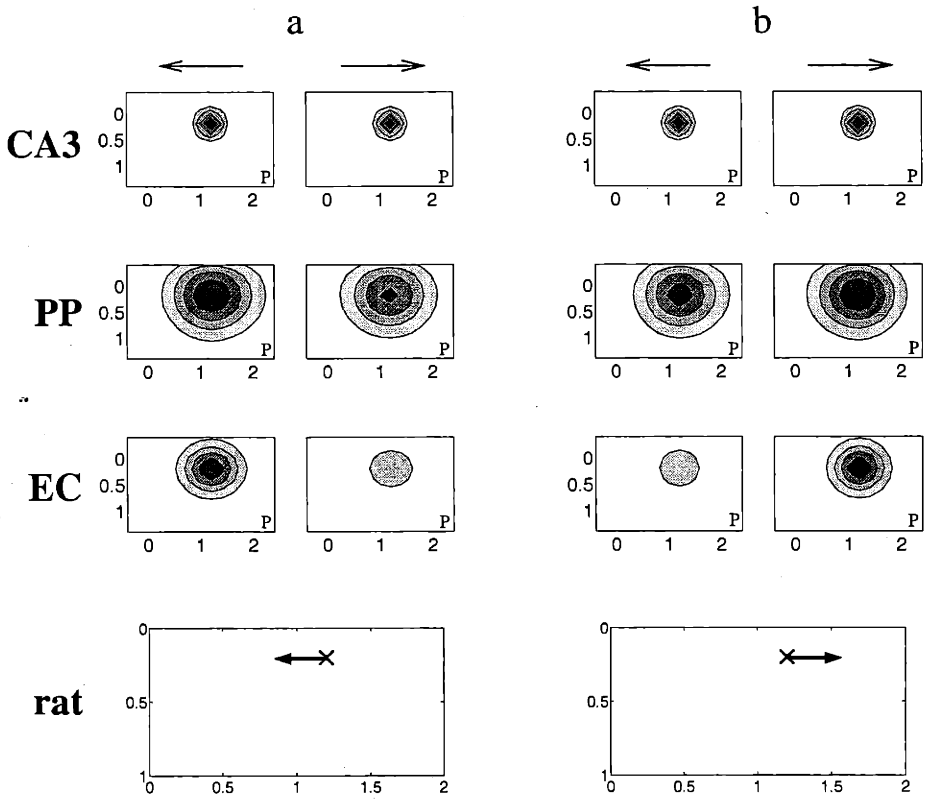


Figure 3.3: The formation of non-directional place fields. **a,b**: The bottom plot in each case shows the actual position (indicated by the cross) and head direction (indicated by the arrow) of the rat in the environment. The other plots are population plots (as defined in the text, and marked with 'P'), and they show, at the location and direction in the bottom plot, the activities of cells in entorhinal cortex (EC), the net perforant path (PP) inputs to CA3 neurons (I_i^{PP}), and the final activities of the same place cells (marked CA3), as a function of the preferred location of the neuron; the two columns in both **a** and **b** are for cells with preferred head direction indicated by the arrow above each column.

the cells with the actual position and heading of the animal kept fixed. In the population plot, we arrange cells with a given preferred direction on an imaginary plane according to their preferred locations (for CA3 place cells, this is defined as the preferred location of their active mossy fiber input). A complete population plot would include 8 subplots, one for each population of cells with a different preferred direction, but we typically show only 1, 2, or 4 of these, depending on the degree of variation with preferred direction in that particular case. Population plots in this paper are marked with 'P' in the lower right corner for easy identification.

The results of the simulations with the "ideal" weights are summarized in the population plots of Figure 3.3, which display activities in EC, net perforant path inputs and final activities in CA3 for all cells with two particular (opposite) preferred head directions, when the model rat is at a given location, facing in a particular direction. Figure 3.3 shows that the final states of the model CA3 network resemble thresholded two dimensional Gaussian bumps of activity in the population plot. This type of solution can emerge spontaneously from the network dynamics even in the absence of external inputs, in which case the location of the bump is random – i.e., determined by the initial neural activities, as well as various other factors including the distribution of preferred locations and directions of the neurons. Even though the network only has a finite number of point attractors (possible stable activity patterns) in the absence of input, when there is even a small perforant path input to CA3, the location of the bump is determined by this input so that the activity profile provides the best possible fit to the input. The position of the peak varies continuously, and the shape of the activity profile is essentially constant. This holds in our model if the net feedforward input to the most active CA3 neurons is between roughly 1% and 30% of the summed input they receive from other CA3 cells; in most simulations we set the relative efficacies of perforant path and recurrent synapses so that this ratio is about 5%.

Figure 3.3a illustrates how inputs are used by the network to effectively select one of the possible final states. First of all, the EC activity pattern (which is determined by sensory features in the environment as already described) gives rise to a pattern of perforant path inputs to CA3 which is centered on neurons with preferred locations close to the actual position of the rat, although the profile is even broader than the activity profile in EC. This is the consequence of plasticity of the perforant path in the learning phase, which establishes an association between EC cells and CA3 neurons with similar preferred locations and head directions. Based on the learned weights, the PP projection also reduces directionality substantially, so that inputs to CA3 already depend less on the preferred head direction of

the cell than neuronal activities in EC. The shape of the final activity profile across place cells is, however, essentially determined by the CA3 internal dynamics, resulting in a spatial activity profile which is much more sharply peaked than the feedforward inputs. Further, the final activities of the cells are essentially independent of their preferred head direction. All in all, the relation of the final CA3 activity pattern to the EC input reflects, first, the action of the learned feedforward projection from EC to CA3 (which is sometimes referred to as heteroassociation), and, second, the attractor dynamics of the CA3 recurrent network (which is biased by the feedforward input, and thus can be seen as an autoassociative operation). The resulting model place fields (i.e., single cell activity maps) possess many of the characteristics of real place cell firing patterns recorded in open environments. As we will see later (e.g., in Figure 3.6a), they are unimodal, approximately Gaussian with circular symmetry, and essentially non-directional.

Figure 3.3 reveals how non-directional place fields result despite the directional input representation in EC. The two parts of the figure compare the activities of EC neurons, the PP inputs to CA3 place cells, and the final activities of place cells as the model rat faces in two opposite directions at the same location. Due to the properties of the PP projection discussed above, place cells receive relatively similar inputs in the two cases. More importantly, however, this leads to the emergence of the same, non-directional, attractor in CA3, making the place fields independent of head direction. It should be emphasized that, given the dominance of internal connections in determining the final state of the system, even PP inputs as similar as those in Figure 3.3 a and b could easily lead to fundamentally different patterns of final activity if the two input patterns biased the system towards different attractors. Indeed, these same two EC input patterns do actually give rise to two very dissimilar final patterns if the weights are set up during a directed search task like the one described later (instead of the omnidirectional random exploration assumed here).

3.2 On-line learning of attractors

So far we have assumed that weights proportional to the spatially averaged correlations between cells had been established by an appropriate learning procedure before spatial activity patterns are measured. We have not yet shown that a neurobiologically standard Hebbian learning rule, applied to the activity patterns occurring in the network during random exploration of an environment, is capable of establishing this kind of weight structure,

within the time window during which place fields are seen to develop in experiments (on the order of 5 minutes; Wilson and McNaughton, 1993).

A general property of attractor networks is that, in order to store more than a single pattern, the recurrent connections need to be suppressed while new patterns are learned. Experimental data and theoretical considerations have been adduced to justify models of CA3 in which the relative strengths and adaptability of mossy fiber input and perforant path and recurrent collateral input is different between initial learning about an environment and recall of information within a familiar environment. We adopt the suggestion of Hasselmo et al. (1996) which is based on experimental data on the effects of septal (cholinergic and $GABA_B$-receptor-mediated) modulation in the hippocampus.

In particular, substances that activate muscarinic cholinergic receptors or $GABA_B$ receptors in the hippocampus were found to selectively suppress excitatory recurrent synapses in area CA3 compared to feedforward excitatory connections (Ault and Nadler, 1982; Hasselmo et al., 1995). In addition, cholinergic input to the hippocampus has been shown to enhance long-term synaptic plasticity (Burgard and Sarvey, 1990; Huerta and Lisman, 1993), and leads to the suppression of inhibition Pitler and Alger (1992) and the direct depolarization of hippocampal pyramidal neurons (Benardo and Prince, 1982). These effects of cholinergic modulation create exactly the right circumstances for the learning of new information in the hippocampus while minimizing interference from previously stored information. This is convincingly illustrated by the associative memory model of Hasselmo et al. (1995), where several moderately overlapping input patterns can be stored and recalled successfully using feedback cholinergic modulation of network parameters.

It turns out that attractor networks with continuous attractors, such as ours, face a more stringent requirement for learning because of potential bias in the sampling of a continuous set of input patterns, and we therefore consider a slightly different model of neuromodulatory control. In the resulting on-line learning procedure, plasticity is gated by familiarity, and we show that it leads to weights similar to those in the "ideal" model described above, and thus a place cell representation similar to ones observed experimentally.

In our model, the hippocampal network has two modes of operation. When the rat first encounters a new environment, learning in both the PP inputs to CA3 and the CA3 recurrent synapses is enabled, synaptic transmission through the recurrent connections is suppressed, inhibition in CA3 is reduced, and inputs through the mossy fiber connections dominate. This state of the network is called "learning mode". On the other hand, when the rat is in

a highly familiar environment, no learning takes place in any of the connections, the MF inputs are relatively less effective than the PP connections and CA3 recurrent synapses, and the intrinsic dynamics of the recurrent network dominates activity in CA3, leading to previously established attractors. This is called "recall mode".

Initial learning in a novel environment is essentially input-driven due to the suppression of recurrent activity, but this phase is responsible for setting up the attractors and feedforward associative projections which determine the patterns of place cell activity seen subsequently. Note that synapses are modified even when their efficacy is reduced to zero by neuromodulation, i.e., when the post-synaptic effect of perforant path and recurrent connections is negligible in the learning phase.

Perforant path and recurrent weights are acquired during the learning phase. The neural dynamics described by Equations 3.1 is simplified substantially in this phase by making the recurrent connections ineffective and neglecting inhibition, leaving $\tau \dot{u}_i = -u_i + I_i^{MF}$. Assuming that the MF inputs change more slowly than the membrane time constant, the membrane potential of CA3 place cells during the learning phase is given by $u_i = I_i^{MF}$. Application of a Hebbian learning rule (with the addition of weight decay to prevent weights from growing indefinitely) to these activities leads to weights which are proportional to the temporally averaged correlations between pre- and postsynaptic cells. The only difference between these weights and the "ideal" ones used in earlier simulations is that while the "ideal" weights were obtained by averaging the product of pre- and post-synaptic activities across spatial locations and headings, this on-line method calculates averages across time. The two processes become exactly equivalent if we assume that, during initial exploration in the environment, the rat receives even exposure to all combinations of location and head direction allowed by the apparatus (and the movement pattern followed). An early version of the model based on this and many other simplifying assumptions is analyzed in Appendix A.

However, there are two potential differences between uniform spatial averaging and temporal averaging over random exploration, the first coming from any systematic spatial bias (which depends on the exploration strategy), and the second coming from random deviations from this ideal, if biased, exploration. Figure 3.4a shows a sample path from a simulation of a common exploration paradigm (which is essentially equivalent to experiment 1 in Markus et al. (1995)). This shows the first five minutes of exploration in a new environment whilst a rat chases food pellets thrown into random locations in a rectangular apparatus. Once it has retrieved one pellet, the next one is thrown in at random. We assume that the rat

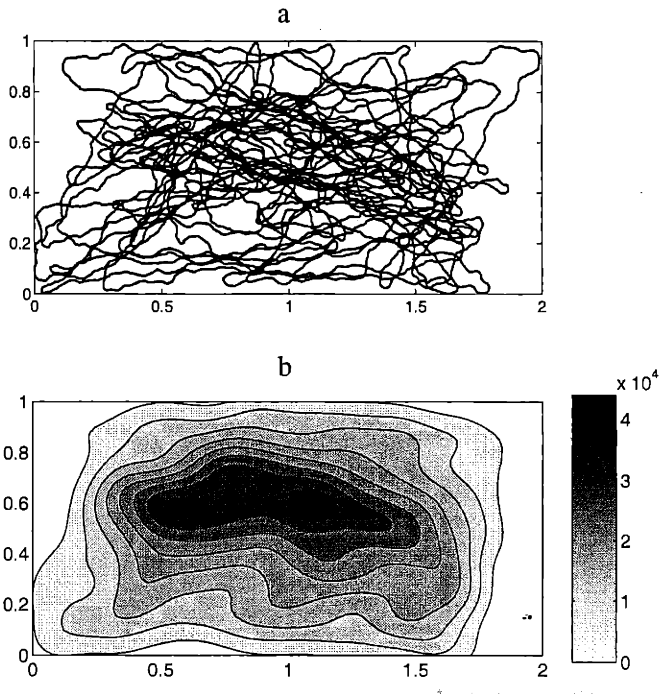


Figure 3.4: Non-uniform sampling of the environment during random exploration. **a**: An example trajectory, showing the first 5 minutes of exploration in our simulation of a common paradigm in which the rat chases food pellets thrown into random locations in the environment. Note that some parts of the environment are visited much more often than others. **b**: Convolution of the path in **a** with a 2-dimensional Gaussian ($\sigma = 0.075$), which measures exposure to locations in the apparatus ($g(\mathbf{x}^*, \phi^*, t)$), summed over all directions, after 5 minutes of exploration.

runs at a constant speed V, and it always heads essentially in the direction of the next food pellet, with random fluctuations in direction. The exact movement laws and parameters were taken from the model of Brunel and Trullier (1998).

Even at a first glance, this exploration strategy clearly results in an inhomogeneous sampling of the environment. We quantify variations in exposure to different locations and directions in the apparatus by convolving the sample path with a Gaussian, yielding the function

$$g(\mathbf{x}^*, \phi^*, t) = \int_0^t e^{-\frac{(\mathbf{x}(t')-\mathbf{x}^*)^2}{2\sigma^2} + \rho\cos(\phi(t')-\phi^*)}\, dt' \tag{3.6}$$

where t is time since the beginning of exploration, $\mathbf{x}(t)$ and $\phi(t)$ are the rat's position and heading at time t, σ is the width of the spatial Gaussian, and ρ is the sharpness of

the circular Gaussian applied to differences in direction. This measures sampling density as a function of position and direction, and an example (after 5 minutes of exploration, averaged over all directions) is shown in Figure 3.4b. There is clear deviation, both random and systematic, from a uniform sampling density. The random aspect of the deviation turns out to be benign, since it does not destroy the overall structure of the attractors. However, the fact that, on average, the animal spends several times as much time at a location near the center of the apparatus than at a location near the edges, causes the naive on-line Hebbian learning procedure to produce a non-uniform weight structure, resulting in a very poor place cell representation. An example of this is given in Figure 3.5a – the network possesses just two or three distinct attractors, and only neurons which are active in one of these attractors ever become active in this environment. This effect cannot be mitigated by increasing exploration time, and is also persistent with respect to the specifics of the movement laws. In particular, even though rats have a tendency to stay close to the walls of the apparatus (e.g., Muller et al., 1987), this is unlikely to precisely counterbalance the effect described above and result in spatially and directionally unbiased exploration.

Systematic differences in sampling density have a profound effect on the resulting attractor structure because of the continuous nature of the set of patterns that need to be represented by the network. This requires the set of recurrent weights to be such that the activity patterns corresponding to all different positions in the environment are equally stable. Continuous attractor networks like ours are generally known to be very sensitive to the regularity of the recurrent weight structure (Zhang, 1996; Pouget et al., 1998), and most such previous models were forced to set these weights by hand.

Using all patterns indiscriminately during on-line learning is also questionable from a computational point of view, especially in the presence of substantial sampling bias. Learning should be gated by familiarity – the more familiar a part of the environment, the less about it that should be learned. Figure 3.4a shows that familiarity is actually a graded quantity – since the animal has more exposure to the center of the environment than the perimeter. Therefore, we use a graded familiarity signal, like the one proposed by Hasselmo et al. (1995). Note that it is not clear how familiarity is measured; for instance, Hasselmo et al. (1996) even suggest that a feedback loop involving the septal nuclei and the hippocampus itself might be responsible. We adopt the simple procedure of using the exposure measure of Equation 3.6, gating learning according to

$$\dot{J}_{ij} = e^{-\alpha g(\mathbf{x}(t),\phi(t),t)}\, u_i(t)\, g_u(u_j(t)) \quad \text{and} \quad \dot{W}_{ik} = e^{-\alpha g(\mathbf{x}(t),\phi(t),t)}\, u_i(t)\, z_k(t) \tag{3.7}$$

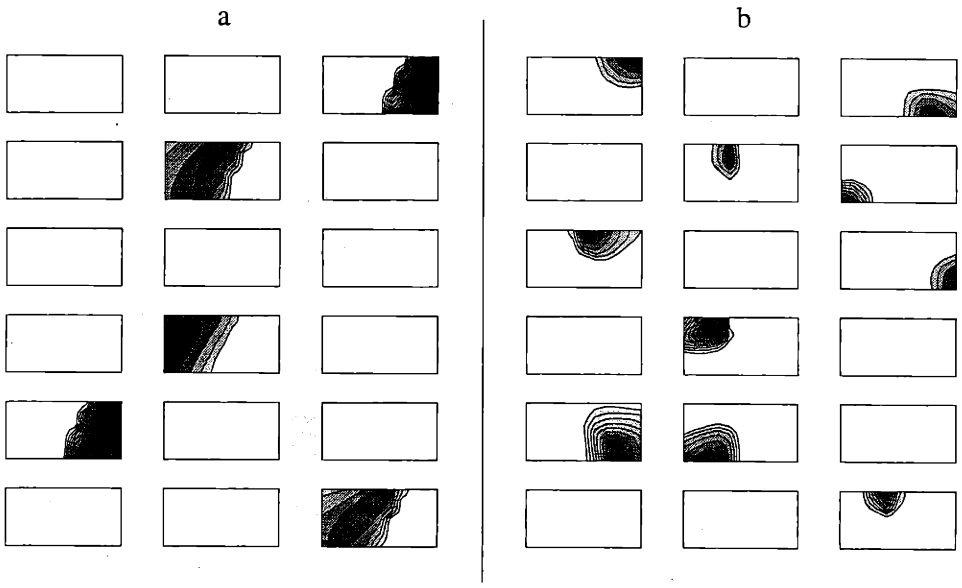


Figure 3.5: Place cell firing patterns during recall, after using different learning procedures. The figure shows the firing rate maps of 12 randomly selected CA3 place cells after the exploration shown in Figure 3.4a, **a**: using simple Hebbian learning, and **b**: using the familiarity-based learning procedure to establish the weights. The place fields in **(b)** closely resemble experimental place fields, and provide good coverage of the whole environment. Conversely, the spatial firing patterns in **(a)** reflect essentially two different attractor states containing only a small proportion of the neurons, perturbed to some extent by the feed-forward input.

where σ, ρ and α were determined so that the amount of learning that occurs in different parts of the apparatus is as uniform as possible after 5, or more, minutes of exploration. The application of this learning procedure results in an attractor structure not very much different from the one defined by the "ideal" weights described earlier, and, as shown in Figure 3.5b, leads to a good place cell representation after just 5 minutes of exploration, in agreement with experimental data. The weight structure becomes increasingly uniform, given more exploration, and the place fields duly become increasingly regular.

Although the efficacies of the different types on inputs to CA3 cells may also be modulated in a graded fashion (this may even involve the same signals which modulate plasticity), we currently use a simple heuristic based on the notion of two distinct processing modes as described above. Recall mode is entered after a fixed amount of exploration per unit area of the environment (by which time learning has saturated essentially everywhere in the environment), or immediately upon entry into the environment if it is similar enough to

Number of CA3 place cells	N	1200
EC tuning amplitude	b	100
EC spatial tuning width	σ_{EC}	0.4
EC directional sharpness	ρ_{EC}	0.5
MF spatial tuning width	σ_{MF}	0.2
MF directional sharpness	ρ_{MF}	1.5
Excitatory time constant	τ	100
Inhibitory time constant	τ'	0
Inhibitory feedback weight	h	3
Pyramidal-to-inhibitory weight	w	0.005
Excitatory gain	β	1
Excitatory threshold	μ	80
Inhibitory gain	γ	1
Inhibitory threshold	ν	12
Running speed	V	0.3
Spatial spread of familiarity	σ	0.075
Directional spread of familiarity	ρ	3
Novelty decay rate	α	250

Table 3.1: Model parameters. The table displays the values of model parameters used in the simulations. σ_{MF}, σ_{EC}, and σ are in units such that the shorter side of the rectangular environment used in most simulations is of unit length. τ and τ' are given in time steps used during simulations of recall, and all other quantities are in their natural units. Note that since the parameters only appear in certain combinations in the equations, some groups of parameters can be changed together appropriately without affecting the behavior of the model.

an environment already explored; otherwise, learning is initiated. More precisely, we skip learning in a new environment only if it shares most sensory features with an environment which is completely familiar to the animal, i.e., one that has been thoroughly explored.

3.3 Modeling more complex paradigms

So far we have shown that an attractor-based model, using weights defined by correlations between the feedforward activations of cells, can account for many of the experimentally observed basic properties of the CA3 spatial representation. We have also described a two-mode on-line learning process which computes an approximation to these "ideal" weights, and results in a very similar, although slightly less regular, place cell representation. In this section, we show how our model can also account for experimental results in a number

of more complex paradigms, including the task-dependence of place field directionality, the co-existence of several orthogonal representations for very different environments as well as overlapping representations for very similar environments, and the transformations of place fields following manipulations of the environment. We ran all simulations using both idealized, correlation-based weights and those resulting from on-line learning, and got qualitatively similar results in all cases. Most figures display results obtained using the "ideal" weights, because these tend to illustrate our points more clearly due to the lack of randomness.

3.3.1 Task-dependence of directionality

We have already described how random exploration in an open environment can lead to non-directional place fields (an example of which is shown in Figure 3.6a), through the establishment of appropriate attractors in CA3. In agreement with the recent modelling study by Brunel and Trullier (1998), we found that the ability of the recurrent network to suppress the directionality of the inputs depends critically on the set of locations and head directions experienced by the rat during learning. Place cells become direction independent only in situations in which the animal is exposed to a wide range of directions at a particular location. On the other hand, when the behavioral task or the environment itself constrains the set of directions experienced at a given location, as in a radial maze or when the rat is required to follow a specific route in an open field, place cells retain their intrinsic directionality. Even in these cases, the width of directional tuning can, however, be modified by the recurrent network. These results are in good agreement with experimental findings (Muller et al., 1994; Markus et al., 1995). The dependence of directionality on movement patterns is illustrated in Figure 3.6b, which shows the place field of the same model CA3 cell that appears in Figure 3.6a, for a rat which has performed a different behavioral task in the same environment. In this task, which can be thought of as a simplified version of the directed search task described by Markus et al. (1995), the rat is required to run back and forth between the two shorter walls of the environment to obtain reward. For the idealized case, we model this by assuming during exploration that the rat is now exposed only to the two directions parallel to the long walls instead of all directions at each location. Everything else in the simulations is left the same. This change affects the correlations between place cells in the learning phase, resulting in altered weight structure, which, in turn, changes the attractors. In agreement with experimental data, the new attractors do not eliminate the directionality of the inputs to the place cells. In fact, two very distinct sets of attractors are

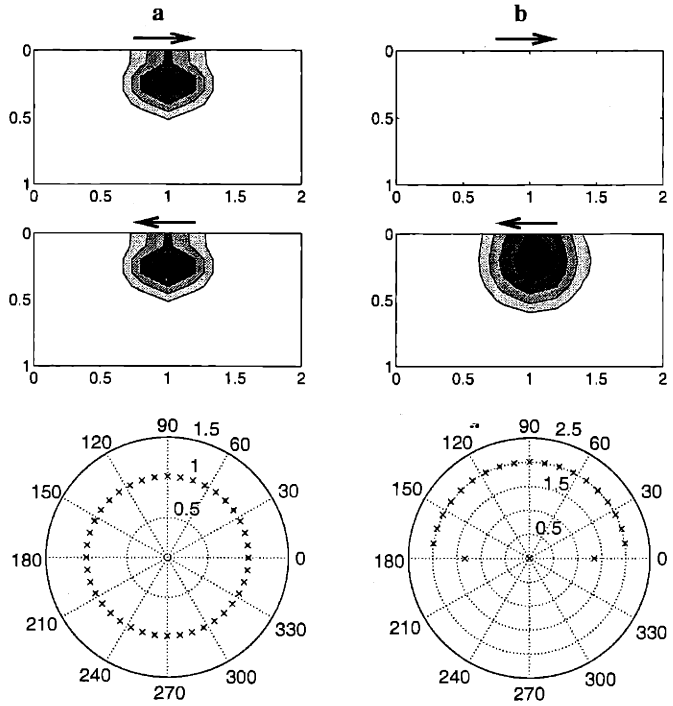


Figure 3.6: The task dependence of directionality. The contour plots show the place field of a CA3 cell which prefers the left direction, when the rat faces in the direction indicated by the arrows, and the polar plots show the maximum firing rate (indicated by the crosses, and relative to the maximum rate when averaged across directions) of the same neuron as a function of head direction; **a**: in a model rat which explored the environment randomly during the learning phase; **b**: in a model animal which always ran in one of the directions parallel to the long walls of the box during learning. The top plot is empty in **b** because the cell does not fire at all in that direction in this case. The effect of the attractor dynamics is very prominent in the all-or-none nature of activity in the directional plot in **b** (all the points in the bottom half of the plot collapsed to the origin). The maxima of the top and bottom contour plots correspond to the crosses at 270 and 90 degrees, respectively, in the polar plots.

established, one corresponding to each of the two directions sampled during learning. To illustrate this point, in the bottom half of Figure 3.6, we plotted the maximum activity of the place cell shown at the top of the figure, as a function of the animal's heading. In the rat trained using random exploration, the activity of the cell is essentially direction independent; however, if we train the rat in the shuttling task instead, the cell fires at a high rate for all directions with a westward component, and is completely silent for all directions with an eastward component. Cells which prefer direction east behave in exactly the opposite way. Once more, the results of simulations with the on-line learning procedure are similar to those obtained using the "ideal" weights, although the differences in directionality between the two training paradigms are generally somewhat reduced, and activity changes with head direction tend to be more graded. (For further discussion on how different behavioral paradigms might lead to spatial representations with different degrees of directionality, see Brunel and Trullier (1998).)

3.3.2 Very different environments

Experiments in which the firing rate maps of place cells are recorded in multiple environments which are similar to a controlled degree can provide valuable information about how input representations depend on details of the environment, how they are transformed into the place cell representation, and also about possible interference between representations of different environments realized by the same network of place cells. The general pattern of results is that radically different environments give rise to very different, and apparently unrelated place cell representations (O'Keefe and Conway, 1978; Muller and Kubie, 1987; Bostock et al., 1991). On the other hand, when a previously familiar environment is subjected to subtle alterations, the place cell representation often stays basically the same (O'Keefe and Conway, 1978; Bostock et al., 1991), or changes according to the transformation of the environment (Muller and Kubie, 1987; O'Keefe and Burgess, 1996).

In order to test our model in the first type of situation, we added another model environment to the one described in the previous section, and tested whether these two environments can be learned and recalled simultaneously without interference. The two environments are very different in terms of visual appearance; the new environment has a circular shape, and is assumed to carry visual features that are dissimilar to the ones in the rectangular box. Therefore, we assume that the spatial characteristics of both EC neuronal activities and mossy fiber inputs to CA3 as well as their relations are completely independent in the two

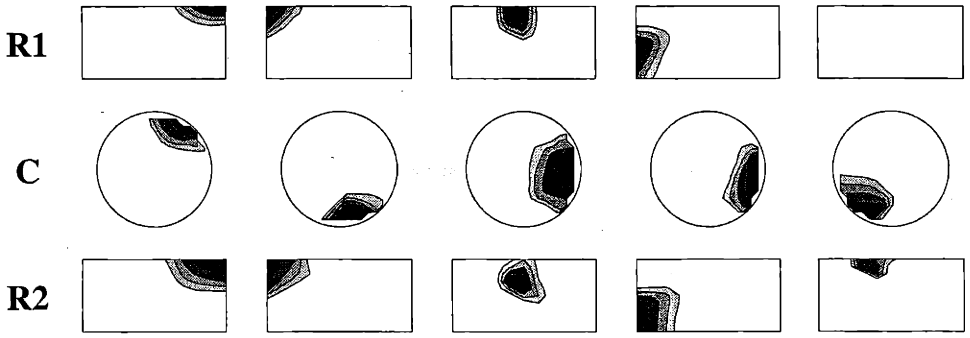


Figure 3.7: Very different environments. This figure shows the place fields of 5 selected place cells in a rectangular and a circular apparatus which have very different sensory features. The top row (marked R1) shows the place fields after learning in the rectangular apparatus but before any experience in the circular one, and the other two rows show the place fields in the circular and rectangular environments (marked C and R2, respectively) after the rat has become familiar with both. There is no obvious relation between place fields of the same cell in the two environments. The effect of encoding a second environment on the place cell representation in the first environment can be assessed by comparing the top and bottom rows of this figure. Although there are some visible changes, these tend to be small, and do not affect the general structure of the spatial representation. One of the few exceptions is shown on the far right, where a place cell which had been silent in the rectangular environment becomes active there after experience in the circular environment.

environments; i.e., for instance, knowing the relative locations of maximum activity for two EC neurons in one environment carries no information about the relation of their preferred locations in the other environment. However, as a worst case scenario, we use exactly the same neuronal populations to represent the two environments; if these populations are distinct to any extent, this can only improve the separability of the two environments. Since we are interested in interactions between different environments, and not in extending our input model to curved walls and other cues, we derive the inputs in the circular environment assuming that there is a very salient square box (which looks very different from the rectangular box) surrounding the circular arena so that the inputs are determined by distances from the walls of the square box in the same way as before.

Initial learning in the rectangular environment is performed using the on-line procedure described in the previous section, and the resulting place cell firing patterns are determined as before. Then the weights are modified by running a learning phase in the circular environment, and spatial firing distributions during recall are determined in both environments to assess interference caused by exposure to the other environment.

Figure 3.7 shows the firing rate maps of 5 model CA3 cells in the rectangular apparatus before any exposure to the circular environment (top row of figure), and in the rectangular and the circular apparatus after learning in both environments (middle and bottom rows). In general, there is no systematic relation between the location of place fields in the two different environments, which indicates that several different sets of attractors can be stored and recalled independently in the model.

Comparing the top and bottom rows of Figure 3.7 reveals that most place cells have very similar firing rate maps in the rectangular box before (R1) and after (R2) training in the circular environment. In particular, for the majority of CA3 cells, the location of maximal firing, the size, shape, and directionality (not shown) of the place field are all virtually unchanged. Consequently, the overall structure of the spatial representation is essentially unaffected by exposure to a different environment. However, for a minority of place cells, experience in the circular environment resulted in a more radical change in the firing rate map in the rectangular box (as in the last example in Figure 3.7). The most commonly observed types of change were the appearance of a new place field and the disappearance of one previously present. These probably occurred when the changes in the net input received by the cell – resulting from the weight changes that took place in the other environment – caused the neuron to cross the dynamic threshold for activation. Learning to represent a new, "orthogonal" environment can be thought of as introducing noise into both the feedforward and the recurrent weights as far as the representation of the original environment is concerned.

In order to quantify the change caused by exposure to a different environment, we computed the overlap between the overall CA3 spatial representations in the rectangular box before and after learning in the circular environment. To obtain a scale against which we can measure differences in overlap, and also to facilitate direct comparison with experimental data, we generated from our firing rate maps a large number of spike count samples, assuming independent Poisson noise for all cells and bins. Maximum firing rates, bin sizes, and session time were similar to those in experiments (e.g., Muller et al., 1987). The correlation coefficient between samples from the R1 and R2 spatial representations was found to be 0.754 ± 0.001 (mean and standard deviation), which is significantly lower (t-test, $P < 0.0001$) than the correlation between different samples from R1 (0.911 ± 0.001), but significantly higher ($P < 0.0001$) than the correlation between samples from R1 and a version of R1 where place cells have been randomly reshuffled (-0.005 ± 0.0005). These figures confirm our observation that, although there is a certain degree of degradation, the spatial

representation after learning in an "orthogonal" environment remains quite similar to the original one. Furthermore, since the number of neurons and connections is much larger in the real hippocampus than in the model, and not all neurons are active in any particular environment, interference between representations of different environments is likely to be less severe, and the number (and perhaps the spatial extent) of environments that can be stored is probably larger.

Finally, our model would also produce orthogonal place cell representations for environments that differ only in shape (Muller and Kubie, 1987), even from non-orthogonal input representations (Quirk et al., 1992), provided that the DG can separate the input patterns effectively, and the two environments are perceived as different so that learning is initiated in both environments.

3.3.3 Geometric Manipulations

We also investigated what happens to place fields in our model if the environment undergoes some simple geometric transformation. We chose to model the experiment of O'Keefe and Burgess (1996) because of its relatively complex pattern of results. In this experiment a rat, which has been thoroughly familiarized with a rectangular box, is transferred into a new box that differs from the original one only in the length of one or both sides. We will concentrate on the case when the second environment is a larger square box which can be obtained by stretching the original box by a factor of two. In this case, stretching the environment had one of the following general effects (O'Keefe and Burgess, 1996): some fields remained fixed with respect to one of the walls of the apparatus; some changed their location and/or shape in correspondence with the transformation of the box; others developed a second peak in the direction of stretching. Many of the cells with two-peaked or stretched fields also developed directional dependence; i.e., the location of maximum activity depended on the heading of the rat, usually in the way that the subfield closer to a wall was more active when the rat was facing away from that wall.

We assume that learning is triggered by exposure to the novel situation of the initial, rectangular box, and that the transformed environment in this case is similar enough to the original one so that no significant learning occurs subsequently. Therefore, the attractors established in the first environment are the final states of the network dynamics in the new environment as well, and place fields are determined by the way that the inputs (as a

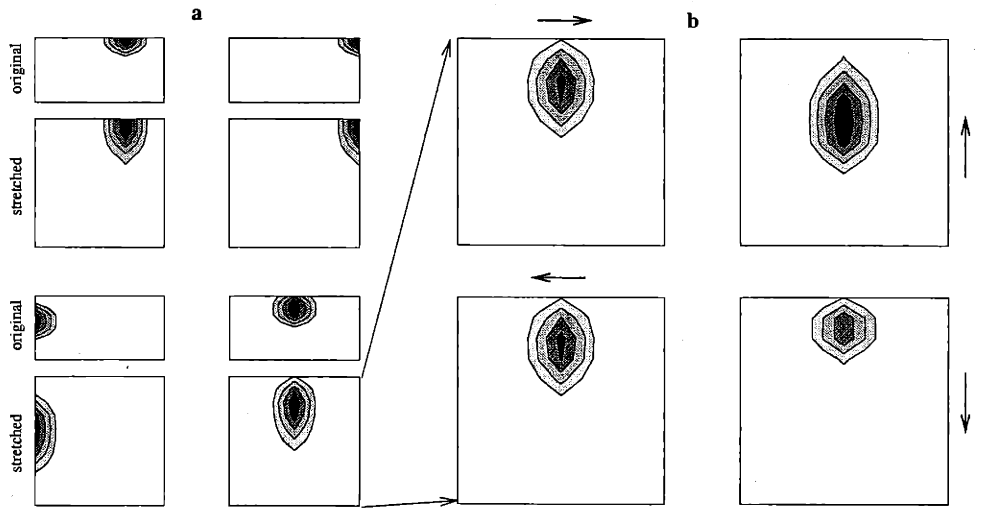


Figure 3.8: Place fields in transformed environments. **a**: The place fields of four selected cells in the original and the stretched environment in our simulation of the experiment by O'Keefe and Burgess (1996); the firing rates shown are averages over all head directions. **b**: Directionality of the place field shown in the bottom right corner of part a; the place field depends on the heading of the rat (indicated by the arrows). This dependence on head direction is induced by the transformation of the environment; place fields in the original environment are essentially non-directional (like the one shown in Figure 3.6a).

function of location and direction) in the new environment *select* attractors established in the old environment.

Figure 3.8a shows the place fields of four model CA3 neurons in the rectangular box which was used during initial learning, and in the larger square box. The place fields follow the transformation of the box; that is, their centers remain at the same relative distance from opposite walls, and their shapes become elongated along the direction of stretching. As revealed by Figure 3.8b, the fields consist of directional subcomponents with the observed relation between subfield position and preferred direction.

We can understand some of the characteristics of transformed place fields by looking at attractor selection in our model. Attractors have a regular, compact shape if place cells are characterized by their preferred locations in the original environment; on the other hand, we have no a priori knowledge about what they look like as a function of preferred locations in the new environment. Thus, it is much easier to understand the transformations occurring in the system if we look at activities in the new environment (the square box) as a function of the neurons' preferred coordinates in the old environment (the rectangle).

This is illustrated in Figure 3.9, which shows the activities of EC neurons, PP inputs to CA3, and final activities (after recurrent processing) in CA3 at three different locations in the square box, all as functions of preferred locations in the rectangular box. The activities of EC cells are determined by multiplying together (Gaussian-tuned) components whose activities depend on the animal's heading, and its position with respect to the walls. Since the walls have moved relative to each other, the different components lead to different estimates of position in the old coordinate system. Combining such inputs conjunctively leads to an EC activity profile which peaks somewhere between the positions indicated by individual walls. For instance, when the rat is halfway between the two walls that have been moved apart, listening to one of these walls would indicate that the animal is located at the opposite wall, and the resulting EC activity profile is centered on neurons which like the middle of the rectangular box (see the bottom left contour plot in Figure 3.9). Since the PP connections were established in the rectangular box, the PP input pattern to CA3 cells is centered around the same location as the EC activity pattern if both are viewed as a function of preferred coordinates in the rectangle (compare the first and the second columns of Figure 3.9). The recurrent connections then sharpen the activity profile considerably, but leave the location of the bump (in the old coordinate system) essentially unchanged. The final activities of CA3 cells as a function of location in the square box define the place fields in the new environment. We can see that as the rat moves around in the new environment, the activity packet also moves smoothly on the plane defined by the preferred locations of place cells in the rectangular box. This results in a smooth transformation of place fields between the two environments. In addition, the activity packet moves more slowly in the stretched direction in the old coordinate system than the actual speed of the rat in the new environment, or, in other words, the rat needs to travel about twice as much in the square box than in the rectangular box for the activity profile to shift by the same amount; consequently, place fields become elongated in the direction of stretching.

The emergence of directional subcomponents can be understood by looking at how the activities of EC cells and the resulting activities of CA3 neurons depend on the head direction of the rat. This is depicted in Figure 3.10, which shows that due to the dependence on head direction of the rat's confidence in the inputs from different walls (as described earlier), conflicting sources of information are weighed differently depending on which way the rat faces. The EC activity profile and, consequently, the CA3 activity profile, shift as the rat turns around in the square box, and the result is that a given place cell fires maximally at different locations depending on head direction. In the example shown in Figure 3.10, the activity profile shifts 'north' when the rat faces 'north', and shifts 'south' when the

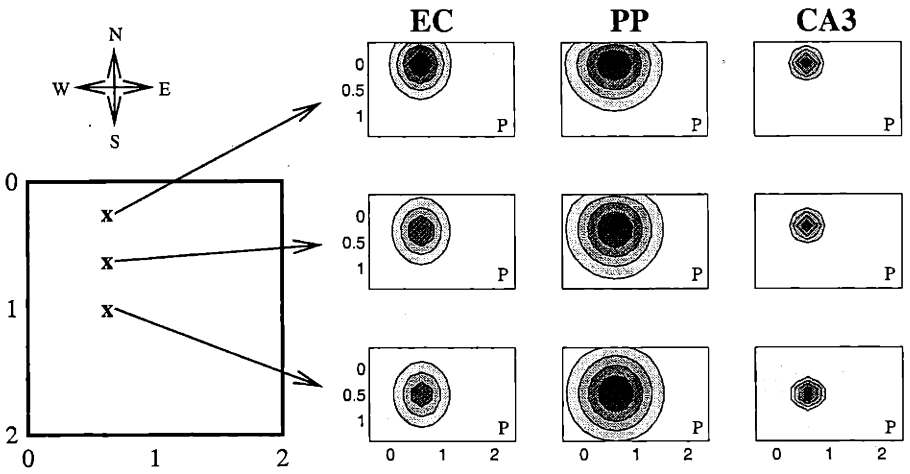


Figure 3.9: Place field stretching. The population plots of this figure show the neuronal activities in entorhinal cortex (EC), the perforant path (PP) inputs to place cells, and the CA3 final activities as a function of the preferred location of the neuron in the original, rectangular box, for three different positions of the rat in the square box, indicated by the crosses in the plot on the left. The plots only show cells with preferred direction 'north', and the model rat faces 'west' in all cases.

rat faces 'south'. In an apparent paradox, from the perspective of a single place field, this actually has the opposite effect (shown in Figure 3.8b), that the center of the place field is further 'south' when the animal faces north, and vice versa. The easiest way to see this is to ask where in the environment the rat has to be when it is facing in a particular direction, to arrange for exactly the population activity across CA3 shown in the middle picture of Figure 3.10. The answer to this will tell us how the favored location of the most active cell in this population depends on direction. When the rat faces 'north', the activity profile shifts 'north', so the rat must be displaced relatively 'south' in order to compensate for this. Thus, the location for the cell's peak response is shifted 'south'. The converse is also true – when the rat actually faces 'south', then the place field moves 'north'.

3.3.4 Very similar environments

Skaggs and McNaughton (1998) conducted an experiment designed specifically to probe the relation between spatial representations in environments with a high degree of similarity. In this experiment, animals explored an apparatus which consisted of two visually identical boxes connected by a corridor. Many place cells were found to have similar place fields in the two regions, whereas others had uncorrelated place fields. This finding challenges

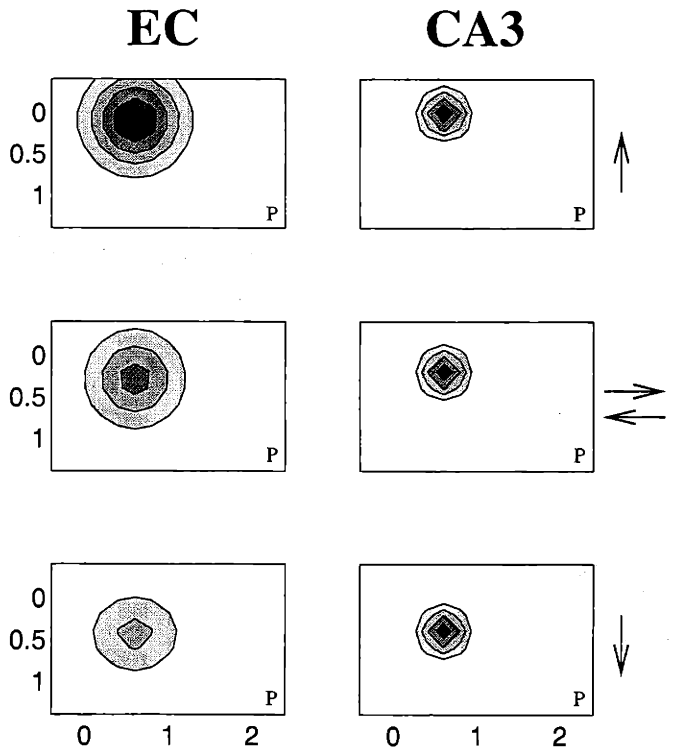


Figure 3.10: Directionality of stretched place fields. Population plots of EC neuronal activities and CA3 final activities (of the same sets of cells as in Figure 3.9), as a function of the neurons' preferred locations in the rectangular box, for different headings of the rat (indicated by the arrows) at a single location in the square box (marked by the middle cross in Figure 3.9). The middle row of plots is for both directions 'east' and 'west' as these lead to the same activities for the neurons displayed here. The position of the input peak changes as the rat faces in different directions (due to the dependence on head direction of the breadths of the input components tied to different walls), and the position of the final activity profile in CA3 changes accordingly. This shift can be compensated for by changes in location as seen in Figure 3.9, resulting in the directional subfields shown in Figure 3.8b.

the idea that there is a predefined set of uncorrelated attractors wired into the recurrent connections in CA3 (Samsonovich and McNaughton, 1997), because such a model would predict either identical or orthogonal firing patterns in different environments or different parts of the same environment. This particular problem may be solved by postulating a hierarchy of fixed attractors with various degrees of overlap (Samsonovich et al., 1998); however, it still remains to be explained why similar representations are selected in very similar environments. On the other hand, the attractors established in our model are input dependent, which in principle allows attractors with an arbitrary degree of similarity, and directly defines the association between attractors and environments. Therefore, we simulated the experiment by Skaggs and McNaughton (1998) in our model to study the spatial representations in very similar environments.

We still do not model the different sources of spatial information explicitly. We assume that there are some inputs (e.g., signals derived from path integration) which allow the two boxes to be distinguished, while other inputs to the system (e.g., local visual cues) are identical at corresponding locations in the two boxes. Since cells in EC are assumed to respond to different inputs to a randomly varying extent and to encode these inputs conjunctively, we applied the following scheme to determine activities in EC at locations inside the two boxes. EC cells are now characterized by a preferred location (and also a preferred head direction) based on visual inputs (this is now actually a set of two locations, one in each box), as well as a polarization index (P), which is defined as the maximum firing rate for the cell in the north box minus the maximum firing rate in the south box, divided by the maximum rate in any of the boxes. P is always between -1 and 1, its magnitude indicates how much that particular cell is influenced by cues that distinguish the two boxes, and its sign shows which box the neuron prefers. We assign P values to EC cells randomly from a uniform distribution. The firing rate of an EC neuron is then given by $z_k = (1 + P_k)z_k'$ in the north box and $z_k = (1 - P_k)z_k'$ in the south box, where z_k' is a function of coordinates within the current box, and it depends on spatial position and head direction the same way as z_k in Equation 3.2. We assume that the MF inputs to CA3 can be characterized similarly; however, due to the orthogonalizing properties of the dentate gyrus, P values do not vary continuously, but only take the values -1, 0, and 1, each with probability 1/3. This means that there is a population of cells in CA3 which receives the same input at corresponding locations in the two boxes during learning, while another population receives different inputs. Since the first time the rat is introduced into the apparatus it is allowed to explore it entirely, we do not treat the two halves of the environment differently during the learning phase.

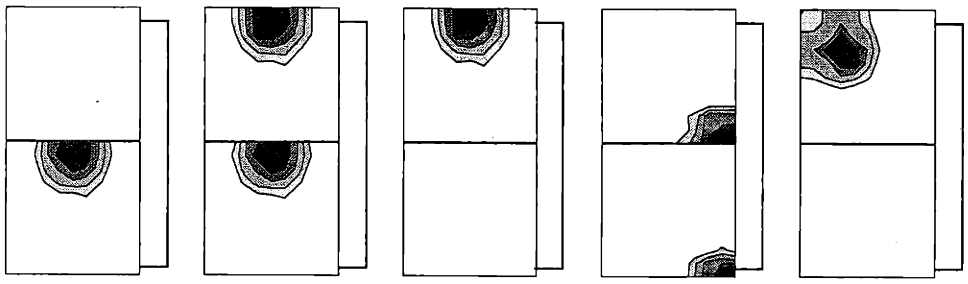


Figure 3.11: Place fields in our simulation of Skaggs and McNaughton (1998). The figure shows the place fields of five CA3 place cells in the two identical boxes; activity in the corridor connecting the boxes was not simulated.

Some examples of the place fields that develop in this model are shown in Figure 3.11. There are cells which have similar firing rate patterns in the two boxes, while others are active in only one of the boxes, in accordance with experimental observations. In other words, our model has no difficulty storing and recalling partially overlapping spatial representations. In the model, the degree of overlap is determined by the extent of orthogonalization occurring in DG, i.e., what proportion of granule cells distinguishes between the two boxes – CA3 cells simply inherit the selectivity of their MF inputs as attractors are established during the learning phase. Most EC neurons are active in both boxes, although to a different extent (see Figure 3.12a). Consequently, all CA3 cells that are active in this environment get a substantial PP input in both boxes (Figure 3.12b); however, the activity patterns encoded during learning are restored by the recurrent connections and feedback inhibition, and the PP input only determines which of these patterns emerges. The figures also show that although EC neurons have relatively broad tuning curves, and this results in CA3 cells receiving feedforward input that is even more broadly tuned, the final tuning of CA3 neurons is considerably sharper due to recurrent activity. The attractor network also renders place cells directionally nonselective, just as before.

3.4 Discussion

3.4.1 Principal findings

We have presented a plastic attractor network model of CA3 place cells which describes how a conjunctive representation of location- and direction-specific sensory information in

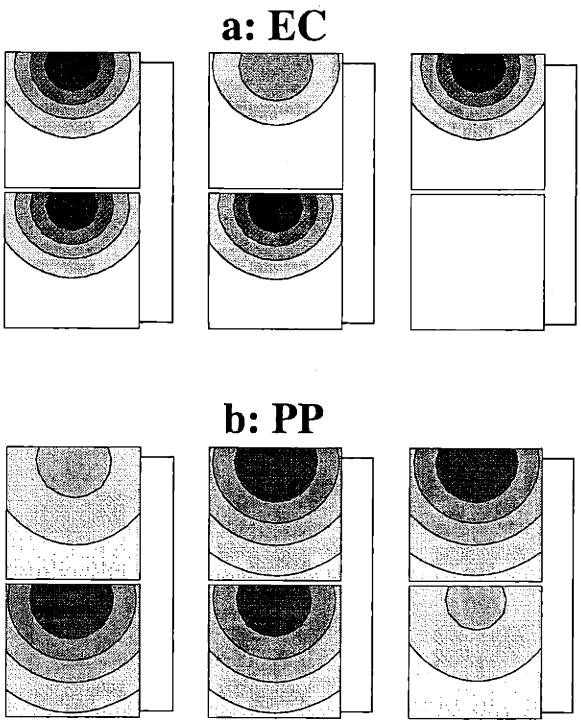


Figure 3.12: Input representation and inputs to CA3 in our simulation of Skaggs and McNaughton (1998). **a**: The activities, as a function of location in the apparatus, of 3 entorhinal neurons which have the same preferred (visual) location within the boxes, but different degrees of polarization (as defined in the text; the polarization indices are -0.01, 0.25, and -0.70, respectively). **b**: This part of the figure, which displays the perforant path inputs to the first three CA3 place cells of Figure 3.11, shows that, as a result of learning in the perforant pathway, some place cells receive similar inputs at corresponding locations in the two boxes, while others receive inputs of different magnitudes, setting the stage for the CA3 recurrent network which makes these differences much more pronounced (as seen in Figure 3.11).

entorhinal cortex can be transformed by feedforward pathways and recurrent processing in the hippocampus, into a place cell representation whose properties match a wide range of experimental observations. In particular, our model (1) accounts for the head direction independence of place cells in open environments as well as their directionality in linearly restricted environments, (2) demonstrates how several different environments can be stored and recalled independently by the CA3 recurrent network, (3) produces place cell activity patterns with an appropriate degree of overlap in visually similar environments, and (4) correctly captures the transformations of place fields after simple geometric manipulations of the environment.

Further, we have shown that the neural connections required for this spatial representation can be computed from the correlations between the input-driven feedforward activations of neurons during initial exploration of the environment, using a familiarity-based on-line learning procedure.

Although the representations formed may be useful for spatial tasks such as navigation (Burgess et al., 1997; Foster, 1999; Foster et al., 2000), a major goal for our model was to show how ideas about how non-spatial information is processed by the hippocampus are in accordance with data on place fields.

3.4.2 Components of the model

The idea of using attractor networks for computations has been applied in various settings (Somers et al., 1995; Zhang, 1996; Pouget et al., 1998); such networks have been shown to be capable of amplifying certain facets of their inputs (Ben-Yishai et al., 1995) as well as creating invariance (Chance et al., 1999). Our model (and Brunel and Trullier's, 1998) displays both behaviors simultaneously; the recurrent network enhances the spatial tuning of place cells while suppressing their directional tuning in open field environments. Under attractor dynamics (which we assume characterizes well the average behavior of CA3 across theta oscillations), it is unwise to invent rules describing how individual place cells respond in various situations; rather, the system is better described collectively, by identifying the attractors and specifying which attractor gets selected for any particular input. The attractor concept also helps explain the persistence of spatial firing patterns in the face of environmental manipulations such as cue removal or cue rotation (O'Keefe and Conway, 1978; Muller and Kubie, 1987) as well as the abrupt changes that ensue for changes of other kinds (e.g., changing the shape of the environment from circular to square; Muller

and Kubie, 1987) or of a larger magnitude. Feedforward models (e.g., Sharp, 1991; Burgess et al., 1997; Hartley et al., 2000), albeit ignoring the recurrent connections, can also be made to exhibit many of the properties we have demonstrated. We have not yet modeled the pathway from CA3 to CA1, assuming that the spatial properties of the latter faithfully reflect those of the former, assuming normal plasticity. CA1 is, of course, the source of the bulk of the experimental data on place fields.

Although we have assumed the CA3 recurrent network as the anatomical substrate for the learned continuous attractors in our model, alternative anatomical localizations of essentially the same computations are certainly possible. In particular, it has been found recently that the dentate gyrus also possesses an extensive recurrent network (Buckmaster and Schwartzkroin, 1994), in which dentate granule cells excite so-called mossy cells in the part of the hippocampus known as the hilus (adjacent to, and sometimes considered to be part of the DG). Hilar mossy cells, in turn, project extensively back onto granule cells via connections that can undergo long-term potentiation (Hetherington et al., 1994), thus forming a plastic recurrent network in the dentate region. This network might support the learning of (discrete or continuous) attractors, and may carry out at least some of the operations attributed to CA3 in our model, in addition to the role of the DG in pattern separation and orthogonalization.

More generally, the existence of two plastic recurrent networks at two subsequent stages of hippocampal processing, albeit with substantially different properties, suggests some kind of division of labor between the two networks, but this division may be different from what we proposed in our model. Two recent computational models provide examples of this by proposing that the DG-hilus system is responsible for most pattern completion in the hippocampus and the role of CA3 is to perform sequence prediction (Lisman, 1999), or, alternatively, that the DG-hilus system is responsible for context selection for the hippocampal place cell representation (Doboli et al., 2000).

The learning rule was chosen as a crude model of experimental long-term synaptic plasticity (LTP), and we have ignored most empirical complexities. We have not taken into account the fact that the sign and magnitude of long-term synaptic modification depends on the relative timing of pre- and postsynaptic activity (Levy and Steward, 1983; Markram et al., 1997) , which has been suggested as a mechanism underlying a navigational role of place cells (Blum and Abbott, 1996). Indeed, the recurrent weights in our model ultimately learn a weight structure similar to the "cognitive graph" described by Muller et al. (1991b, 1996b).

Similar proposals to ours have been put forward in associative memory models of the hippocampus (Treves and Rolls, 1992) as to the separate roles for the indirect pathway to CA3 via the dentate gyrus (which defines attractors during the learning mode) and the direct perforant path (which selects attractors during recall mode). However, the activity patterns representing location and direction are intrinsically continuous, and thus strongly overlapping, so the patterns that are retrieved can differ in systematic ways from all the patterns encountered during learning (*eg* being insensitive to head direction in open field environments). The relationship between attractor networks storing discrete vs. continuous sets of patterns, particularly regarding their storage capacity, has been studied by Samsonovich (1997) and Battaglia and Treves (1998).

Maintaining such overlapping attractors requires a learning rule which compensates for systematic spatial biases during exploration by gating learning through familiarity in a *graded* rather than a *binary* fashion, a subtlety not necessary for the very distinct attractors assumed by memory models. Gating of synaptic effectiveness (among the MF, PP, and recurrent collateral connections) and plasticity may either be mediated by closely coupled or by more distinct mechanisms (e.g., acetylcholine vs. GABA Hasselmo, 1999; Sohal and Hasselmo, 1998), but there is very little evidence to distinguish between these possibilities at this point.

Exactly how entorhinal and dentate neurons encode features in the environment, and how they respond to manipulations of the environment, is not experimentally clear. Our choice was necessarily somewhat arbitrary – the aim has been to show that there exists at least one reasonable choice that results in place fields consistent with experimental data in a wide range of experimental situations. Our entorhinal representation is similar to that of Burgess et al. (1997), except that their units are directionally non-selective, and each is tied to exactly two orthogonal walls of the environment. In their model, place cell firing patterns are then determined through the feedforward weights connecting EC to CA3; these weights are set up using a competitive learning scheme similar to the one used by Sharp (1991) to model the formation of place fields. Competitive learning supports the separation of different input patterns in these models; in our model, the same task is thought to be accomplished through processing by the dentate gyrus. It is possible, of course, that both of these processes contribute to pattern separation in the hippocampus.

3.4.3 Comparison with other models

Apart from (Brunel and Trullier, 1998), the models of Samsonovich and McNaughton (1997) and Burgess et al. (1997) are closest to ours. The most important distinction from Samsonovich and McNaughton (1997) is that it relates the position of the activity profile (or 'packet') in CA3 (as an attractor network) to external coordinates in a different way, assuming a hard-wired system which is capable of updating the position of the CA3 activity packet based on self-motion information, and a learned association with sensory representations which can be used to correct for accumulated errors in path integration. In other words, the CA3 network is an integral part of the path integration system in their model. In fact, experimental evidence on the direct involvement of the hippocampus itself in path integration is controversial (Alyan and McNaughton, 1999; Maaswinkel et al., 1999).

Learning works differently in our model, and the metric of the place cell representation reflects the way in which the EC representation depends on external coordinates, including sensory features of the environment and, to account for the formation and maintenance of place fields in darkness, self-motion information. One important consequence of this is that many of the problems associated with multiple pre-defined continuous attractors (also known as "charts") are solved in a trivial manner. In particular, there is no need to consider what happens when the animal reaches the edge of the chart, or how to stitch charts representing different parts of the environment together; the continuous attractor is extended by learning as needed, and consistency is ensured by the similarity structure of the input representation. The phenomena of multiple place cell representations associated with the same external external input depending on task or past experience (which formed the basis of multi-chart theories; Samsonovich and McNaughton, 1997; Redish, 1999) can be explained in our model by a combination of learning effects and the persistence of active neural representations (both within the CA3 attractor network and due to feedback from the hippocampus to EC).

Burgess et al.'s (1997) model also accounts for some of O'Keefe and Burgess's (1996) data. Their results are complementary to the ones we presented here, in that their model captures the behavior of those place cells which remain fixed with respect to one wall or develop a second place field after stretching the environment, while our model correctly describes those place fields that follow the transformation of the environment, and also explains the acquired directionality of stretched place fields in the transformed environment. A modified version of our model, which incorporates random variations in the extent to which

input cells respond to different spatial cues, reproduces all the observed classes of place field transformation. Due to its randomness, it offers less insight into the underlying mechanisms than the model described here. Our model also accounts for other properties of place cells, such as directionality and non-directionality.

3.4.4 Extensions of the model

Our model of spatial representations in the rodent can be extended in at least three different ways: first, by testing whether the model can capture the effects on place cells of further experimental manipulations; second, by including in the model some more detailed features of hippocampal processing in order to expand its scope, particularly in the temporal domain; and third, by building explicit models of areas other than CA3, starting from the dentate gyrus and CA1, but potentially including other areas such as the subiculum and entorhinal cortex. Since these last two kinds of extensions are also important for models of the hippocampus outside the spatial domain, they will be discussed in Chapter 5.

For testing the model in novel experimental situations, a particularly interesting set of data is from experiments which look at long-term changes (on the time scale of hours and days) in the place cell representation resulting from various manipulations (e.g., Bostock et al., 1991; Knierim et al., 1995; Jeffery, 1998; Jeffery and O'Keefe, 1999). These experiments clearly indicate that, at least in situations which involve uncertainty or conflicting information from sources, place cell firing patterns continue to change over a long time period. These changes are likely to require plastic changes in feedforward and recurrent connections in the hippocampus. It would be important to find out whether these observations can be accommodated by our current model, either by assuming some residual plasticity after initial exploration in an environment, or by identifying the conditions which might lead to a relapse into learning mode at later stages.

3.4.5 Critical experiments

Various experiments could, in principle, test the key assumptions and predictions of our model. First, pharmacological or molecular biological blockade of plasticity in the CA3 recurrent connections should prevent the formation of a new representation in a novel situation. According to our model, the system would either remain trapped in learning mode, which would be indicated by, among other things, retained directionality of place fields in an

open field, or recall attractors from one or more environments explored before the blockade, resulting in irregular or fragmented place fields. Direct manipulations of the neuromodulatory control mechanisms governing the choice of learning versus recall mode should have a similar effect. Unfortunately, there exist many different forms of experimental plasticity, and it is not clear which in particular are most relevant for learning in vivo.

Our model predicts that the CA3 place cell representation should be different during the first few minutes of exploration in a new environment from the time after the animal has become familiar with its surroundings. In particular, place cells are expected to be directional in any novel environment immediately after entry, and become non-directional later in open environments.

Analysis of our model also indicates that the amount of training in a given environment might have a significant effect on the place cell representation in a similar environment encountered subsequently, since only well-established attractors are assumed to be capable of being recalled. For instance, we would expect to see a less obvious relation between place fields in different environments in the experiment of O'Keefe and Burgess (1996) if, instead of training the rat in one size of box before allowing it to explore the others, they had made it explore all four environments in quick succession, especially if the rat is prevented from using extramaze cues.

Chapter 4

A model of hippocampal-cortical interactions and memory consolidation

4.1 Introduction

4.1.1 Basic phenomenology and general observations

As we have seen in the Background chapter, the hippocampus appears to be involved in a wide variety of behaviors, and substantial effort has gone into identifying common themes and organizing principles in the data. One such theme, the focus of this chapter, is the putative time-limited role of the hippocampus in certain types of memory. Evidence for such a role comes mainly from studies of retrograde amnesia (RA). RA can be defined as impaired performance (compared to untreated controls) of subjects that were affected by some kind of manipulation (having a well-defined time of onset; usually focal brain damage) on a task that was acquired before the manipulation started. RA following hippocampal damage is often found for tasks whose acquisition depends on the hippocampus (i.e., those that are also affected by anterograde amnesia), and sometimes also for tasks unaffected by anterograde amnesia (for a review, see Sutherland et al., 2001). In many cases, retrograde amnesia has been shown to be temporally graded; i.e., remote memories are recalled better than recent ones, the exact opposite of the pattern found in normals. This phenomenon

has been known for more than a century (for historical references, see Alvarez and Squire, 1994). An important insight into the underlying mechanisms was afforded by the discovery that lesions to the medial temporal lobes (MTL) in humans can give rise to temporally graded RA (Scoville and Milner, 1957). This finding has now been confirmed by studies of a large number of patients, as well as by controlled lesion studies in several animal species.

It should be noted, however, that retrograde amnesia does not always seem to be temporally graded. As discussed earlier, whether RA is graded (as well as the temporal extent of the gradient when it is graded) depends on the anatomical location and extent of the lesion, as well as on the nature of the task. In particular, ungraded RA for all types of declarative memories has been associated with damage outside the medial temporal lobes (e.g., in lateral temporal cortex) in humans. It has also been suggested that, even with lesions limited to the hippocampus, retrograde amnesia for certain types of material (in particular, personally experienced episodes, and perhaps detailed spatial information as well) may always be temporally ungraded (Nadel et al., 2000). In animals, the amnesic effects of lesions outside the MTL (and associated subcortical areas) have not been explored systematically. However, ungraded RA (sometimes also called retrograde amnesia with a "flat gradient") has been reported for some tasks after MTL lesions of varying specificity (Murray and Bussey, 2001; Sutherland et al., 2001).

Nevertheless, the evidence is quite strong that temporally graded RA does occur for at least some types of material following relatively selective lesions to medial temporal lobe structures (including the hippocampus) in several species including humans. Accepting this fact, we can now ask the question: What are the implications of temporally graded RA? It was suggested a long time ago that temporally graded RA reflects a gradual reorganization of the neural substrate of memory, and this notion then became linked with the medial temporal lobe through the data described above. Temporally graded RA after hippocampal (or other MTL) lesions has been taken to imply that memories which initially depend on the hippocampus for their successful retrieval are gradually reorganized (or *consolidated*) through some hippocampally-dependent process so that retrieval becomes possible without the hippocampus.

Several points are worth emphasizing here. First of all, based only on the numerical dependence of performance levels (in amnesics and controls) on time between training and testing (with lesions occurring shortly before testing in amnesics), very little can be said about the nature of the memory representation at different times after acquisition. Even in normals, memories before and after consolidation might have very different characteristics as long as

they can both support task performance. An example of such a qualitative change might be the observed "semanticization" of episodic memories in humans (Nadel et al., 2000). Similar examples of apparent qualitative differences have been found in animal studies (Nadel and Bohbot, 2001).

Second, there are some points concerning the relationship between the observed shapes of the time-performance curves (as described above) and the underlying deficit. It has been noted repeatedly that only the case in which performance is significantly worse for recently acquired memories than for more remote ones (at least from some time period) in lesioned subjects can be interpreted as unequivocal evidence for hippocampal involvement in memory consolidation (Squire, 1992; Alvarez and Squire, 1994; Redish, 1999). Otherwise, an alternative explanation, according to which memory has both a quickly fading, hippocampally dependent and a more permanent, neocortical component, cannot be ruled out. As briefly mentioned before, the converse is not true; i.e., even if both normals and hippocampals show better performance for more recently acquired memories, it is possible that hippocampally-dependent consolidation is occurring, but its effect gets masked by some hippocampally-independent short-term component (provided that hippocampals are significantly worse than normals at recent memories, and this difference vanishes for remote time periods). By the same token, it is not required even for a strict test of the consolidation hypothesis that performance for recently encountered material should be worse than performance for material learned a very long time ago if both these putative short-term effects and normal forgetting are taken into account.

Finally, the hippocampus appears to play at least two separable roles in memory. Since the memory is lost if the hippocampus is lesioned directly after successful acquisition, the hippocampus must have a role in either the storage or the retrieval of the memory at this stage. On the other hand, the consolidation process itself must depend on the hippocampus, since no recovery from retrograde amnesia can normally be observed in the absence of the hippocampus. A detailed analysis of the possible ways in which the hippocampus may contribute to each of these operations is presented next, in the context of a more general examination of the computations that support memory.

4.1.2 Computational analysis of the role of the hippocampus in declarative memory

General characterization of memory systems

Before going into the specific roles of the hippocampus in declarative memory, first let us step back and consider memory storage from a computational point of view (in the sense of Marr (1982)). At this level, memory may be defined as using information from past experiences to aid subsequent performance, and the argument is about optimality, independent of how the required computations are performed. One of the advantages of this level of analysis is that it can make very clear the ways in which memory tasks can differ. One of the important variables is the structure of information in the learning episodes; in particular, whether there is a single or multiple learning episodes, what the nature of the stimuli constituting a single episode is, and how the experiences in the different learning episodes relate to each other. The other fundamental factor is the information required to solve the task at the time of retrieval, and how this relates to the information in the set of learning experiences. This relationship determines the amount and type of information that needs to be preserved about the original experiences. Of course, we generally expect that information from a set of experiences will need to be used in different ways in various tasks, and therefore the information stored during initial exposure must be able to support all these possible future uses. For example, consider the case when there are multiple learning experiences, each of them consisting of the concurrent presentation of a pair of visual stimuli. Then the information that needs to be preserved about these presentations is very different depending on whether the task later will be recognition of the pairs as a whole, recall of paired associates, generalization to new instances of the rule describing the relationship between the stimuli within a pair, or being able to say whether at least one of the stimuli was always a household item.

The argument so far has not depended crucially on the way experiences, recalled memories, and other information are represented, or the form in which memories are stored. However, if we want to extend our analysis to what Marr (1982) referred to as the representational and implementational levels, in order to find out how the brain carries out the computations that support memory, we need to take into account the different ways in which information can be represented in the brain. In particular, suppose we want to find the optimal form for the storage of information in a given memory task. This will clearly depend on a number of factors and their associated costs. First, the form in which the original learning episodes

are represented is relevant, as are the representations utilized by the output systems, since information needs to be converted from the input format to the stored form during exposure to the learned material, and from the stored to the correct output format during retrieval. Depending on the storage format (assuming that the input and output formats are given), one or both of these conversion operations may be computationally expensive (requiring a lot of time or resources). For example, if the task requires the integration of information across different learning episodes, use of a stored format which itself integrates different traces might be appropriate since this could make retrieval much simpler. However, this might come at the cost of having to take into account all previous experiences when storing (the relevant aspects of) a new one. Alternatively, one might consider keeping the stored representations of the different experiences separate, and performing the integration at the time of retrieval. On the retrieval side, one of the most important contributions to the cost is probably the time required to generate an answer.

Second, an obvious contribution to the cost of solving a memory task in a particular way comes from the cost of making errors. A third, somewhat less obvious contribution is the cost of maintaining the stored information between learning and retrieval. The inclusion of this term provides a simplistic explanation for why short-term memory is mostly activity-based, while long-term memory tends to be weight-based. The idea is that storage based on maintained activities is likely to involve lower conversion costs between the stored representation and input and output formats due to their fundamentally similar nature, but weight-based storage prevails for longer retention intervals because of the lower metabolic costs associated with maintenance. However, other factors like the vulnerability of the stored memory are also likely to play a role. Finally, as already described at the computational level, an efficient memory system should be able to utilize information from a given learning experience in different circumstances, and the form of the stored representation is at the core of such versatility. Of course, it is possible that a single stored representation cannot support all the relevant uses of the information in a learning experience, and this might explain some observed instances of different learning systems apparently processing the same type of information (albeit in different ways).

An example of such a possible conflict between different utilizations of the same experiences (one that is very much relevant to the main topic of our discussion) is provided by the analysis of McClelland et al. (1995). These authors suggest that there may be a computational reason for a division of labor between the hippocampus and neocortex. They argue that discovery of shared structure in an ensemble of inputs (which is thought to be necessary

for successful learning in most complex tasks) crucially depends on making slow, gradual changes to the internal parameters of the learning system based on multiple, interleaved presentations of representative samples from all areas of the input space. In a system designed to extract general structure, attempts to learn new, specific information without interleaving it with examples that conform to the structure already learned result in the phenomenon called "catastrophic interference", a dramatic impairment for previously acquired information following even moderate amounts of new learning. Thus the neocortex, which is assumed to be responsible for the acquisition of general information, is unsuited for rapid learning of specific events, and needs to be supplemented by the hippocampus that performs this latter task (but is not capable of extracting generalities).

Next, we will turn to the more specific issues of how the hippocampus might be involved in the early recall and the consolidation of declarative memories. However, we cannot address these issues without assuming something about the nature of input representations and storage mechanisms. Our earlier arguments about RA already assumed that both representation and storage are local to specific brain areas. However, if we are to evaluate the hypotheses regarding the specific roles of the hippocampus in memory processes, some additional assumptions (most of which are widely accepted in the neuroscience community) have to be made. First, we assume that perceptual and other inputs, cues for memory retrieval, and the retrieved memories themselves are all represented as specific patterns of activity across large numbers of neurons, often distributed in several different brain areas. Experiences of events activate neurons in a large number of brain areas that process stimuli of different modalities, and recall of the episode leads to the same (or very similar) pattern of activation in those same areas (perhaps excepting "low-level" sensory areas). Second, it is assumed that memories are stored in the efficacies of the synaptic connections through which one neuron can influence the activity of the other; this may include connections within a brain area as well as connections between different areas. Third, learning involves changes in these synaptic efficacies which depend mainly on the activities of the particular neurons making the connection. Finally, there exist global factors, such as inputs from subcortical neuromodulatory centers, that can influence both neuronal activity and synaptic plasticity in a manner that is not specific to particular patterns of activation.

Hippocampal involvement in early recall

Let us first consider the role of the hippocampus in early recall. In other words, how does the hippocampus contribute to the recall of a memory directly after it has been established?

There appear to be two (not mutually exclusive) basic possibilities. One possibility is that some of the information required to identify the memory to be recalled uniquely is actually stored in connections involving the hippocampus. This storage may involve intrahippocampal connections as well as the connections to and from the hippocampus, whose information content is also lost if the hippocampus is damaged. The other possibility is that although all the information about the memory may be stored outside the hippocampus, the hippocampus is somehow required to access this information and translate it into the activity pattern representing the memory. In principle, there would also be a third possibility, namely, that the hippocampus, rather than playing a role in memory, is instead critical for other aspects of the task (such as perceptual processing or the performance of necessary actions); however, this possibility can be ruled out quite easily in most cases.

The simplest version of the storage deficit hypothesis is that the hippocampus stores the memory in its entirety, i.e., a representation of all the information that can later be recalled. This is in fact the position taken in several theories. For example, Marr (1971) suggested that the hippocampus could act as a "simple memory", a temporary store for all incoming information, that would assist the neocortex in the categorization of stimuli by replaying these experiences later (during sleep). In this theory, connections to, from, and within the hippocampus act as storage sites for memorized patterns. As pointed out by McClelland et al. (1995), the stored representation would not have to be a copy of neocortical activation patterns, but rather a compressed, summary version of them, which may nonetheless carry the same information due to the redundancy of the neocortical representation. Compression and decompression would be carried out by the convergent neocortical projections to the medial temporal lobe and on to the hippocampus, and the divergent back-projections from the hippocampus through the medial temporal cortical regions to other areas of neocortex, respectively.

This version of the proposal is actually closely related to "indexing theories" of hippocampal function (e.g., Teyler and DiScenna, 1986), which hold that the role of the hippocampus is to store an index, or list of pointers to the neocortical locations where the different components of the memory itself are stored. Closer scrutiny reveals that the "compressed storage" and "indexing" proposals are in fact completely equivalent in relation to the role they assign to the hippocampus during episodic encoding and retrieval. In particular, according to both Teyler and DiScenna (1986) and McClelland et al. (1995), very little change in neocortical connections occurs during initial learning; instead, synaptic changes within the hippocampus mediate initial storage, and the original neocortical representation of the

memory can be reinstated via the existing mapping between hippocampal and neocortical representations, embodied by the hippocampo-cortical backprojections. Another alternative is that initial learning occurs in the hippocampus and in the backprojections to neocortex, which could maintain the correspondence of hippocampal and neocortical representations. During retrieval, cues activate the index or compressed representation of the memory in the hippocampus, which in turn activates the corresponding neocortical representation.

Another idea mentioned above, namely that the hippocampus binds together information from disparate cortical areas, is also central to several theories of hippocampal function. As described earlier in the section on anatomy, the hippocampus (and, in fact, entorhinal cortex and, to some extent, the perirhinal and parahippocampal cortices as well) can be seen as a convergence point in the hierarchy of neocortical areas, which, taking into account both forward and backward projections, provides a route of communication between any areas within the hierarchy. In fact, this anatomical binding property is a simple way in which the medial temporal lobe may be critical for memory retrieval independent of the plastic changes happening within the structure – if the areas representing the memory (some of which may also represent retrieval cues) cannot communicate with each other, retrieval will fail, even if all the information stored about the memory is intact. In many theories, this binding role of the MTL is combined with storage of information within the MTL.

On the other hand, there have also been several proposals according to which the hippocampus does not store memories at all. Instead, it may serve as a modulatory center; its role may be to enable "chunking" (the formation of new associations) in neocortex (Wickelgren, 1979), or to "imprint" or "rehearse" memories in neocortex (Mishkin, 1982). Alternatively, it was proposed that the hippocampus may implement the "orienting subsystem" in "adaptive resonance theory", processing information about novelty of stimuli, and signalling to neocortex the need to form new representations (Carpenter and Grossberg, 1993).

A final issue I will consider regarding initial storage of hippocampally dependent memories is the locus and nature of the underlying neural changes. Since at least some of these memories (including episodic memories is humans) can be acquired in a single presentation, there must be large enough changes occurring in some connections as a result of a single episode to support reliable recall of the specifics of that episode. In the storage theories discussed above, at least some of these "fast" changes happen in the hippocampus. In fact, several authors have proposed that this ability to rapidly store information about specific occurrences is a unique property of the hippocampus; in contrast, the neocortex may be specialized in "slow" learning of general information which can only be acquired

by integrating across a large number of experiences (Marr, 1971; Milner, 1989; Alvarez and Squire, 1994; McClelland et al., 1995).

Consolidation mechanisms

Now let us turn to the role of the hippocampus in consolidation. First of all, consolidation should involve changes within neocortex, whereby whatever role the hippocampus initially played in the recall of the memory (be it storage- or retrieval-related) can now be assumed by the neocortex itself, independent of the presence or absence of the hippocampus. Second, as already discussed, consolidation should require the active participation of the hippocampus. Issues now concern the locus and nature of neocortical changes, as well as the nature of hippocampal involvement. Regarding the nature of cortical changes, if we assumed that information crucial for the recall of the memory was initially stored within the hippocampus, this information must somehow be transferred to neocortex during consolidation. This transfer should not be thought of as direct copying of synaptic efficacies from one brain area to another; as well as being physically unfeasible, such a process would also be useless since a given pattern of synaptic efficacies may have a completely different meaning in different neural contexts. Rather, the transfer of information should occur through the spread of activation from one brain area to another, resulting in activity-dependent synaptic changes in the participating areas.

One way in which such transfer can work is the following. Imagine that the information initially stored in the hippocampus can be used to reconstruct the neocortical activity pattern that characterized the original episode. This is obviously the case if the hippocampus stores a compressed representation of the episode that can be decompressed by hippocampocortical projections and cortical backprojections, but similar arguments may apply to some other cases as well. During several such reactivations, perhaps spread over a relatively long time, associations between neocortical areas or neurons can be built up gradually (in line with the proposal discussed above that the neocortex learns slowly), and these associative connections could then support memory retrieval. Some version of this idea plays a major role in most theories of consolidation (Marr, 1971; Squire et al., 1984; Teyler and DiScenna, 1986; Milner, 1989; Alvarez and Squire, 1994; McClelland et al., 1995; Murre, 1997).

Most of these accounts are not very specific about the mechanism by which new neocortical associations become established. Alvarez and Squire (1994) suggest that, as a result of

multiple, hippocampally-dependent reactivations of the scattered cortical areas representing the memory, direct connections are established between the participating neurons in different areas through Hebbian associative learning. Although Hebbian type learning is thought to occur between pairs of connected neurons in the brain (indeed, LTP/LTD has been given a Hebbian interpretation), areas of neocortex which process the different modalities that make up an episode do not normally have large numbers of direct connections, and it appears unlikely that cortical learning would require the formation of a large number of new direct long-range intracortical connections. Murre (1997) also points out this difficulty, and suggests that, instead of direct connections between the participating neurons, the formation of new associations involves the strengthening of existing connections within a chain of neurons connecting the two sites. In addition, it is noted that, rather than being densely or randomly connected, neocortex is better characterized as a loose hierarchy of areas (perhaps with multiple roots or convergence points), with the medial temporal lobe as one of the convergence points (Felleman and Van Essen, 1991). This observation suggests that the way to establish arbitrary new cortical associations should perhaps be sought in the context of hierarchical architectures.

As mentioned earlier, hippocampal reactivation of neocortical memory representations, leading to the gradual establishment of new cortical associations, is not the only way in which consolidation might work. First of all, the role of the hippocampus does not even need to be specific; i.e., it does not necessarily have to exert an influence that is selective to the cells participating in the memory in neocortex. In theories in which the role of the hippocampus is not to store memories but rather to modulate cortical processing (e.g., Carpenter and Grossberg, 1993), consolidation may be the consequence of the time-limited need for this modulatory influence, perhaps until the final cortical representation of the memory becomes established. However, in the absence of hippocampal reinstatement, these theories need to rely on the outside world to provide repetitions of the inputs for the hippocampus to act on. Therefore, theories such as Carpenter and Grossberg (1993) may not be very appropriate to describe the consolidation of memories for unique experiences (episodes).

On the other hand, if the role of the hippocampus in consolidation is specific to the cortical areas or neurons involved in the memory, there are several possibilities as to the degree of "supervision" in the process. At one extreme, it might be possible that the hippocampus is involved in working out the actual internal representations to be used by neocortex for the memory (this is close to the view taken by Gluck and Myers, 1993; Gluck et al., 1997). And if the hippocampus just assists the cortex in reinstantiating the original episode,

without specifying how the neocortex should use this information, there are still several possibilities regarding the way in which neocortical learning might proceed. For example, in the model of McClelland et al. (1995), hippocampal reinstatement provides full training examples (i.e., both inputs and desired outputs) to a multilayer neocortical neural network, which uses these examples to learn the desired input-output mapping using the error back-propagation algorithm. This kind of "supervised" cortical learning is somewhat different from the "unsupervised" approach described earlier, where no distinction between input and output areas is made as any area can provide training data, cues for retrieval, or be the subject of a query depending on context. In addition, the error-driven learning rule used in back-propagation is somewhat different in nature from the associative, Hebbian form of learning used in other models (e.g., Alvarez and Squire, 1994), although various links between the two have been discovered. In particular, an approximate form of the error back-propagation algorithm called recirculation (Hinton and McClelland, 1988), which has a higher degree of biological realism than the original form, has been shown to be equivalent to an essentially Hebbian form of learning (O'Reilly, 1996).

4.1.3 Memory consolidation and sleep

There is currently very little direct experimental evidence that would allow us to distinguish between the different possibilities outlined above regarding the nature of the initial involvement of the hippocampus in recall, its role in consolidation, and the nature of neocortical changes during consolidation. However, there is an abundance of data that provide indirect clues about these issues; some of these data (e.g., on cortical connectivity, hippocampal anatomy and physiology, plasticity, neuromodulation, effects of lesions and other manipulations in animals and humans) have already been reviewed here or in the Background chapter. Here I will briefly summarize two sets of data, one from physiology and one from behavioral studies, which together suggest the existence of special brain states (occurring during sleep) where both reactivation of hippocampal memory traces and neocortical memory consolidation seem to take place, perhaps signifying a functional link between the two.

On the physiological side, there is now considerable evidence that the spatio-temporal activity patterns of hippocampal neurons that occur during waking are replayed during subsequent slow wave sleep (SWS) (Pavlides and Winson, 1989; Wilson and McNaughton, 1994; Skaggs and McNaughton, 1996; Kudrimoti et al., 1999) and rapid eye movement sleep (REMS) (Louie and Wilson, 2001). SWS reactivation happens during physiological conditions when hippocampal activity patterns are created largely autonomously, and then

serve to influence activity in neocortex (unlike during waking, when hippocampal activity is strongly driven by cortical input) (Buzsáki, 1996). Sharp wave activity in the hippocampus (characteristic of SWS), driving activity in neocortex (Siapas and Wilson, 1998; Collins et al., 1999), probably underlies the observed reactivation of neocortical activity patterns (Qin et al., 1997), and is also thought to provide good conditions for neocortical synaptic plasticity (Buzsáki, 1989, 1996). These experiments have been reviewed in more detail in the Background chapter.

On the behavioral side, there appears to be a link between sleep and memory consolidation (for reviews, see Hennevin et al., 1995; Smith, 1995; Stickgold, 1998). The relevant data can be summarized briefly as follows. Behavioral experiments in animal learning indicate that training in a wide range of tasks results in a subsequent increase in REMS duration within a limited period of time (called "REM window") beginning a few hours after training. REMS deprivation during these REM windows can significantly reduce the effect of training on subsequent testing. The tasks affected include both appetitive and aversive, and both hippocampally-dependent (explicit) and hippocampally-independent (implicit) ones. Finally, electrical stimulation during REM windows of brain centers that are thought to play a role in controlling REMS (such as the mesencephalic reticular formation) can enhance subsequent performance (Hennevin et al., 1995).

In humans, REMS deprivation also impairs the retention of information after learning. However, just like in animals, REMS deprivation does not affect all tasks equally; in particular, it appears to affect implicit (procedural) tasks to a much greater extent than explicit (declarative) tasks. In support, it has been reported that sleep during the second half of the night (which is relatively rich in REMS but poor in SWS) enhances subsequent performance in a procedural task (mirror tracing) but not in a declarative task (recall of paired-associate lists); conversely, sleep during the first half of the night (which is poor in REMS but rich in SWS) enhances retention of declarative, but not procedural learning (Plihal and Born, 1997). On the other hand, another implicit (visual discrimination) task (Karni and Sagi, 1993), in which overnight improvement was known to require REM sleep (Karni et al., 1994), was also found to be dependent on SWS (Gais et al., 2000; Stickgold et al., 2000), so that actual task improvement was proportional to the product of the amount of SWS in the first quarter of the night and the amount of REMS in the last quarter of the night (Stickgold et al., 2000).

All in all, it appears that both REMS and SWS are likely to be important for memory consolidation, although the degree to which each of them is involved is probably task-

dependent. Taken together with the physiological data, the evidence seems to support a theory in which consolidation requires both SWS, during which the hippocampus initiates the reactivation of hippocampal-cortical memory representations, and REMS, during which some kind of reactivation may also occur, but is likely to be initiated in neocortex. On a more fundamental level, the apparent basic differences in physiology, and specifically in the nature of information flow in the hippocampal-neocortical network, between awake exploratory states and the states implicated in consolidation (particularly SWS) suggest that the roles played by the hippocampus in recall and in consolidation may be distinct rather than identical.

4.1.4 Goals of research

It should be clear from the foregoing analysis that the role of the hippocampus in memory consolidation is far from being firmly established, and that a large number of theoretical proposals regarding this role have been put forward. However, as I hope I have also demonstrated, there are many different aspects of this hippocampal involvement, most of which are debated, and thus any satisfactory account of memory consolidation would need to be explicit about all of these issues. Most of the theories discussed so far do not satisfy this requirement, making suggestions about some aspects while completely ignoring some other, closely related questions. Part of the reason why all these theories could afford to remain silent on some crucial issues is that, with a few notable exceptions (e.g., Alvarez and Squire, 1994; McClelland et al., 1995), these theories were never implemented, or tested quantitatively. Even the model of McClelland et al. (1995) is used mainly to support their computational argument, and therefore lacks any substantial degree of biological realism. Finally, the model of Alvarez and Squire (1994), even though it embodies many of the general principles of memory consolidation discussed above, suffers from several rather serious limitations. These are partly due to the model's spartan simplicity, which also makes it hard to test comprehensively.

In this chapter, we consider consolidation using a model whose complexity brings to the fore consideration of computational issues that are invisible to simpler proposals. In particular, it treats cortex as a hierarchical structure, with hierarchical codes for input patterns acquired through a process of unsupervised learning. This allows us to study the relationship between coding for generic patterns, which forms a sort of semantic memory, and the coding for the specific patterns through consolidation. It also allows us to consider consolidation

as happening in hierarchical connections (in which the cortex abounds) as an alternative to consolidation only between disparate areas at the same levels of the hierarchy. The same general framework also allows us to model other related forms of memory, including familiarity-based recognition and repetition priming. The next section describes the model in detail; we then provide an analysis of its mechanisms and performance.

4.2 The model

4.2.1 Main concepts

Let us first consider our model of neocortex in isolation. It consists of a set of neocortical areas which are collectively assumed to provide an internal representation of the perceived or remembered state of the world. These areas may be thought of as higher order association areas of different modalities; we do not explicitly consider lower sensory areas which are assumed to be involved in extracting relevant information from sensory inputs, and, at least for now, we also neglect frontal areas involved in processes like working memory and executive functions. We treat neocortical areas in the medial temporal lobe (entorhinal, perirhinal, and parahippocampal cortices) separately, because these areas integrate information from all processing streams in each modality, and interact directly with the hippocampal formation. Based on anatomical evidence (Felleman and Van Essen, 1991; Lavenex and Amaral, 2000), we conceptualize neocortex (perhaps excluding frontal areas) as an approximately hierarchical structure, with primary sensory areas at the bottom, and entorhinal cortex at the apex. Connections between areas are assumed to be reciprocal, again in good agreement with anatomical data. Importantly, there are typically no direct connections between areas at lower levels of the hierarchy (although there are some notable exceptions), and these areas can only communicate through areas at the higher levels. In our work so far, we have used one of the simplest structures that embodies these anatomical principles. The network we will consider here has only two layers; the bottom layer contains several generic neocortical areas, which are connected bidirectionally to a single area in the top layer (Fig. 4.1). This model region at the top of the cortical hierarchy is assumed to correspond to the neocortical regions of the medial temporal lobe, and we tentatively call this region the entorhinal/parahippocampal/perirhinal area (E/P). The neocortical areas on the lower level will be referred to as areas A, B, C, etc.

Within each area of the model network, there are a large number of abstract (binary),

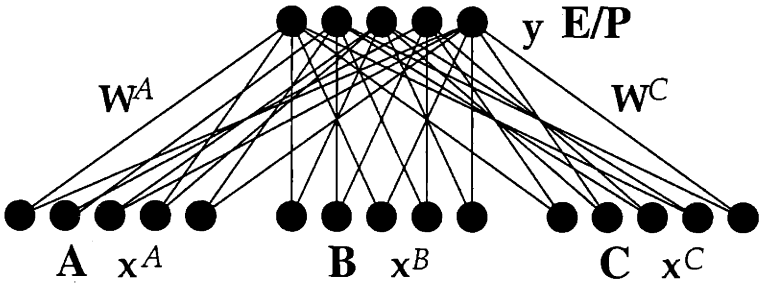


Figure 4.1: Architecture of the neocortical model. All units in neocortical areas A, B, and C re connected to all units in area E/P through bidirectional, symmetric weights, but there are no connections between units on the same level.

neuron-like units. A given pattern of activities of neurons in one of the lower level areas encodes a particular stimulus, but only a small subset of all possible activation patterns corresponds to stimuli that actually occur in the world. The *vector* of activations in area A caused by stimulus i ($i = 1, 2, \ldots, r$) will be denoted by $\mathbf{x}_i^A$, and activity in E/P by $\mathbf{y}$. Areas A, B, and C can be thought of as representing different modalities, and thus, at least in principle, we may encounter any combination of the possible patterns in each area (e.g., $\mathbf{x}_1^A\mathbf{x}_1^B\mathbf{x}_1^C$, $\mathbf{x}_2^A\mathbf{x}_2^B\mathbf{x}_2^C$, $\mathbf{x}_1^A\mathbf{x}_3^B\mathbf{x}_2^C$, ...). For now, we even make the further simplifying assumption that all possible combinations of stimuli are experienced equally often; this also implies that the identity of the pattern seen in some areas (say, A and B) in general carries no information about the identity of the pattern in a different area (say, area C).

Now we have to specify the dynamics of activities in our hierarchical network, i.e., how the activity of a neuron is updated as a function of the inputs it receives, as well as the dynamics of learning, i.e., how the efficacies of the connections are updated as a function of unit activities. Here, the model borrows from the extensive theories about the formation of hierarchical representations in cortex that comes from work in unsupervised learning (e.g., Hinton and Sejnowski, 1999). This approach is based on the reasonable premise that raw sensory information (e.g., the activities of photoreceptors in the retina, or even the activities of neurons in primary sensory areas of neocortex) is not in a form that is particularly well-suited for the operations assumed to underlie higher cognitive functions (such as reasoning, decision-making, or memory), and therefore has to be re-represented in a more suitable format. One particularly useful way to represent information about the world may be to identify the causes underlying perceived inputs. Since these underlying causes are responsible for the statistical structure (i.e., regularities) in the input, identification of these causes can lead to a very efficient description of non-random aspects of the input data. In

order to account for the variability of natural stimuli, and the resulting uncertainty about the underlying causes, most unsupervised learning models are probabilistic, specifying the probabilities that various causes underlie particular stimuli.

A basic assumption in unsupervised learning is that the underlying causes are not identified explicitly when the inputs are presented; instead, possible causes need to be inferred from the statistical structure in the collection of input examples. It also follows from this assumption that a good way to judge the performance of such a model is to see how well the causes it extracts can explain and reproduce (statistically) the inputs they are supposed to represent. In the class of models considered here, this argument is turned around such that the primary goal is now to specify a model (called the generative model) which creates synthetic data based on assumed underlying causes. This generative model has a collection of parameters and a general form (or structure) that determines how the parameters specify the distribution over the inputs. Over the course of learning, the parameters are adjusted until the real input distribution and the synthetic distribution specified by the generative model are as similar as possible. In other words, the goal of learning is to maximize the likelihood that the actual data could have been created by the generative model.

Once the generative model provides a good fit to the distribution of the input data, the model can also be used for "recognizing" novel inputs (i.e., identifying the probable underlying causes for any new instantiation of the inputs). The reason for this is that the recognition model (i.e., the probability of the possible causes given the input data) is the statistical inverse of the generative model (which gives the probability of inputs given the underlying causes), and as such can be obtained from it by applying Bayes' theorem (although this operation often cannot be performed directly due to practical limitations). Furthermore, a good statistical model of the input distribution also makes it possible to do probabilistic inference about some inputs based on others. For example, in our case we may denote the real (joint) distribution over possible input patterns by $P[\mathbf{x}^A, \mathbf{x}^B, \mathbf{x}^C]$, and its approximation by the generative model by $P[\mathbf{x}^A, \mathbf{x}^B, \mathbf{x}^C; \mathbf{W}]$ (since it is determined by the parameters of the generative model, which in our case are the network weights and biases $\mathbf{W}$). Then $P[\mathbf{x}^C \mid \mathbf{x}^A, \mathbf{x}^B]$, the probability of observing some pattern $\mathbf{x}^C$ in area C given that we know that the patterns in areas A and B are $\mathbf{x}^A$ and $\mathbf{x}^B$, respectively, can be approximated as $P[\mathbf{x}^C \mid \mathbf{x}^A, \mathbf{x}^B; \mathbf{W}]$, which in turn can be calculated from the generative model.

Thus, we model the neocortical network as a hierarchical generative model, which is trying to extract from inputs $\mathbf{x}^A, \mathbf{x}^B, \mathbf{x}^C$ the underlying causes in the form of E/P activations

$\mathbf{y}$. The type of information acquired by the network through unsupervised learning is probably best characterized as general semantic knowledge, and includes information about categories, tendencies, and correlations – what we may refer to as statistical structure in the world. On the other hand, it should be noted that this kind of learning is generally not very well suited for storing individual input patterns, especially if it needs to be accomplished in a single presentation. However, storage of individual patterns is a crucial aspect of episodic memory, which may be conceptualized as the ability to store specific patterns in a single learning episode, and later recall them based on partial or noisy input. How is it then that the brain is obviously capable of episodic remembering?

Part of the answer may come from the proposal (described earlier) that initial fast storage of specific patterns occurs not in neocortex, but in the hippocampus. It has also been suggested that the hippocampus is very well suited for the pattern completion operation underlying the ability to recall patterns based on partial or noisy cues. However, episodic memory recall may eventually become independent of the hippocampus (Rempel-Clower et al., 1996; Reed and Squire, 1998) (cf. Nadel et al., 2000). And what about memory for facts, which is also not probabilistic in nature, initially also depends on the hippocampus, but almost certainly becomes independent of the hippocampus later on? If such consolidation does happen, the neocortex eventually has to be able to support the storage and recall of specific patterns.

A clue about how this might be possible comes from the observation that probabilistic inference and pattern completion are not so different computationally as they might seem. In fact, the inference example described above, which involved generating samples from a conditional distribution such as $P[\mathbf{x}^C \mid \mathbf{x}^A, \mathbf{x}^B]$, can also be viewed as probabilistic pattern completion. Indeed, it preferentially produces samples $\mathbf{x}^C$ which were likely to be presented together with $\mathbf{x}^A$ and $\mathbf{x}^B$. And if the distribution $P[\mathbf{x}^C \mid \mathbf{x}^A, \mathbf{x}^B]$ is very peaked, i.e., if there is only one likely $\mathbf{x}^C$ for a given $\mathbf{x}^A$ and $\mathbf{x}^B$, then the network essentially performs deterministic pattern completion. Therefore, if the neocortical network (which normally performs probabilistic unsupervised learning) can be biased so that the probability distribution that it represents becomes very peaked around certain patterns, these patterns will have been stored by the network and can be recalled by presenting partial or noisy versions. Finally, since the model's estimate of the input probability distribution is based on counts of occurrences of input patterns, biasing the network can be accomplished by simply providing more repetitions of the specific patterns to be stored. Consequently, the hippocampus may effect the consolidation of specific patterns (whether these represent episodes or factual knowledge) by reinstating the patterns repeatedly in neocortex while learning in neocortex

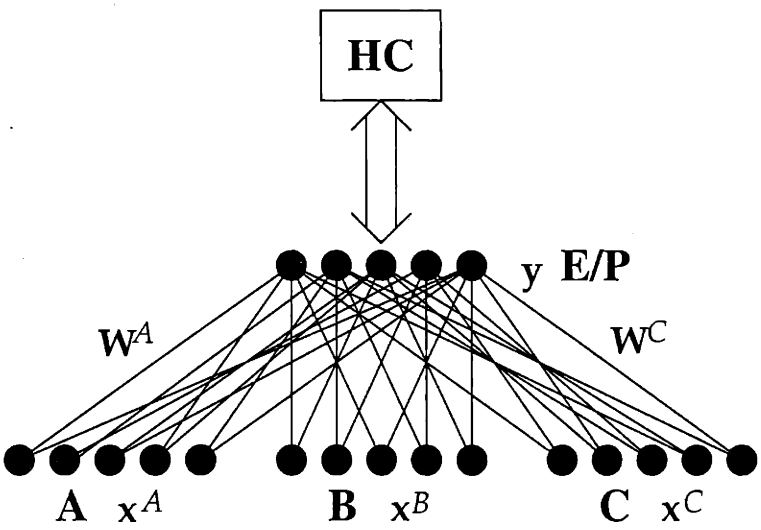


Figure 4.2: Architecture of the full model. The hippocampus (HC) is not directly implemented, but it can influence and store the patterns in E/P. All communication between the HC and the input areas is via area E/P.

proceeds in the same way as it does during external presentation of input patterns. In addition, due to the existence of the neocortical generative model, input patterns need not be recreated directly; instead, the hippocampus can recreate just the compressed pattern corresponding to the input pattern at the highest level of the hierarchy (area E/P), and rely on the cortical backprojections to reinstate the original input pattern (see Figure 4.2).

Thus, our qualitative model of the roles of the hippocampus and neocortex in declarative memory can be summarized as follows. First, the neocortex is continuously involved in learning a hierarchical probabilistic model of the set of patterns appearing in its input areas. This process results in the formation of a knowledge base of general semantics. Second, the hippocampus, interacting with the top level of the cortical hierarchy, can store specific patterns instantaneously, and supports early recall of these patterns by using its intrinsic pattern completion capability. The completed hippocampal pattern is then used to complete the activity pattern in areas E/P, which in turn leads to the recall of the full pattern in other neocortical areas through the cortical generative model. Finally, these memories are consolidated through hippocampally-driven reinstatement of the corresponding neocortical representations, which biases the neocortical probabilistic model in a way that eventually allows hippocampally-independent cortical recall of specific patterns. In order to determine the feasibility of this general scheme, we implemented it in a network model, which we

describe next.

4.2.2 The neocortical model and semantic learning

The restricted Boltzmann machine

For the implementation of the two-level cortical hierarchy we chose the unsupervised learning model known as the restricted Boltzmann machine (RBM). The Boltzmann machine is an often used and statistically well-founded way to approximate probability distributions over large sets of binary variables (Hinton and Sejnowski, 1986). It is also closely related to the Hopfield net, the paradigmatic autoassociative memory (Hopfield, 1982), and it is this relationship between probabilistic generative models and autoassociative memory that will allow the model to represent both episodic and semantic information.

The Boltzmann machine consists of a set of binary nodes and real-valued symmetric weights connecting some of the nodes. The Boltzmann machine that we consider is *restricted* because the nodes are separated into layers, and all connections are between layers (rather than within the layers). The probability distribution represented by the Boltzmann machine is encoded in the weights. More specifically, if we denote the activity state of the i'th input unit (irrespective of which input area it belongs to) by x_i, the state of the j'th unit in E/P by y_j, and the weight connecting the two by W_{ij}, we can define the *energy function* of the RBM as

$$E(\mathbf{x}, \mathbf{y}; \mathbf{W}) = -\sum_{i,j} W_{ij} x_i y_j - \sum_i b_i x_i - \sum_j c_j y_j, \tag{4.1}$$

which depends explicitly on the x_i's and y_j's, and contains the weights (including bias terms b_i and c_j) as parameters. The probability distribution over possible states of the network is then defined as

$$P(\mathbf{x}, \mathbf{y}; \mathbf{W}) = \frac{1}{Z} e^{-E(\mathbf{x},\mathbf{y};\mathbf{W})} \quad \text{where} \quad Z = \sum_{\mathbf{x},\mathbf{y}} e^{-E(\mathbf{x},\mathbf{y};\mathbf{W})} \tag{4.2}$$

and the last sum is over all possible instantiations of the x_i's and y_j's. $P(\mathbf{x}, \mathbf{y}; \mathbf{W})$ is called the Boltzmann distribution, and Z is referred to as the partition function. Under the Boltzmann distribution, states with lower energies are more likely. Of course, the joint distribution over E/P states (also called "hidden" states) and input states represented by

Equation 4.2 cannot be compared directly with data (where only the activities in A, B, and C are given). However, Equation 4.2 also implies a distribution over just the input states, which can be obtained by summing over all possible configurations of the hidden states,

$$P(\mathbf{x};\mathbf{W}) = \sum_{\mathbf{y}} P(\mathbf{x},\mathbf{y};\mathbf{W}) = \frac{1}{Z}\sum_{\mathbf{y}} e^{-E(\mathbf{x},\mathbf{y};\mathbf{W})}. \tag{4.3}$$

This is essentially equivalent to ignoring the states of the E/P units for the purpose of determining the input distribution generated by the network.

One way to obtain samples from this distribution (called Gibbs sampling method) is to use an appropriately defined dynamics of the activities (i.e., the states of the binary nodes). In this scheme, activity updates depend on the activities of units connected to the unit to be updated and on the values of these connecting weights. In the RBM, since there are no connections within layers, units within a layer can be updated synchronously, and the dynamics of activity in the network consists of alternating updates in the two layers. At each step, units within the updated layer are set according to the Gibbs sampling rule

$$x_i = \begin{cases} 1 & \text{with probability } \sigma\left(\sum_j W_{ij}y_j + b_i\right) \\ 0 & \text{with probability } 1-\sigma\left(\sum_j W_{ij}y_j + b_i\right) \end{cases} \tag{4.4}$$

for units in the input areas, where $\sigma(x) = 1/(1+\exp(-x))$ is the standard sigmoid function, and similarly

$$y_j = \begin{cases} 1 & \text{with probability } \sigma\left(\sum_i W_{ij}x_i + c_j\right) \\ 0 & \text{with probability } 1-\sigma\left(\sum_i W_{ij}x_i + c_j\right) \end{cases} \tag{4.5}$$

for units in area E/P in the alternate steps. The states of the units after a sufficiently large number of iterations provide a sample from the probability distribution defined by Equation 4.2; ignoring the states of the hidden units and only looking at the states of the input units yields a sample from the distribution of Equation 4.3.

Probabilistic inference of the kind discussed earlier can be carried out in the Boltzmann machine as follows. Input to a set of units *clamps* the activities so that they do not change. Activities of all other units are updated many times according to Equations 4.4 and 4.5. Given clamped activities in, say, $\mathbf{x}^A$ and $\mathbf{x}^B$, the resulting activities in $\mathbf{x}^C$ represent a *sample*

according to a distribution

$$P[\mathbf{x}^C \mid \mathbf{x}^A, \mathbf{x}^B; \mathbf{W}] \tag{4.6}$$

where $\mathbf{W} = \{\mathbf{W}^A, \mathbf{W}^B, \mathbf{W}^C, \mathbf{b}, \mathbf{c}\}$. Given appropriate weights $\mathbf{W}$, this is also how the model can perform autoassociation, completing $\mathbf{x}^A, \mathbf{x}^B$ to the best fitting $\mathbf{x}^C$.

The weights $\mathbf{W}$ are assumed to be subject to slow plastic changes in order to fit distributions such as that in expression 4.6 to the statistics of the patterns presented. The learning rule is based on the standard Boltzmann Machine learning algorithm (Hinton and Sejnowski, 1986) using Gibbs sampling, with a modification that was introduced recently in the context of a general class of generative models called products of experts (of which the RBM is a simple example) (Hinton, 2000). This learning rule involves one phase of Hebbian learning driven by activity patterns from the world, and one phase of anti-Hebbian learning driven by patterns generated in response by the network. In the positive (Hebbian) phase, an input pattern is presented to the input layer (i.e., areas A, B, and C). The corresponding activity pattern in E/P (the so-called "hidden" layer) is then determined by applying the activity update rule (equation 4.5), and the weights between the two layers are *increased* in proportion to the product of the activities of the nodes connected. The negative (anti-Hebbian) phase starts with an update of the activities in the input layer based on the activities in E/P determined in the positive phase. Then, based on these activities, the activities in E/P are updated for a second time, and the weights are *decreased* in proportion to the product of the activities generated internally in the negative phase. If the training examples come from a distribution that is constant in time, learning causes the probability distribution represented by the generative model to approach the true input distribution (with the distance between two distributions defined as the Kulback-Leibler divergence). Learning stops when the distribution generated by the Boltzmann Machine (as defined by Equation 4.3) is as close as possible to the true input distribution $P(\mathbf{x})$.

Simulations of the isolated neocortical network

In the first set of simulations using this architecture, we addressed some basic issues regarding general (semantic) learning. In these simulations, there were three input areas (denoted by A, B, and C), and each of the four cortical areas (A, B, C, and E/P) contained 100 units. For each of A, B, and C, we generated 10 random binary patterns (denoted $\mathbf{x}^{A_1}$–$\mathbf{x}^{A_{10}}$, $\mathbf{x}^{B_1}$–$\mathbf{x}^{B_{10}}$ and $\mathbf{x}^{C_1}$–$\mathbf{x}^{C_{10}}$, each bit of which is turned on with probability 1/2).

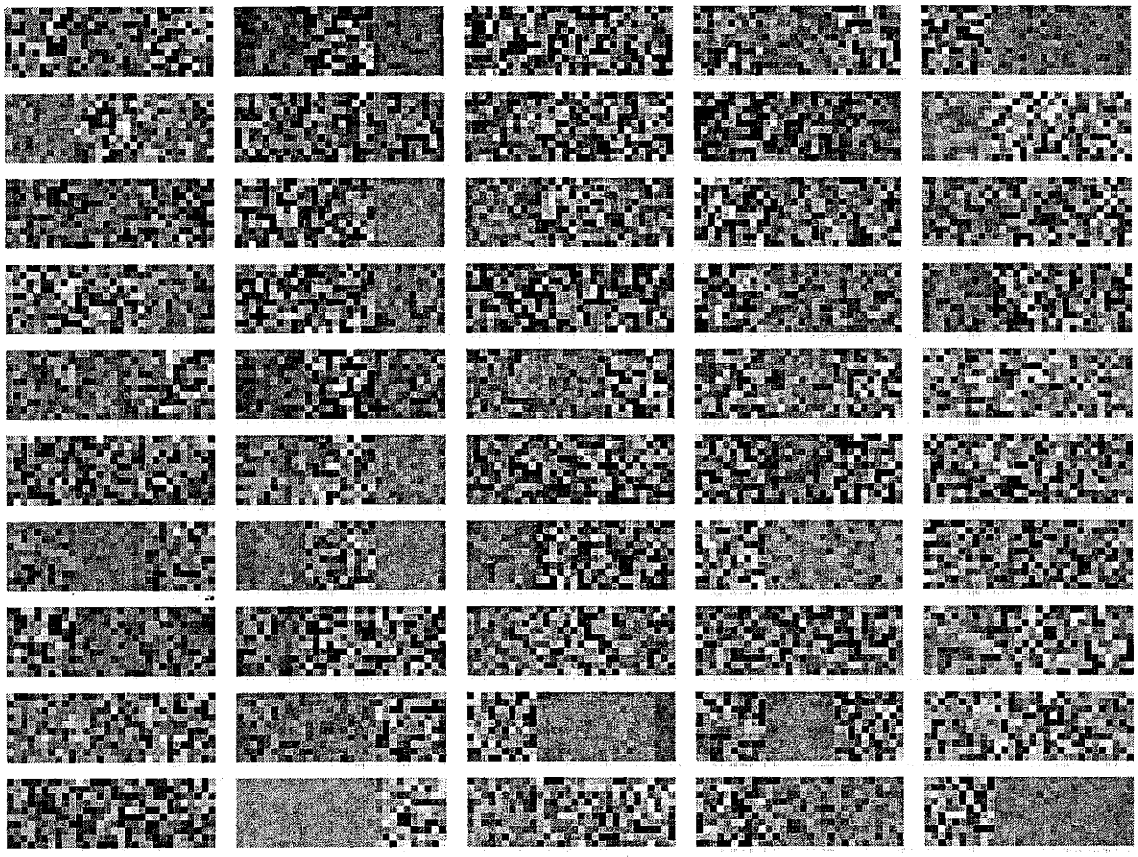


Figure 4.3: E/P receptive fields. Examples of the weights received by units in E/P from areas A-C after pre-training. Each rectangle represents one unit in E/P, and each little square inside the rectangles represents a weight; the shade of the square indicates the magnitude of the weight (black – large positive, white – large negative). The spatial arrangement of the weights is only for visualization purposes; the weights from areas A, B, and C are shown in the left, middle, and right third of the rectangle, respectively.

These stand for the different stimuli that can be represented in any one of these areas. During the semantic learning phase, these patterns are presented to the network in random combinations (e.g., an input example could be $\mathbf{x}^{A_1}\mathbf{x}^{B_6}\mathbf{x}^{C_3}$), which corresponds to prior exposure of the system to events involving different combinations of stimuli. In all, 50,000 presentations were made and cortical weights were modified using the RBM learning algorithm. This leads to a population code in E/P for the patterns presented in the input layer, and establishes a correspondence between the representations in areas A-C and those in E/P in the form of a generative model.

One way to represent the result of this learning is to look at the weights going into (and also out of) individual units in E/P, which can be thought of as the "receptive fields" of the E/P units in the space of input activities. Some examples for different E/P units are shown in Figure 4.3; weights are coded by gray-scale, and weights from a specific input area are blocked together. Some of the units seem to specialize in representing just a single input area and ignore the others, while other units are completely global. It is interesting to note the appearance of these global units, since all inter-area correlations are small and incidental in this case due to the independent presentation of patterns in different input areas, and therefore the input probability distribution could be modelled perfectly with receptive fields that are strictly localized to one of the input areas. On the other hand, these units with global weight vectors are exactly the ones that allow the network to produce valid activity samples in an area in the absence of external input to that area. More importantly, such higher order units with multi-modal input become critical in the more realistic case when there are meaningful correlations between activities in different areas.

In order to see that these weights have indeed established a correspondence between areas A-C and E/P, we can look at the quality of one-step reconstructions of input patterns based on their "code" in E/P. This can be done by presenting to areas A-C valid input patterns (i.e., patterns which are generated the same way as the training patterns and therefore conform to the same statistical structure), determining the corresponding E/P pattern through application of the activity update rule to E/P units, and then determining what reconstructed A-C pattern this E/P activities would imply in the absence of external inputs by updating the activities of input units. We can then calculate the average one-step reconstruction error by summing over input units the squared differences between the original and the reconstructed activations, and averaging this quantity over many presentations of different valid input patterns. Now we can examine how the average one-step reconstruction error changes as we train the network. The result is shown in Figure 4.4. The average reconstruction error decreases fairly rapidly, and, as can be seen in part b of the figure, continues to decrease steadily with further training. Another interesting fact that can be seen in the log-log plot in Figure 4.4b is that, after an initial slower phase, the decrease of the average reconstruction error follows a power law (hence the linear section in the log-log plot) whose exponent is around -2. In other words, the error is inversely proportional to the square of training time. Therefore, as a result of semantic training, the model quickly establishes compressed codes in E/P for valid input patterns.

We can also check whether the trained network can do more than just represent the full pat-

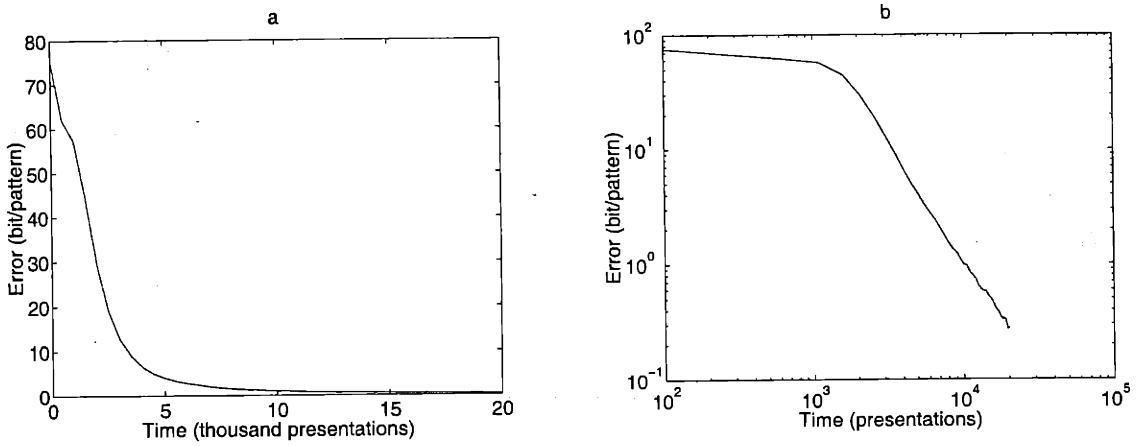


Figure 4.4: Quality of general (semantic) representation. Average error in one-step reconstruction by the network of all valid input patterns (i.e., all possible combinations of the valid patterns in each of the input areas) as a function of the length of training on randomly selected valid patterns. The two parts of the figure show the same data, **a** on a linear scale and **b** on a log-log scale.

terns that it has seen. In particular, the model should be capable of performing probabilistic inference as discussed earlier. This was evaluated in the following way. We initialized the network with valid patterns in areas A and B, and a random binary pattern unrelated to the training patterns in area C. Next, 20 updates of the full network were run with activities in A and B clamped, which in this case corresponds to 20 cycles of alternating updates in area E/P and area C. The resulting activity pattern $\mathbf{x}^C$ was regarded as the completion of the pattern $\mathbf{x}^A\mathbf{x}^B$ as inferred by the network. The first thing we can check is whether this pattern corresponds to any of the valid patterns in area C. This was done by comparing the pattern with all the valid patterns for area C, and a squared distance of less than 5 was considered a match (the squared distance between two independent random patterns has an expected value of 50 and standard deviation of 5, so a random match between two patterns is extremely unlikely). The percentage of samples generated by the network that matched one of the valid patterns is shown in Figure 4.5a as a function of the amount of preceding semantic training. The model learns to generate, in response to partial patterns, complete patterns which are similar to the full patterns it was trained on.

However, if the network has truly learned the input probability distribution, it should also produce the different valid patterns $\mathbf{x}_i^C$ in proportion to the actual probability $P[\mathbf{x}^C \mid \mathbf{x}^A, \mathbf{x}^B]$ of their appearing together with the current $\mathbf{x}^A\mathbf{x}^B$ during training. In our case, since

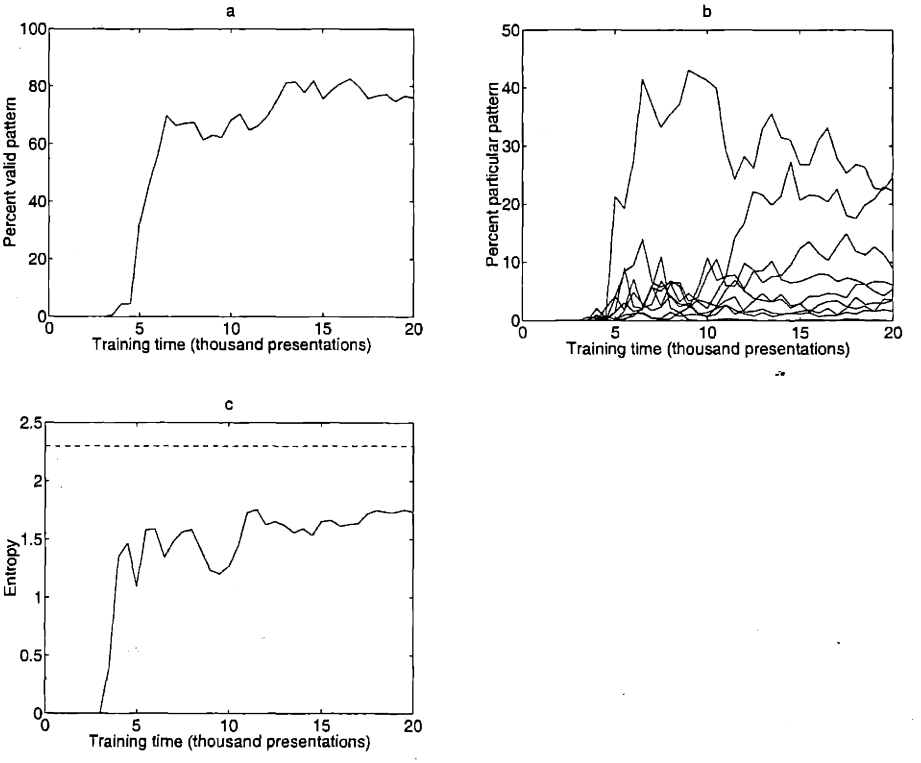


Figure 4.5: Inference in the semantically trained network. **(a)** Percentage of times a partial pattern (a valid pattern in two input areas) was completed by the network with a valid pattern in the third area, as a function of the amount of preceding general training. An error of 5 was allowed in assessing the validity of the completed pattern. **(b)** Percentage of times when completion resulted in a particular valid pattern (represented by the individual curves). **(c)** Entropy (diversity) in valid completions. The maximum possible entropy (about 2.30, corresponding to the case when all valid patterns are generated with equal probability) is represented by the dashed line.

all combinations of valid patterns were presented with equal probability during training, all valid patterns in C should be generated by the network equally often in response to any valid partial pattern in A and B. The actual percentages of completions to different patterns in a case where there were 10 valid patterns in each area (and therefore each of the 10 patterns in C should appear in 10% of the cases) are shown in Figure 4.5b. It is clear from the figure that, after some semantic training, there is substantial diversity in the completed patterns. It is also clear, however, that the actual percentages of the different patterns do not become equal, although there appears to be some tendency towards equality as training goes on. In order to quantify the diversity of completion, we calculated the observed entropy of the distribution of completed patterns, which is shown in Figure 4.5c. The entropy generally increases with more training, and it goes up to about 3/4 of the maximum possible entropy (about 2.30) during the initial part of training shown in the figure. Finally, it should be noted that even those valid patterns that are rarely (or never) generated by the network in response to ambiguous partial cues can nevertheless be represented and reconstructed when presented in full, as we saw before, and they can be also generated from partial cues that are sufficiently similar to them.

We can also demonstrate that this neocortical network by itself is incapable of long-term remembering of specific patterns in the face of ongoing processing of general information. We started with the network that has undergone general semantic training as described above, and which has therefore learned to represent and, at least to some extent, make inferences about probable input patterns. One specific valid input pattern (say, $\mathbf{x}_1^A\mathbf{x}_1^B\mathbf{x}_1^C$) was designated as a pattern to be stored episodically, i.e., a pattern that the network should be able to recall reliably based on partial cues. Then we provided the pretrained network with different amounts of training (assumed to take place during a single presentation) on this particular pattern, and monitored the network's ability to reconstruct this pattern based on partial cues, as well as its ability to represent other valid patterns. Recall performance on the episodic pattern was assessed by presenting the network with the partial pattern $\mathbf{x}_1^A\mathbf{x}_1^B$, and letting the network infer $\mathbf{x}^C$ in the same way as described above. Recall performance was defined as the percentage of times when the network completed the pattern to $\mathbf{x}_1^C$ (again with a final squared distance of less than 5). The ability to represent other patterns was again measured by the average one-step reconstruction error for valid patterns. As shown in Figure 4.6a, recall performance on the episodic pattern increases steadily with more training on that pattern, and eventually reaches 100%. However, paralleling this improvement on that specific pattern, the ability of the network to represent other patterns declines, as evidenced by the increase in the average one-step reconstruction error shown in

Figure 4.6b. This result was fully expected; however, it is interesting to note that the effect of episodic training on general representational capacity is far from being catastrophic – the average reconstruction error remains below 1, which corresponds to a single bit flipped per 300 bit pattern. This is somewhat surprising given the devastating magnitude of the interference effect resulting from new learning described in other studies (e.g., McCloskey and Cohen, 1989; McClelland et al., 1995); the crucial difference here is that the pattern used for episodic training is not entirely new, since it is one of the patterns used during pretraining; more importantly, it conforms to the general statistical structure represented by the network, and therefore relatively subtle changes in the weights can bias the retrieval probability for the episodic pattern without having a catastrophic effect on the processing of unrelated patterns. On the other hand, the fact that episodic training did have a substantial detrimental effect on the quality of general representations indicates the need for some kind of maintenance process, which could most simply take the form of ongoing training on general patterns.

Therefore, we tested the persistence of the established episodic memory in the face of subsequent processing in the neocortical network. For this, we took the network that had been trained to asymptotic (100%) performance on episodic recall of a single pattern following general pretraining (represented by the endpoints of the graphs in Figures 4.6a,b). This network was then subjected to further general semantic training similar to pretraining, and episodic recall and general representational ability were monitored as training went on. First looking at the average one-step reconstruction error displayed in Figure 4.6d, we see further evidence that episodic training did not change the existing semantic representation in any fundamental way. Once general training resumed, the average reconstruction error returned to previous levels (or lower) almost instantaneously, indicating that only a small change in the weights was required. However, as shown in Figure 4.6c, recall performance on the episodic pattern also declined rather rapidly once training on that single pattern was replaced by general training (which includes the episodic pattern as well as all related and unrelated valid patterns). In other words, this scheme does not seem to support the long-term retention of memories for specific patterns, at least not in parallel with learning or maintaining representations for general patterns. Even though such a statement may be difficult to prove formally, we believe that this incompatibility between general and specific learning is a general characteristic of connectionist unsupervised learning systems such as ours.

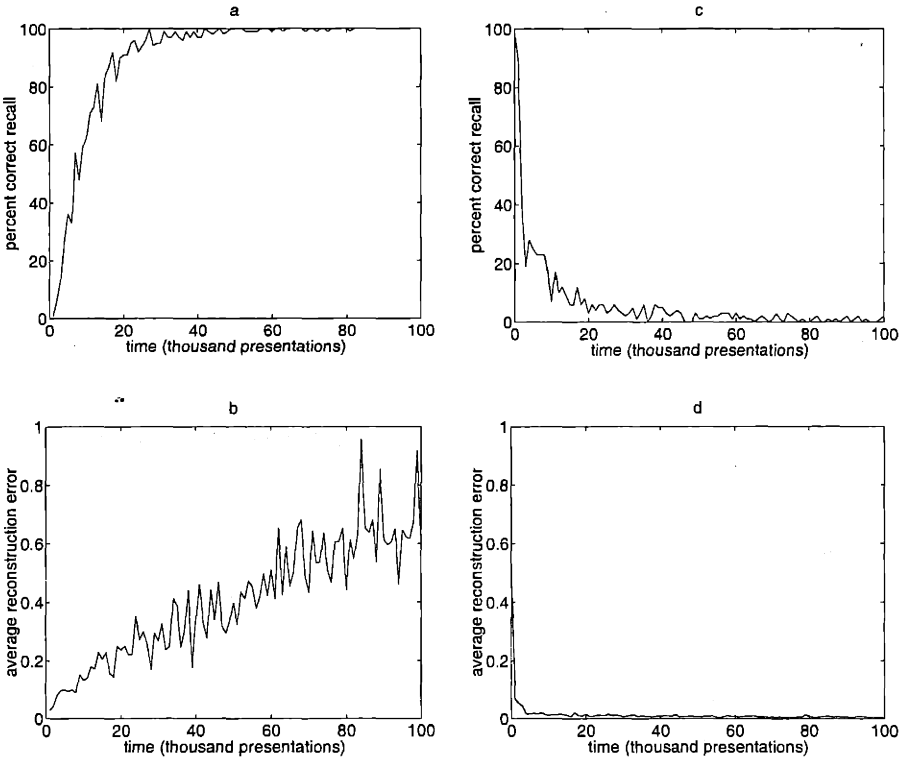


Figure 4.6: Interactions between episodic and semantic learning in the isolated neocortical network. **(a)** Percentage of times when a particular input pattern was successfully recalled based on a partial cue (the correct pattern in two input areas) as a function of the amount of training on that particular pattern (after pre-training on general valid patterns); **(b)** the effect of this episodic training on the error in one-step reconstruction of all valid input patterns. **(c)** and **(d)** The same measures of successful recall of the specific pattern and error in the reconstruction of general patterns, respectively, as a function of the amount of training on all valid patterns after the end of training on the specific pattern (corresponding to the endpoints of the graphs in parts (a) and (b)).

4.2.3 The hippocampal-neocortical network and episodic learning

Adding the hippocampal component

So far we have described how a hierarchical model of neocortex, implemented as a probabilistic generative model, can learn the general statistical structure of its inputs. We have also shown that such a network by itself is not capable of both maintaining this general information and acquiring and retaining memories of specific input patterns. As discussed earlier, it has been suggested that the hippocampus (contributing to both initial storage and subsequent consolidation) may be necessary for the learning and retention of specific information. In order to explore this idea, we added a hippocampal component to our model, and examined how this affects the processing of both specific and general information.

The hippocampus is not modeled explicitly in the current model. Instead, based on substantial experimental and theoretical evidence, it is assumed to be capable of three operations: (1) under appropriate conditions (*eg* when the current stimulus configuration is salient or important for some other reason), it stores a representation of the current E/P pattern, which allows the reinstatement of that E/P pattern if the corresponding hippocampal memory state is activated; (2) if the current E/P pattern is sufficiently similar to a previously stored pattern, the hippocampal representation for the stored pattern is activated (hippocampal pattern completion), and, consequently, the stored E/P pattern gets reinstated; (3) during consolidation, the patterns stored in the HC are activated intrinsically and randomly (this may happen during slow wave sleep Wilson and McNaughton (1994)), which leads to the reinstatement of cortical patterns in the same way as during completion, at least given appropriate weights.

This characterization of the various ways in which the hippocampus interacts with neocortex is supported by several lines of evidence, both theoretical and experimental. First, as argued earlier, the above scheme represents one of the most parsimonious ways to account for the basic findings on temporally graded retrograde amnesia. Second, there is now substantial evidence from computational studies which suggests that it may be crucial for fast learning of specific information to be anatomically and functionally separated from the representation of knowledge abstracted from a large number of examples, and that these instances of initially separately stored information can later be integrated into the general knowledge representation through reactivations spread over a longer time period. Failure to separate memories of specific instances acquired in one presentation from long-term representations of general knowledge may lead to a catastrophic breakdown of the

latter following new one-shot learning, as in McClelland et al. (1995), or to rapid forgetting of novel specific information due to subsequent general learning, which is the case in our neocortical network as discussed above. Third, the proposed roles for the hippocampus are entirely consistent with the available anatomical and physiological data, both on the internal characteristics of the hippocampus and on the ways it interacts with neocortex. Finally, there is a long tradition of modeling the hippocampus, and several of these studies suggest ways in which different parts of the above proposal (e.g., one-shot learning, pattern completion, and autonomous reactivation) might be implemented (Marr, 1971; McNaughton and Morris, 1987; Treves and Rolls, 1992, 1994; O'Reilly and McClelland, 1994; Hasselmo et al., 1996; Shen and McNaughton, 1996).

Modeling memory consolidation

Using this simple model of the hippocampus in conjunction with the neocortical network described earlier, we first modeled an experimental paradigm often used in animal studies of retrograde amnesia (e.g., Zola-Morgan and Squire, 1990; Cho and Kesner, 1996; Wiig et al., 1996; Anagnostaras et al., 1999). In this paradigm, animals learn about several different sets of stimuli (e.g., in the context of a sensory discrimination task) at different times before the hippocampus (or some other associated structure) is lesioned. Performance on all sets of stimuli is tested after recovery from surgery. Temporally graded retrograde amnesia is often found in these experiments; i.e., lesioned animals are often more impaired on recently learned material than older material, which is the opposite of the pattern seen in normals.

In the basic simulation, each of the four cortical areas (A, B, C, and E/P) contained 100 units. For each of A, B, and C, 20 random binary patterns (denoted $\mathbf{x}_1^A - \mathbf{x}_{20}^A$, $\mathbf{x}_1^B - \mathbf{x}_{20}^B$, and $\mathbf{x}_1^C - \mathbf{x}_{20}^C$) were generated, which represent different stimuli in the modality of that area. A general training phase was run first, which was identical to the general training described earlier for the neocortical network, and consisted of 50,000 presentations of random combinations of the valid patterns in the three input areas. Next, 18 specific input patterns ($\mathbf{x}_1^A \mathbf{x}_1^B \mathbf{x}_1^C - \mathbf{x}_{18}^A \mathbf{x}_{18}^B \mathbf{x}_{18}^C$) were designated as episodic patterns to be memorized. These were introduced at different times during subsequent training (with 50,000 pattern presentations between the initial storage of two adjacent episodes), whereupon their corresponding E/P population codes were determined and stored in the hippocampus. These stored representations can support the initial hippocampally-dependent recall of the episodic patterns as we will describe shortly, and are also used during consolidation.

Consolidation itself was modelled as follows. In the presence of stored patterns in the hippocampus, the model network alternated between two types of learning events. The first was identical to those in the general training phase, and corresponds to continued exposure to the same kinds of stimuli (while awake, and perhaps also during a sleep stage such as REM sleep, characterized by independent random activations of cortical areas). The second type of learning event started from hippocampal reactivation in E/P of one of the stored memory patterns. The selection among the stored patterns was random, with the probability of a given pattern being selected proportional to the "strength" of that hippocampal memory trace, which decayed exponentially in time with a time constant of 200,000 presentations. The resulting E/P activation led to the reactivation of the input areas using the usual top-down activation update rule.

A small modification of the neocortical dynamics was introduced in this phase. In particular, we assume that general learning establishes a relatively weak local attractor structure within each cortical input area, so that local interactions (which are not implemented explicitly) bias representations within that area slightly towards previously experienced patterns. Only the assumed end-result of this local processing was implemented; specifically, whenever feedback from E/P, in the absence of feedforward activation, resulted in a pattern in an input area that was close to (again, a squared distance of less than 5 was used) one of the valid patterns in that area, the activation pattern was changed (locally) to that valid pattern. This assumption actually is not critical for any of the basic results presented here; however, it was found to have a stabilizing effect during very long periods of consolidation, preventing a drift in the E/P representation corresponding to episodic patterns.

Finally, after local attractors had modified the activity patterns in the input areas resulting from reactivation by E/P, the weights were modified using the same learning algorithm that was responsible for general learning. Actually, two different versions of the algorithm were tested, and were found to perform very similarly. The difference was in the identity of the activities used in the Hebbian and anti-Hebbian parts of the learning rule. In one version, the hippocampally reactivated E/P pattern and the input pattern resulting from activation by E/P (supplemented by local attractor activity) are used in the Hebbian part, and two steps of autonomous updates within the neocortical network (the first in E/P, the second in A, B, and C) are used to generate the activities for the anti-Hebbian part. In the other version of the learning rule, the initial E/P activations resulting from hippocampal reactivation are discarded after activation has spread on to the input areas, and the resulting input patterns are used as a starting point for learning in the same way as during learning

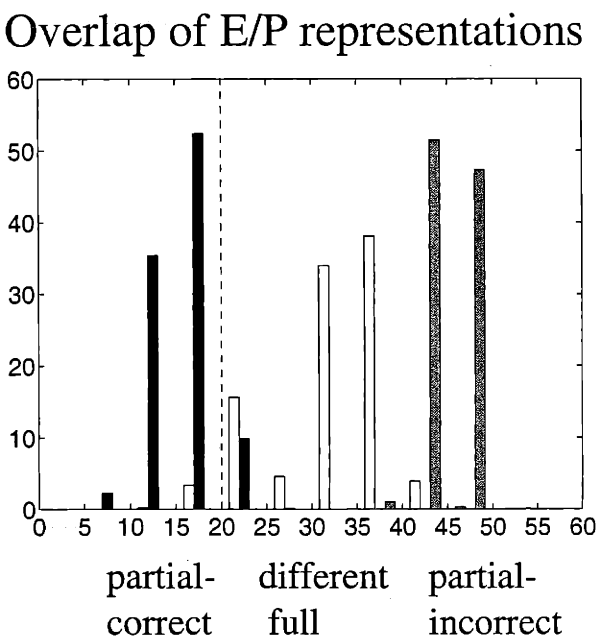


Figure 4.7: Threshold selection for hippocampal completion. The black bars represent the frequency histogram of the squared distances between the E/P representations of partial patterns and the correct full pattern; the gray bars are the same for the partial patterns and all the stored patterns other than the correct one; the white bars represent the distances between the E/P representations of the stored patterns and all other valid full patterns. Setting the threshold for completion of the E/P pattern to a stored pattern to 20 (dashed vertical line) makes it possible to correctly complete 90% of partial patterns while leaving the representations of full patterns mostly unchanged.

driven by feedforward activation (i.e., an E/P pattern is calculated first for the Hebbian part of the learning rule, followed by two updates yielding the activations used in the anti-Hebbian part). Throughout the consolidation period, blocks of 900 hippocampally-initiated learning events alternated with blocks of 100 cortically-initiated (general training) events.

After each phase of training, the performance of the network was tested in several ways, both in the presence and in the absence of the HC. The ability of the network to recall the stored memories was assessed by presenting parts of these patterns as inputs to see whether they can be completed. Partial patterns consisted of the correct pattern in two input areas (say, $\mathbf{x}_1^A \mathbf{x}_1^B$), and random activation in the third area ($\mathbf{x}_{\text{random}}^C$). Next, the activities in E/P were computed. In cases when the HC was present, the E/P pattern was then compared to each of the stored representations, and, if it was sufficiently close to one of them (using an arbitrary threshold of 20 on the squared distance), it was replaced by that stored pattern

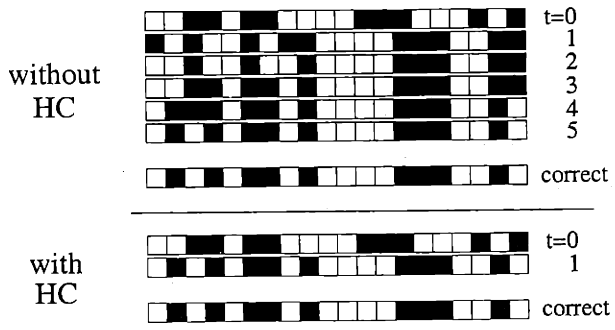


Figure 4.8: Convergence to the memorized pattern in area C (after consolidation) while a partial pattern is presented in areas A and B. The top part of the figure shows the first five iterations in the absence of the hippocampus and the correct pattern at the bottom; the bottom part shows the first iteration in the presence of the hippocampus. Only the first 20 units in area C are shown; the top row in both cases is the initial random pattern.

with a probability that was proportional to the strength of the hippocampal trace (recall that this strength decays exponentially with the age of the memory); otherwise, it was left unchanged. As shown in Figure 4.7, this allows the selection of the correct stored pattern, since the E/P representation of a partial pattern is always closer to the representation of the full pattern it was derived from than to the representation of any other full pattern (compare the distributions represented by the black and gray bars in the figure). It is also important that the representations of partial cues are typically closer to the representations of the corresponding full patterns than the representations of different valid input patterns are to each other (compare the black and the white distributions in the figure). This ensures that, provided that the completion threshold is set appropriately, partial patterns almost always get completed by the hippocampus, while the representations of complete patterns are almost always left unchanged. This gating of hippocampal participation based on similarities of currently processed and stored representations is intended to correspond to familiarity-based modulation of hippocampal processing.

Next, the cortical network was allowed to run for 20 iterations (with $\mathbf{x}_1^A \mathbf{x}_1^B$ clamped), each iteration consisting of an activity update in E/P followed by an update in area C. In most cases, activity in area C converged to a final pattern during these iterations. Some examples of the convergence to the correct pattern in area C (after consolidation of that pattern, with and without HC) are shown in Figure 4.8. Convergence is often instantaneous in the presence of the hippocampus, while it takes longer and is more gradual in its absence.

Figure 4.9 shows the main results from our model experiment with sequentially learned

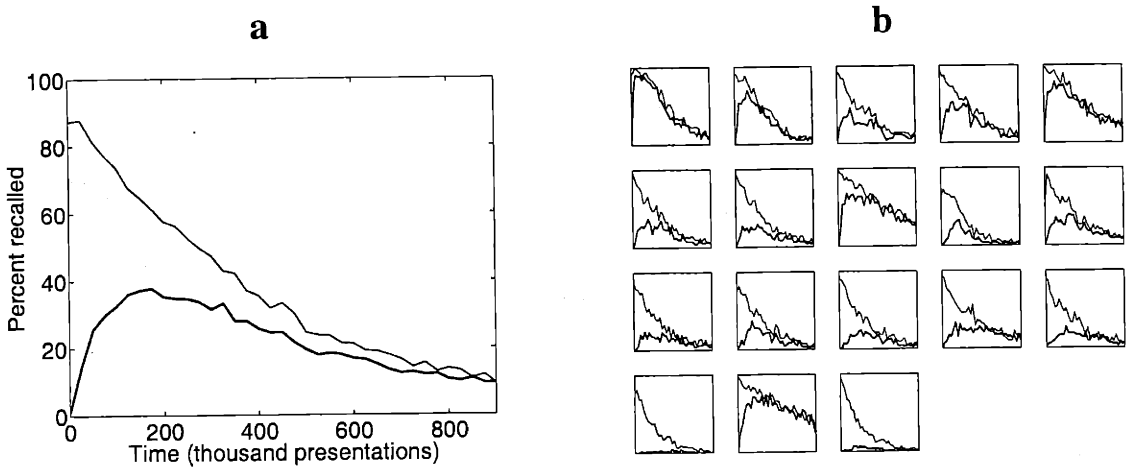


Figure 4.9: The consolidation of episodic memories. Recall performance on specific (episodic) patterns as a function of time between the initial presentation of the episodic pattern and testing (or, equivalently, time between training and lesion in hippocampals) in the simulations. The thin (upper) curve represents normal controls, and the thick (lower) curve is for the case when the hippocampal module is inactivated for testing. **(a)** shows an average over all episodic patterns used, aligned at the initial presentations (also the start of consolidation) of the particular patterns; **(b)** shows all the 18 individual patterns.

episodic patterns. Since hippocampal lesions are "reversible" in the model, we could test performance both in the presence and in the absence of the hippocampus at each time step. Scores obtained with the hippocampus correspond to the performance of normals, and scores without the hippocampus correspond to performance directly after the removal of the hippocampus on previously acquired material. The horizontal axis of the plots shows time between training and testing (or, equivalently, time between training and lesion for hippocampals) for a given pattern. On the vertical axis is the percentage of times when the pattern was successfully recalled from a partial cue using the recall procedure described above. Figure 4.9a shows the average over the performance curves for all episodic patterns, with the individual curves shifted so that the time of initial hippocampal storage (and the start of consolidation) corresponds to time 0 for each pattern.

The averaged time-performance curves for normals and hippocampals replicate many of the important characteristics of the corresponding experimental data. First of all, normals show the highest performance directly after training, and forgetting occurs gradually as time goes by. Hippocampals, on the other hand, perform at floor level if the hippocampus is removed directly after training, confirming the essential role of the hippocampus in early

recall. Furthermore, the performance of hippocampals becomes much better as more time intervenes between training and lesion, and the difference between hippocampals and normals becomes negligible for the most remote time periods tested, signalling the successful consolidation of these memories in our model.

On a computational level, these results can be understood as follows. Before memorizing the episodic patterns, or, equivalently, if the hippocampus is removed before any consolidation can occur, only semantic knowledge about the input patterns (embodied by the neocortical weights) is available. Consequently, given the partial pattern $\mathbf{x}_1^A\mathbf{x}_1^B$, all valid patterns in area C ($\mathbf{x}_1^C - \mathbf{x}_{20}^C$) are still equally likely, and are ultimately generated in approximately equal proportions by the network. This results in low recall performance for any particular input pattern, including $\mathbf{x}_1^A\mathbf{x}_1^B\mathbf{x}_1^C$. In addition, as we will shortly discuss in more detail, the network often does not even settle into one of the valid patterns within the first 20 iterations. This situation changes dramatically after the selected patterns are stored by the hippocampus. The E/P representations corresponding to the partial patterns are now recognized and completed by the hippocampus in most cases, which leads to good completion in the input areas due to the existing generative model relating activity in E/P to activity in A, B, and C.

The subsequent consolidation process alters the cortical weights so that the cortical network now represents a different probability distribution over the inputs; in particular, the probabilities corresponding to the memorized patterns are increased relative to all other patterns. This would result in some further improvement in the pattern completion performance of the full network (neocortex + hippocampus) if the probability of hippocampal pattern completion did not decay with time; as it is, the improved potential performance of the combined network is masked by the decreasing efficiency of hippocampal recall. More importantly, however, the changed cortical weights can support the recall of the stored patterns on their own, so that the removal of the hippocampus at this stage hardly affects the recall of consolidated patterns.

Figure 4.9b shows the individual time-performance curves for each of the 18 episodic patterns used (in the order that they were introduced during training). Although the above characteristics are apparent in most of the graphs, there are also substantial differences between the individual patterns, mainly in the magnitude of the consolidation effect. There also seems to be a tendency towards less consolidation for later patterns, which is probably due to the fact that the neocortical network was loaded close to its storage capacity for individual patterns in this simulation. We will return to the issue of storage capacity

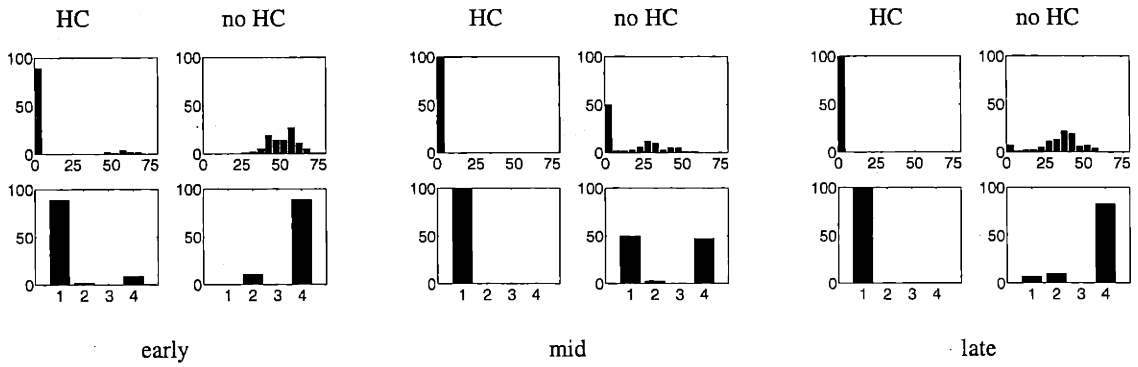


Figure 4.10: Classification of recalled patterns at different stages of consolidation. The three groups of plots represent three different times along the time axis of Figure 4.9, corresponding to time 0 (just after the initial presentation of the episodic pattern, before any consolidation), time 250,000 (around the peak of the recall curve for hippocampals, and time 750,000 (after a long period of consolidation and forgetting), respectively. The plots on the left in each group (marked HC) represent normals, and those on the right (marked "no HC") represent hippocampals. The top row of plots shows the frequency distribution of the distance between the pattern recalled (after 20 iterations) and the correct pattern; the bottom row shows the relative frequencies of the qualitatively different possible results of the recall process: (1) correct pattern; (2) other memorized pattern; (3) other valid pattern; (4) none of the above.

shortly.

Some of the points mentioned earlier were actually borne out of a more detailed analysis of the patterns recalled at different stages by the model (with or without the hippocampus). Final activation patterns in area C were classified as follows: correct pattern ($\mathbf{x}_1^C$), other memorized pattern (one of $\mathbf{x}_2^C - \mathbf{x}_{18}^C$), other (not memorized) valid pattern ($\mathbf{x}_{19}^C, \mathbf{x}_{20}^C$), and none of the above. Small errors in the recalled patterns (a squared distance of less than 5) were allowed in assessing the first three classes. The frequency distribution of errors (squared distance from the correct pattern) was also determined. For the purpose of this analysis, we ignored the decay in the probability of hippocampal completion of E/P patterns, because we were interested in the potential pattern completion ability of the model with or without the HC.

Figure 4.10 shows the class histograms and the full error distributions, both with and without the hippocampus, at three time points (corresponding to 0, 250, and 750, all in thousands of presentations after the start of consolidation, in Figure 4.9) along the performance curves for a particular episodic pattern: (a) before any consolidation, (b) after

substantial consolidation has occurred, around the peak of the amnesic performance curve, and (c) much later, after considerable forgetting has occurred. The first thing to note is that performance in cases when the hippocampus is active starts at around 90% after initial learning, increases further to 100% during consolidation, and stays at this high level subsequently. Therefore, in this model, forgetting in normals occurs mainly as a result of the decay of hippocampal memory. Second, the plots representing hippocampals indicate that most of the time when the network does not converge to the right pattern, it does not converge to any other valid pattern, either; in fact, it probably does not converge to any stable pattern at all. Third, consolidation shifts the frequency distribution of the distance of incorrect final patterns from the correct full pattern. The initial position of the peak is around 50, which corresponds to patterns uncorrelated with the correct pattern, but then the distribution shifts to the left and becomes flatter, indicating final patterns that are closer to the correct pattern. It would seem reasonable to assume that in these cases, although activity has not converged to the correct pattern in 20 iterations, it would eventually do so given more time. Indeed, we have tested this assumption by allowing more iterations, and found that the fraction of correct recall events increased as the cortical network was given more time to settle. However, even after 200 iterations, convergence seemed to be incomplete in some cases. Finally, as the hippocampal trace decays and the frequency of hippocampally-induced reactivations of this pattern decreases, the distribution of error magnitudes shifts back towards 50, accompanied by a decline in successful recall.

An important point about the consolidation process is that it should not impair the ability of the network to represent valid input patterns other than the ones memorized and consolidated, or to reconstruct these other input patterns from their E/P representations. It is relatively easy to come up with a learning algorithm which completes *any* input pattern to one of the patterns stored, but, among other things, such a network would find it difficult to memorize any additional input patterns subsequently. Our model, on the other hand, retains its ability to represent arbitrary combinations of the valid input patterns after consolidation (and, indeed, at all stages of training). We checked this by presenting input patterns different from the memorized episodes, and calculating the one-step reconstruction error after processing in E/P and, if applicable, the HC. We found that the general representational ability of the network was very good at all stages of training (after pre-training), both with and without the HC. Indeed, the average reconstruction error kept decreasing during episodic training (except for a small transient increase at the start of training; see Figure 4.11).

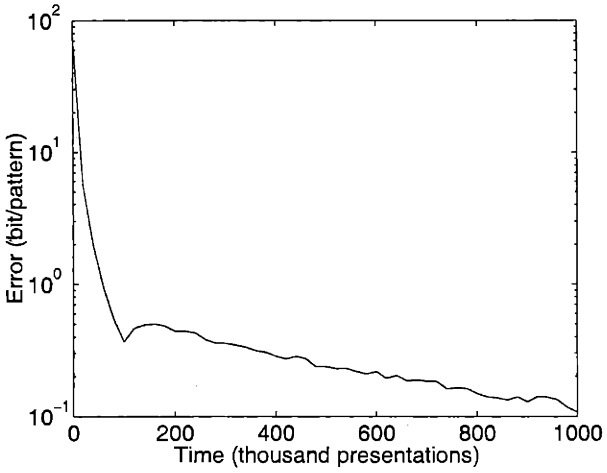


Figure 4.11: Quality of general representations during episodic training and consolidation. The plot shows the average one-step reconstruction error for all valid patterns (on a logarithmic scale) as training progresses; the first 100,000 presentations represent the general semantic training that is also shown in Figure 4.4, and the rest of the curve shows the effect of simultaneously storing and consolidating 8 specific (episodic) patterns. The corresponding recall performance on the episodic patterns is shown in Figure 4.12a.

In order to gain a better understanding of the computational requirements for memory consolidation, we systematically manipulated some of the components of our model to see how these changes affect performance. We looked at the effect of having a little bit less idealized (i.e., noisy) implementation of the hippocampus, requirements for the receptive fields of E/P neurons, and the relationship between the number of units in area E/P and the number of episodic patterns that can be stored by the neocortical network. Once more, since we were more interested in the model's potential for representation and recall than the exact speed at which learning occurs, and also to simplify the interpretation of the results, we set the decay rate for hippocampal memories to 0 (so that there is no decay) in the following simulations. We also started training on all episodic patterns simultaneously. In the baseline condition, there were 100 units in each area (including E/P), 10 valid patterns in each of the three input areas, and 8 patterns used for episodic training. The resulting time-performance plots are shown in Figure 4.12a, with the large plot showing an average over all episodic patterns, and the smaller plots some randomly selected individual patterns.

We first conducted some preliminary investigation of whether the exact implementation of the hippocampal component has a major effect on the performance of the full model. Of course, the ultimate goal is to have an explicit implementation of the hippocampus with a

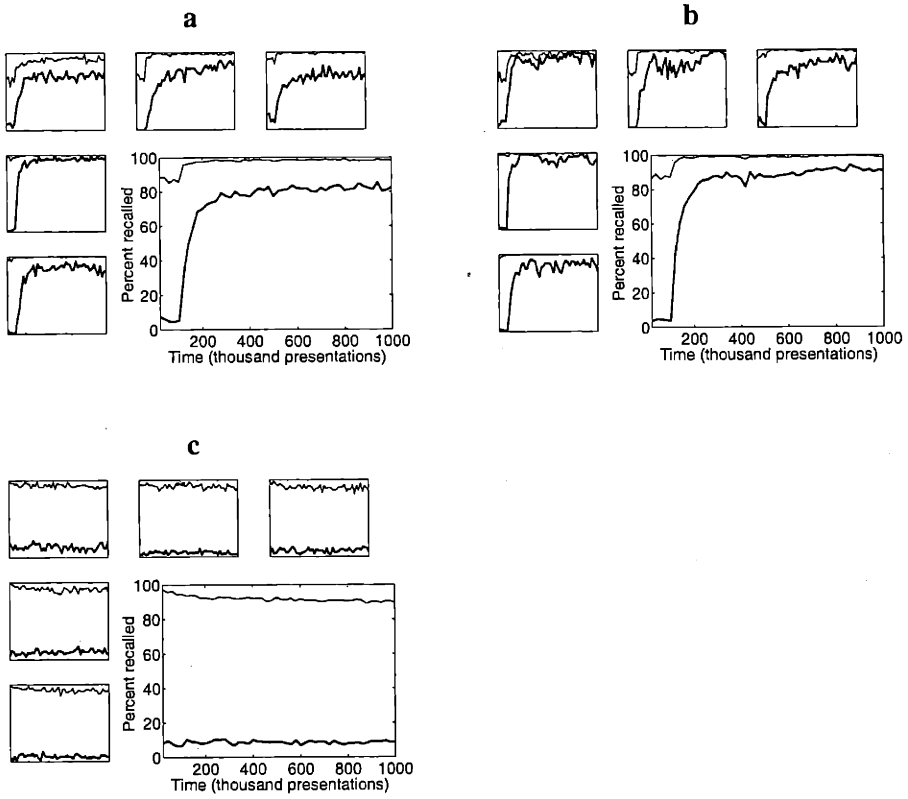


Figure 4.12: Effects of model details on episodic recall and consolidation. All the plots in this figure as well as the following one show percentage of correct recall of episodic patterns as a function of training time, both with and without the hippocampus (represented by the upper thin, and lower thick curves, respectively). In each part of these figures, the large plot displays an average over all episodic patterns, and the smaller plots represent randomly selected individual training patterns. All 8 episodic patterns were introduced simultaneously, after 100,000 presentations of randomly selected valid patterns (combinations of 10 valid patterns per input area), and there was no decay of memory traces in the hippocampus. Part **(a)** shows the baseline condition, where all characteristics of the simulation (with the exceptions noted above) were identical to those used to obtain Figure 4.9. Part **(b)** shows the case where hippocampal reconstructions of E/P patterns were assumed to be noisy (both during recall and during consolidation). In part **(c)**, each individual E/P unit was connected to only a single input area, whose identity varied across the units.

fair degree of biological realism, which would also force us to show how the hippocampus carries out all the operations that we have ascribed to it. However, this is a rather ambitious goal at this point since our current understanding of the internal workings of the hippocampus is still quite rudimentary. Therefore, as a useful first step, we looked at the effect of having a hippocampal component whose performance is not completely ideal. In particular, we tested a version of the model in which hippocampal completion and reinstatement of E/P patterns was slightly imperfect. This was achieved by adding noise (independent Gaussian with zero mean and standard deviation of 0.02) to the recalled probabilities in E/P. The resulting performance curves are shown in Figure 4.12b. Comparison with part a of the figure reveals very little difference between the noisy and the noiseless versions of the model; if anything, the noisy version might achieve slightly better consolidation.

We also looked at how critical the nature of the cortical semantic representation that develops during general training is for performance on episodic recall. As described earlier, the effective weight structure shows a substantial degree of convergence, with many E/P units sensitive to activity in more than one input area (and influencing activity in those same areas). It is intuitively clear that these multimodal units will be critical for cortical pattern completion that characterizes episodic recall, since they provide a link through which anatomically disparate areas can exert influence on each other. This prediction was fully confirmed by our simulations, during which we constrained the connections from different input areas to different parts of area E/P, so that no multi-modal units could develop. General (semantic) learning was basically unaffected by this manipulation since (at least in the simple version we used) it does not involve any inter-area correlations. Consequently, hippocampally-assisted pattern completion was perfectly normal since it only depends on a bidirectional mapping between E/P and the input areas (Figure 4.12c). On the other hand, no consolidation of these memories occurred because the separated representations of the different modalities in E/P could not support the learning of inter-area correlations which are crucial for episodic recall.

Finally, we investigated the storage capacity of the neocortical network by varying the number of episodic patterns to be stored and the number of units in area E/P (which is also proportional to the number of adjustable parameters, or weights in the model). The effects of varying both of these quantities are displayed in Figure 4.13 in a 2×2 format. There were either 10 valid patterns in each cortical area and 8 episodic patterns to be stored or 20 valid patterns and 15 episodic patterns. Along the other axis, the number of units in area E/P was either 50 or 100. The basic results can be summarized very easily: the higher the

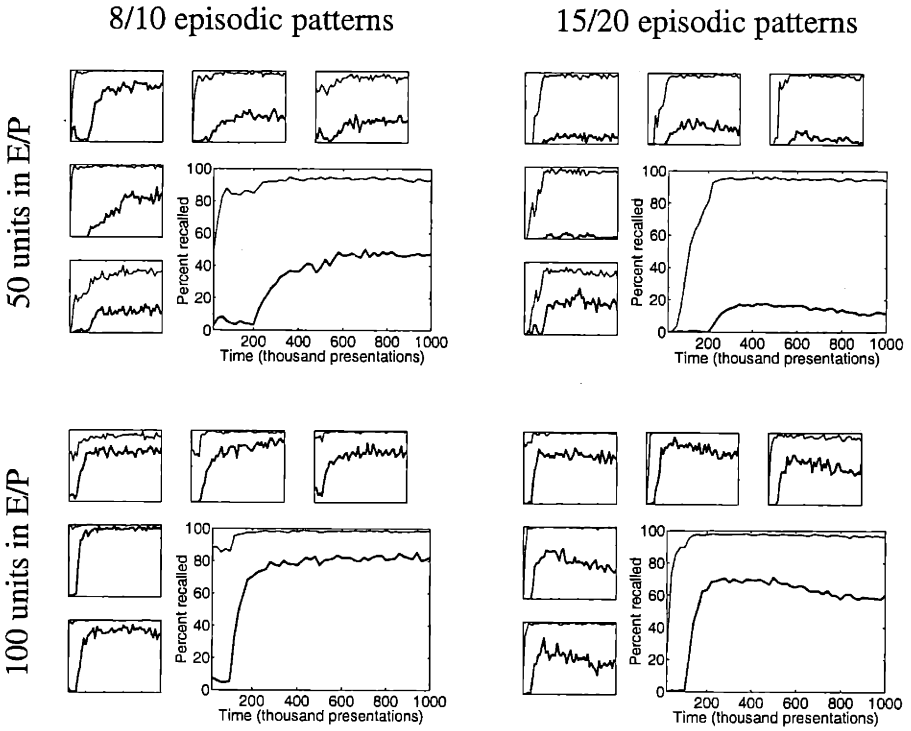


Figure 4.13: Episodic storage capacity. The variables plotted, the details of the plots, and simulation conditions are as described in the legend of (Figure 4.12, except that the number of hidden units in area E/P and the number of episodic patterns to be stored (along with the number of valid patterns per area) were varied independently as noted in the figure (e.g., "8/10 episodic patterns" means 8 episodic patterns and 10 valid patterns per input area).

representational capacity of the network (i.e., the number of E/P units and the number of weights), and the lower the number of actual patterns to be memorized, the higher the recall performance on these stored patterns. However, it is interesting to note that the network (with a given number of E/P units) does not seem to have a sharply defined storage capacity, i.e., the decrease in performance as the number of episodic patterns increases appears to be gradual rather than abrupt. This is in contrast with the behavior of other architectures like the Hopfield net which are characterized by a rapid deterioration of recall performance as their storage capacity is approached and exceeded. It is also important that, although its performance on episodic recall is rather poor, even the network with 50 E/P units can learn to represent combinations of 20 patterns per area, as evidenced by its good performance on 15 episodic patterns with the assistance of the hippocampus (direct measures like average

one-step reconstruction error on all valid inputs patterns support this conclusion). We may also note that, perhaps counter-intuitively, the smaller network takes substantially longer to learn, both in the semantic and the episodic phase of training, than the larger network. The reason is probably that it is more difficult to find an internal representation that supports good performance on all patterns using fewer E/P units.

4.2.4 Modeling repetition priming

One of the advantages of having an explicit model of neocortex which by itself has fairly sophisticated learning abilities is that we may hope to capture forms and expressions of learning which are believed to be relatively independent of the hippocampus. One example of such learning is the group of phenomena known as priming. As described in the Background chapter, priming involves the same kind of material that is also processed by declarative memory systems, and is defined as an increased facility for detecting or identifying stimuli as a result of prior exposure to these (or similar) stimuli, independent of the ability to recall earlier experience with them. Although priming is largely unaffected by medial temporal lobe lesions that result in amnesia, and has been suggested to be subserved by a different "memory system" (Squire, 1992), several models of priming have been proposed whose basic assumption is that the neural changes causing the priming effect happen in the same neocortical connections (outside the medial temporal lobe) as the ones which support the processing and storage of declarative information (McClelland and Rumelhart, 1985; Becker et al., 1997; Stark and McClelland, 2000).

The basic idea of these models can be described briefly as follows. The specific patterns stored in the network are stable states (attractors) of the neural dynamics to which activity can converge. When the network is in one of these states (which can happen during any kind of processing that involves these patterns), Hebbian learning between nodes that are active in that particular pattern leads to a strengthening (or "deepening") of the corresponding attractor. As a result of this strengthening, subsequent convergence to this pattern becomes faster, and is more likely to happen for starting conditions (cues) that could also lead to convergence to different patterns. These changes, i.e., faster processing and increased probability of recall, are exactly the psychological hallmarks of priming.

Since our neocortical model is quite similar, both in spirit and in actual implementation, to some of the above models (e.g., Becker et al., 1997), we expected our model to exhibit a similar capacity for priming. However, it turned out that the particular training protocol

used in the above simulations prevented the occurrence of a measurable priming effect. In particular, all simulations so far performed learning on large batches of different patterns, and the weight changes attributable to a single presentation of any particular pattern were rather small. This is also reflected in the large overall number of presentations required for substantial learning in these simulations. Since priming depends on weight changes occurring in response to a single stimulus presentation, we slightly modified our learning procedure to see if we can then capture this effect. The difference compared to earlier simulations was that a single presentation of an input pattern triggered multiple consecutive learning events in the cortical network, equivalent to about 100 presentations (albeit with a slightly smaller learning rate) in the old scheme. The net effect of this modification is a larger weight change in response to a single presentation, as well as a reduction in the noise (due to the randomness of activity updates during learning) in this change as a result of the averaging effect of multiple presentations.

Priming was assessed by measuring the effect on recall probabilities of previous presentations of a particular pattern. First, a general training phase on all valid input patterns was run, which was identical to the general training described earlier, except that the modified learning algorithm was used throughout. The results of this general training were also similar to those described earlier, except that convergence (as indicated by the decrease of the average one-step reconstruction error) was considerably faster (if measured by the number of pattern presentations). Next, the network was trained sequentially on another 100 pattern presentations; these 100 patterns were either all selected at random from all valid patterns (in the same way as during general training), or only 90 of them were selected randomly and the remaining 10 (none of which were among the last 5 to be presented) were constrained to be instances of one particular valid pattern (the primed pattern). Finally, cued recall of the primed pattern was tested in both the primed and the unprimed case (using the usual recall procedure), and the probability of recalling the primed pattern was compared between the two cases.

The result confirmed that our cortical network could exhibit a robust priming effect. Recall probabilities for any particular pattern (including the primed pattern) were relatively low in all cases; this was due to the fact that (as described earlier) after general training of the set of patterns that we used, all valid patterns in an area should be produced by the network in approximately equal proportions when asked to complete a partial pattern presented in the other areas. However, recall probabilities for the primed pattern were consistently higher in the primed case than in the unprimed case; in one representative case, the recall probability

rose to 6.2% from a baseline of 2.8%. Overall, although the actual recall probabilities varied considerably between patterns, the difference between the primed and the unprimed case was highly significant ($P < 10^{-5}$ in a t-test). It is also important to note that the network's general representational ability was unaffected by priming, since priming did not cause any significant change in the average one-step reconstruction error for all valid input patterns. Finally, it may be worth pointing out that priming is obviously independent of the presence or absence of the hippocampus in our model (although it does depend on the presence of area E/P).

4.3 Discussion

4.3.1 Main findings

The main feature that distinguishes our current model from all previous models of hippocampal-cortical interactions and memory consolidation is our model of the neocortical network, which is intended to capture some fundamental characteristics of neocortical information processing. In particular, our model is based on a large body of work on unsupervised learning, and the computational idea that a major goal of neocortical processing is to build a model of the statistical regularities present in the set of inputs that characterize the observations about the world. Besides providing us with a plausible computational foundation for our model, this framework also allows us to interpret the model's behavior in a theoretically sound, quantitative (probabilistic) manner. In addition, the connectionist implementation that we use ensures that basic principles of neocortical information processing like distributed representations and locality of learning are adhered to. Critically, our model also attempts to capture the large-scale anatomical structure of neocortex by assuming a hierarchy of areas with restricted inter-area connectivity. Direct interaction with the hippocampal component of our model is also restricted to a single neocortical area at the top of the hierarchy.

One of the advantages of having such a well-understood and generic model of neocortex is that, at least in principle, it becomes possible to test the consequences of different assumptions about the role of the hippocampus in learning and the nature of its interaction with neocortex. So far, we have examined one such possibility, which is in many ways a natural extension of earlier models and qualitative theories of hippocampally-dependent memory consolidation. In particular, with reference to the classification outlined in the Introduction

to this chapter of the possible ways in which the hippocampus might be involved in the early recall and consolidation of memories, our model can be characterized as follows. The hippocampus is directly involved in the early recall of memories for specific information by providing initial storage (in the form of a compressed representation) for information that uniquely identifies the episode, and (using this stored information) by initially assisting adjacent areas in medial temporal neocortex in completing the representation of the episode based on representations of retrieval cues. The hippocampus itself does not perform the anatomical binding of disparate cortical areas in our model; this role is played instead by medial temporal neocortex (area E/P). During consolidation, the role of the hippocampus is again specific; in particular, it reinstates the representation of the memory in E/P, which in turn reinstates the corresponding input representation, which is then used essentially as an additional training example by the neocortical network.

By implementing this scheme in a network simulation, we have been able to demonstrate the feasibility of hippocampally-dependent memory consolidation in a much more general and more realistic setting than any of the earlier models. In particular, we have shown that hippocampally-assisted fast learning and consolidation are feasible in a hierarchical model of neocortex, without the need for establishing direct connections between areas at the lower levels of the hierarchy, and without the need for the hippocampus to directly communicate with any but the topmost areas.

Our model also clarifies the relationship between memory (and representations) for general (semantic) and specific (episodic) information. In our model, both kinds of memories are eventually stored in the same neocortical connections and retrieved by the same neocortical processes. Although our models of episodic and semantic memory are certainly rudimentary, especially with respect to recall processes, and there are experimental data indicating that there might exist brain areas that are selectively involved in the recall of either episodic or semantic memory, our results point towards the possibility that these two kinds of declarative memory may share at least part of their underlying substrate.

Despite these similarities, there are also important differences in the ways semantic and episodic information are processed in the model, and their relationship is further complicated by an intriguing set of interdependencies. In particular, both early recall and consolidation of episodic memories, besides being hippocampally-dependent, also rely crucially on an existing mapping between representations at the highest and lowest levels of the cortical hierarchy, which in our case is just another manifestation of the cortical generative model that can also be viewed as the instantiation of semantic memory. In other words, episodic

memory may be restricted to material that can be semantically encoded. This dependency may also explain the psychological finding that new specific information (such as facts) is much easier to remember when it conforms to familiar structures.

In our model, the hippocampus is not essential for the acquisition of general semantic information. Such memories can be established via pure neocortical learning, provided that the network receives repeated exposure (in the form of representative examples) to the relevant information. On the other hand, episodic learning provides another, often more efficient way of incorporating new experience into the existing knowledge representation. Specific examples of a general rule may be stored by the hippocampus following a single presentation, and subsequent reprocessing (recall and consolidation) of these examples can lead to their gradual incorporation into neocortical representations, which also involves the neocortical extraction of generalities from the individual examples as before. Even before consolidation has been completed, episodic recall of these individual examples can help work out possible answers to general queries.

4.3.2 Comparison with other models and theories

The organization of memory

How does this view of the organization of declarative memory relate to other accounts in the literature? First of all, despite similarities between our model and that of Alvarez and Squire (1994) (which will be described shortly), our view of the relationship between episodic and semantic memory (and the hippocampal involvement in these) is clearly distinct from that of Squire and colleagues (Squire, 1992; Squire and Zola, 1998). They suggest that episodic and semantic memory depend similarly on the hippocampus and other medial temporal lobe structures (although episodic memory may additionally depend on the frontal lobes), and that semantic memories are always abstracted from episodic memories. We agree that information is always presented initially as part of an episode, but this information can influence semantic representations directly without the episode ever being stored in its entirety (although the final impact might be larger if an episodic trace is also formed). Our view also emphasizes the crucial dependence of episodic memory on existing semantic representations as described above. This emphasis is partly shared by the Serial Parallel Independent (SPI) model of Tulving and colleagues (Tulving, 1995; Tulving and Markowitsch, 1998), which also regards the semantic system as a gateway to episodic memory. However, in the SPI model, episodic retrieval is independent of the semantic system,

and the model does not include ways in which episodic memories may influence semantic representations (e.g., through reinstatement and consolidation). In addition, Tulving and Markowitsch (1998) view episodic memory as quite distinct from other types of declarative memories (although they are also proposed to have common features), with defining features such as a special kind of ("autonoetic") conscious awareness of "mental time travel", which essentially restrict episodic memory to a specific experience in humans. While we acknowledge that such mental experiences may accompany the retrieval of specific episodes, we do not see this as necessarily relevant for understanding the processes that underlie the storage and recall of declarative memories. In fact, we do not see episodic memories as special in this respect; rather, we see these detailed memories of specific experiences as representing one end of a virtually continuous spectrum of different types of declarative memories with different degrees of specificity and detail. At the other end of the scale we have general semantic knowledge, which pertains to a large number of different situations, and whose exact source often cannot be determined. These two extremes may be seen as corresponding to the two kinds of memories ("episodic" vs. "semantic") examined in our modeling work. However, there are also other types of declarative memories that are in some ways intermediate between these; for example, memories of distant events often lack the vivid detail of recent episodic memories, and may be merged with memories of other similar episodes; personal as well as general facts have even less specificity and detail; and some explicit linguistic information (like word meaning) is typically acquired in a large number of experiences. We do not conceive of any of these types of memories as being processed fundamentally differently by the memory system that is responsible for declarative memory; however, as shown earlier through comparing episodic memories with general semantic knowledge, the differences between types of memories that are very different in terms of detail and specificity can be considered to be qualitative.

Models of memory consolidation

As mentioned in the Introduction to this chapter, there exist a large number of (sometimes quite vague) theoretical proposals about how hippocampally-dependent consolidation might work, but relatively few explicit implementations of these ideas have been put forward. There are only two attempts I am aware of that explicitly implement and quantitatively test a theory of memory consolidation. One of these is the model of Alvarez and Squire (1994). This model consists of two "neocortical" areas (of 8 units each) connected directly through slow-changing connections, and a "medial temporal lobe" (MTL) area (of 4 units)

which is connected to the neocortical areas via fast-changing connections. Unit activities are updated using leaky integration of inputs (weighted by connection strengths), followed by a winner-take-all operation within competitive groups of 4 units (which simulates the effect of inhibition). Learning in all connections follows a covariance-based (modified Hebbian) rule, complemented by exponential forgetting (with different rates of modification for the two different types of connections). The task of the network is to reconstruct two, completely orthogonal (non-overlapping binary) patterns from partial cues (i.e., the correct input to one of the neocortical areas). Training consists of presenting the full patterns to the network twice, and letting activity cycle in the network while all connections are modified. Consolidation involves random activation of the MTL, leading to the reinstatement of the learned patterns in neocortex, which allows the gradual development of cortico-cortical associations mediated by the direct intracortical connections. As a function of the length of the consolidation period, the performance of the full network could be characterized by a normal forgetting curve, while the network in which the MTL was rendered inactive before testing showed the characteristic inverted U-shaped curve of temporally graded retrograde amnesia.

Although the model of Alvarez and Squire (1994) provides some evidence for the general feasibility of the consolidation idea, the authors themselves acknowledge that the model has some rather serious limitations, largely due to its simplicity, which also makes it hard to test comprehensively. The account also appears somewhat arbitrary in some aspects; in particular, within the computational domain investigated in the paper (i.e., the storage and recall of orthogonal, random patterns), there is no need for the intracortical connections to change slowly; storage would work just as well, and without any need for consolidation, if these connections were allowed to change as rapidly as the MTL connections. The need for slow cortical learning (and consolidation) appears only when the set of inputs has between-item correlations, and the cortical network tries to build representations that exploit this correlational structure. Furthermore (a feature that is also shared with some qualitative models such as Murre (1997)), the model requires some way of establishing and/or strengthening functional connections between neurons in disparate areas of neocortex (representing different aspects of the same episode) which would not normally be expected to enjoy substantial reciprocal anatomical connections. In addition, the hippocampus is directly connected to all neocortical areas; more importantly, this limitation would be difficult to overcome within the framework of Alvarez and Squire (1994) due to the absence of a general (semantic) neocortical model that would allow a generic mapping between different areas in neocortex.

The neocortical model of McClelland et al. (1995) in principle allows the processing of both general and specific information (indeed, it is used for both purposes in different examples), but the relationship between these is left unexplored. The hippocampus in this model has direct access to both input and output areas (although not to intervening "hidden" areas), and its role in consolidation is to reinstate training examples consisting of an input and the corresponding correct output, which are then used to train the neocortical network using the error back-propagation algorithm.

One potentially serious problem with the model of McClelland et al. (1995) is that interleaving new learning with rehearsal of old information requires some way of reinstating representative examples of all the relevant stored information. We cannot in general expect the outside world to continue providing examples, and relying on the hippocampus to retain and keep replaying for neocortex everything we ever experienced is bound to cause serious capacity problems (besides being inconsistent with data on amnesics). A possible solution to this problem (termed "pseüdorehearsal") has been suggested by Robins (1995, 1996). The idea is that the information stored in a distributed network may be characterized by the way that the network responds to each one of a sufficiently large set of random input patterns. Therefore, the stored information can be preserved by noting down these arbitrary patterns and the corresponding network responses (collectively called "pseudoitems"), and using these pseudoitems along with new training examples in subsequent training.

A dual-system model of memory based on the pseudorehearsal idea has been proposed by (French, 1997). The model, which is referred to as a "pseudo-recurrent network", consists of two subnetworks which interact with each other via pseudoitems. In particular, one part of the network, called the early-processing subnetwork, learns new items interleaved with pseudoitems generated by the other subnetwork (called the final-storage subnetwork). Later, the representation learned by the early-processing subnetwork is transferred to the final-storage subnetwork either by directly copying weights or (less efficiently, but with a higher degree of biological plausibility) by training on pseudoitems generated by the early-processing subnetwork. This architecture is claimed to alleviate the problem of catastrophic interference; however, the correspondence with the brain, and particularly with the hippocampal system, is not entirely obvious.

Finally, the TraceLink model developed by Murre (1997) shares many of the features of our hippocampal-cortical model. It consists of three modules: a trace system (roughly corresponding to the neocortical input areas in our model), a link system (roughly corresponding to the medial temporal lobes, including the hippocampus), and a modulatory system (which

comprises various subcortical "plasticity-control centers"). There are direct connections between units in the trace system and the link system, which can change rapidly, and sparse, slow-changing connections between the trace units. The modulatory system controls the plasticity of both kinds of connections, and can be activated either "externally" by arousal and attention, or "internally" based on the novelty of stimuli, measured by the level of competition in the link system in response to a pattern presented in the trace system. It is suggested that initial storage of patterns relies on both the trace system and the link system, and repeated exposure to stimuli later leads to the development of new functional circuits in the trace system which can then support recall without the link system. However, no detailed suggestion is made as to how the consolidation process itself might operate.

4.3.3 Outstanding issues

The work described in this chapter represents our initial effort to understand the mechanisms underlying long-term explicit memory through detailed quantitative modeling. However, the work as presented here is incomplete in several important ways. Some issues that are closely related to the ones considered here have been left unexplored; to some of the other questions that we do ask, the answers provided here are to some extent inadequate. In some cases, this is due to the fact that some aspects of the implementation of the model are rather over-simplified and clearly need to be extended to be able to capture a wider range of phenomena. However, there also exists a set of (mostly recent) experimental findings, as well as some results from our modeling study itself, that may potentially require a more fundamental reconsideration of the premises of the current model.

Extensions of the model

Let us start with some desirable extensions of the current model. First, it would be important to show that the neocortical network can be made more realistic, e.g., by expanding it to have more than two levels. An even closer match to gross cortical anatomy would also require us to allow some direct connections between areas that are either on the same level of the hierarchy or separated by one or more levels. The within-area connections (underlying the local attractors in the model) should also be implemented and treated in a common framework with long-range connections. All of these extensions are in principle straightforward while still remaining in the domain of Boltzmann machines; on the other hand, both learning and inference would have to be redefined if the model was no longer

"restricted", and the technical difficulties involved may be substantial. Ultimately, it is also desirable to find an implementation of the double dynamics of the neocortical network (neural activities and synaptic plasticity) with a higher degree of biological realism.

Our current model does not include an explicit implementation of the hippocampus, treating it as a "black box" with a set of properties that are necessary for the proper functioning of the full hippocampal-neocortical model. Of course, these hippocampal properties were not chosen in an arbitrary way, only to make the predictions of the model be consistent with a narrow set of observations on the behavioral level. Instead, our assumptions about what the hippocampus may be capable of doing are based on a wide array of experimental data on the function of the hippocampus (as reviewed in the Background chapter) as well as a long tradition of modelling work on the hippocampus. In particular, the ideas that the hippocampus is capable of fast learning of arbitrary patterns and of pattern completion based on these stored patterns have been explored in detail and are believed by many to be defining properties of hippocampal function. A number of modeling studies have shown how these functions may be carried out by the hippocampal network (O'Reilly and McClelland, 1994; Treves and Rolls, 1994; Rolls, 1996; Hasselmo et al., 1996; Hasselmo and Wyble, 1997). Nevertheless, no "hippocampus module" has been described that could simply replace the black box in our model. In particular, the completion of neocortical paterns by the hippocampus has not been modeled in detail, and the way in which the novelty of the input pattern determines the nature of hippocampal-cortical interaction also needs to be addressed (the work of Hasselmo and colleagues should provide a good starting point). In addition, forgetting within the hippocampus has not been modeled on the neural level, whether it happens by decay or by interference from new patterns. Therefore, it would be crucial for a full proof of the plausibility of our model to develop an appropriate implementation of the hippocampus.

For now, although a full implementation is beyond our reach in the short term, we can still deduce some of the basic properties such a model would need to have. For example, for the hippocampus to be able to support pattern completion in E/P after a single input presentation, a new attractor would need to be established in the hippocampus, which can be activated by appropriate partial patterns in E/P, and can be used to activate the full E/P pattern through backprojections. The formation of the new hippocampal attractor certainly requires fast learning in the connections between the neurons taking part in the attractor state. However, the connections forming the inputs to and the outputs from the hippocampus may also need to learn rapidly; in particular, if the hippocampal representation is not

a simple function of the entorhinal representation (and the place cell data on remapping certainly suggest that it might not be), then new associations between hippocampus and entorhinal cortex will have to be learned for every new stored pattern, and this would need to happen concurrently with intrahippocampal learning for hippocampally-assisted pattern completion in neocortex to work.

The kinds of relationships represented in semantic memory are certainly much richer than what we have explored in our model so far. In particular, our current semantic representation only defines what stimuli and combinations of stimuli are possible in the world, and how likely these are. In fact, this is very useful information and is able to support a surprisingly wide range of computations; however, it is unlikely to be sufficient to support the full range of declarative knowledge that we described earlier. In addition, so far we have not even exploited the full range of possibilities with the current type of semantic representation. Specifically, all valid patterns in a given input area were considered to be equally likely, and, more fundamentally, there were no correlations between inputs to different areas (except when we modelled the consolidation of episodic patterns). Allowing such correlations would substantially enhance the representational ability of the cortical network; however, the range of computations this scheme could support, as well as how it would affect interactions with episodic memory and consolidation will need to be explored.

Familiarity-based recognition

So far, the performance of our model was always tested using a single paradigm, namely, cued recall. However, a lot of experiments measure performance using different paradigms, many of which involve presenting stimuli in exactly the same form as they had been experienced before, and testing whether they are recognized. As described in the Background chapter, it is quite likely that there are at least two routes to recognition; something can either be recognized by explicitly recalling an episode of earlier experience with it, or simply by judging the stimulus to be familiar despite not being able to recall any particular context in which the stimulus had been encountered. Controlled lesion studies indicate that while the hippocampus itself may be critical for recall, it may not be crucial for familiarity-based recognition, which may instead depend on the integrity of perirhinal cortex. A strong hint about the existence within perirhinal cortex of circuits specialized for the processing of familiarity information comes from *in vivo* electrophysiology; in particular, a large fraction of neurons in perirhinal cortex has been described as responding differentially to novel and familiar visual stimuli.

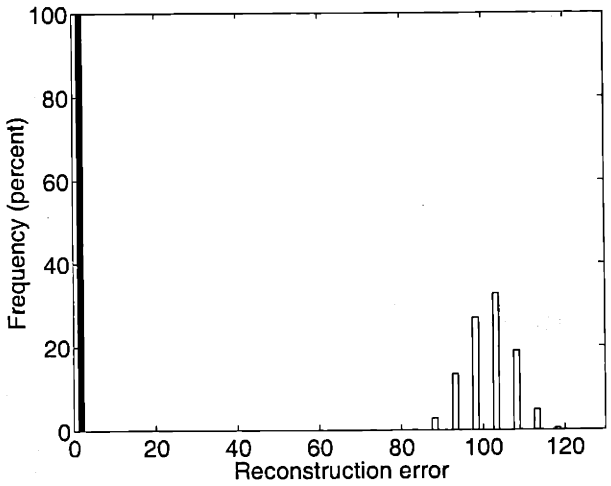


Figure 4.14: Novelty detection by the semantically trained neocortical network. The plot shows the distribution of one-step reconstruction errors for two classes of patterns: the black histogram (just a single bar at this scale) is for patterns randomly selected from the same set of patterns on which the network had been trained (valid, or familiar patterns); the white histogram is for random input patterns (invalid, or novel patterns).

Thus, it is plausible that specialized circuits are responsible for the assessment of familiarity in the brain, and it is not clear how much our generic cortical model can be expected to be able to mimic this ability. It should certainly be possible to supplement our current model with a special-purpose component for familiarity detection; even existing models like Bogacz et al. (2001) could easily be adapted since they only require as input summary representations of the different stimuli, which is provided by the E/P representation in our cortical network.

However, a closer look at our model reveals that, even in its current form, it processes novel and highly familiar input patterns in a different way, and therefore allows some form of discrimination between these. The simplest way to see this is to look at, after general semantic training, the network's ability to represent two different classes of patterns: patterns that were used for general training, and unrelated random patterns. Figure 4.14, which shows the distributions of one-step reconstruction errors for these two classes of patterns, reveals a fundamental difference in the way the network processes novel and familiar patterns. Whereas the reconstruction error for familiar patterns is almost always below 1, the error for novel patterns has a mean of around 100 with a standard deviation of 5.5, and therefore the two classes can easily be discriminated with high certainty based on this measure.

Of course, this fundamental difference is also reflected in any other measure that reflects how well the network is able to represent a given pattern; in particular, the initial rate of change in the E/P representation or the time required for convergence in E/P as the network processes the input pattern could be used instead. In fact, one of these latter measures may be better than reconstruction error in the input areas since they can be measured locally in higher-level areas. However, independent of the way familiarity is measured in the model, the above classes of patterns are not particularly realistic examples of novel and familiar patterns. In particular, the "novel" patterns in the above example are not just unfamiliar, but also "impossible" in the sense that their corresponding activity patterns have a completely different statistical structure from all previously encountered patterns. Besides, the difference in the number of previous experiences between the two classes is quite large; ideally, a single experience should be sufficient to render a particular input pattern familiar. We may conclude from this analysis that whereas some elementary form of familiarity detection is certainly possible within the framework of our current model, extensions including some kind of specialized module may be required if we want to model familiarity-based recognition in a more realistic manner.

Challenges to consolidation theory

Finally, we may ask the questions: Do our results overall support the idea of memory consolidation? Are there any fundamental holes in our computational understanding? Are there any experimental data that are conspicuously at odds with the model? As we argued in the modeling sections and earlier in the Discussion, our model proves the general feasibility of many aspects of consolidation theory, and provides a good qualitative fit to the data from a large number of experiments. Nevertheless, there are a few important issues where the agreement between the available experimental evidence and our model (and, indeed, other formulations of consolidation theory) is not so clear.

First of all, there is the issue of the long-term stability of neocortical memory traces. Qualitative theories of consolidation assume that once a neocortical memory becomes fully established (consolidated), it is potentially permanent and immune to disruption from normal cortical processing (although it may be subject to some form of forgetting) even if the hippocampus is subsequently damaged. Data from amnesic patients apparently support this assumption, since patients with hippocampal damage can typically retain indefinitely those memories that were not lost immediately with the lesion, independent of any kind of mental activity that they engage in.

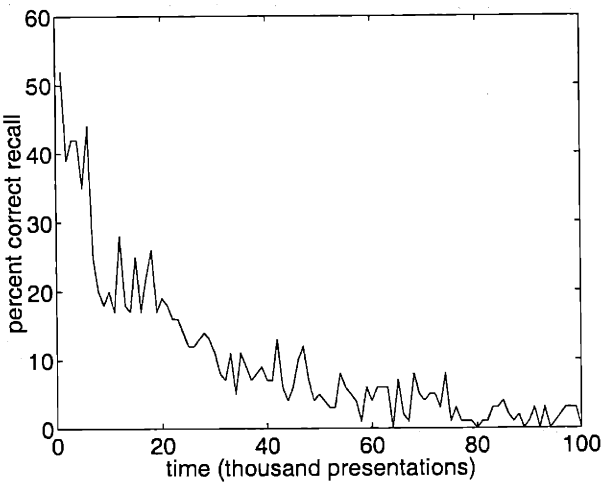


Figure 4.15: Neocortical forgetting of specific patterns. This plot shows how recall performance on a specific "consolidated" pattern (one of the patterns in Figure 4.9b) declines with further general training if hippocampally-initiated rehearsal stops. The initial state of the network for this figure is taken from the simulation of Figure 4.9, 250,000 presentations after the consolidation of this particular pattern started (near peak performance on this pattern without the hippocampus).

This permanence of neocortical traces in the face of ongoing processing (even in the absence of the hippocampus) is not characteristic of our current model. In our description of episodic and semantic memory storage in the isolated neocortical network, we pointed out that although specific (episodic) patterns could be stored and retrieved by this network, recall performance on these patterns rapidly deteriorated when general (semantic) training replaced the rehearsal of the episodic patterns. In our full model, hippocampal reinstantiation of the episodes in neocortex prevents this rapid decay. However, if the hippocampus is lesioned, even seemingly consolidated memories (where the removal of the hippocampus has very little immediate effect) become susceptible to disruption by ongoing general processing (or rather, general learning) in neocortex. This can be seen quite clearly in Figure 4.15, which shows the rapid decay of recall performance on one of the "consolidated" patterns of Figure 4.9b when the hippocampus is switched off in the model, and the network is subjected to general training on all valid input patterns. The episode is forgotten very quickly (although a bit more slowly than after it was trained in one shot; cf. Fig. 4.6c); indeed, the forgetting rate is much higher than that of normals in Fig. 4.9, indicating that the decay rate of hippocampal traces determines the normal forgetting rate in our model.

This result appears to be at odds with experimental data on amnesics, where no evidence

of vastly accelerated loss of old memories after hippocampal damage has been reported. There are several possible reasons for this discrepancy. First, it might be the case that neocortical memories of specific patterns are indeed rapidly forgotten after hippocampal lesions, and therefore they always remain dependent on the integrity of the hippocampus, as proposed for episodic memories by Nadel et al. (2000). However, this proposal is heavily debated; in addition, there appear to be other kinds of memories (like factual memories) that may require the recall of specific patterns across cortical areas, and whose consolidation appears to be well supported by experimental evidence. Second, it might be possible that neocortical learning somehow stops following hippocampal damage, so that it does not interfere with previously stored memories. However, evidence for various kinds of cortical learning in amnesics, including their ability to acquire new declarative memories with much repetition, counts against this possibility (although such learning abilities have been attributed to residual hippocampal function). Finally, the most likely explanation may be that neocortical memories do become permanent and to some extent resistant to subsequent learning, and some additional mechanisms need to be incorporated in our model to account for this. This might be accomplished by allowing neocortical synapses to "freeze", thus making them resistant to further activity-dependent changes, or perhaps by postulating some hippocampally-independent memory maintenance process in neocortex that actively preserves existing memory traces in the face of new learning. Clearly, these possibilities will need to be investigated, both experimentally and through further computational modelling.

There is also some recent experimental evidence that indicates that consolidated memories may in fact become vulnerable under some circumstances. Land et al. (2000) looked at the effect of hippocampal lesions after the reactivation of memories that appeared to be consolidated. They found that a simple reminder treatment made these old memories susceptible to amnesia, similarly to recently acquired memories. As an explanation, the authors suggest that perhaps the activity state of the memory rather than its age determines its vulnerability, and reactivated memories may need to be reconsolidated; however, this idea seems to be inconsistent with some data on memory consolidation, and with the fact that amnesics do not seem to lose memories when they recall them. Thus, it would be very important to understand these results better, and particularly to find out the exact circumstances in which memories require reconsolidation.

A second experiment by Land et al. (2000) also challenges the notion of memory consolidation. In this experiment, they showed that a reactivation treatment (a non-contingent footshock) could restore the memory of earlier training that appeared to be lost due to a

hippocampal lesion that followed the learning experience. The result shows that, at least in some circumstances, the hippocampus may initially be involved in the recall of memories which are stored elsewhere and which, using appropriate procedures, can also be recalled without the hippocampus. However, it is important to note that the acquisition of the task used by Land et al. (2000) is unaffected by hippocampal lesions, and the nature of the hippocampal involvement in the retention of the task in normals is not well understood.

Chapter 5

General discussion and conclusions

This final chapter considers the results presented in the thesis in a more general context. First, we discuss the significance of the results and the related arguments for theories of hippocampal function, both regarding what the hippocampus might be computing and how it might carry out the computation. Second, we will consider the general directions in which our results could be extended, highlighting areas which appear to be particularly promising.

5.1 Theories of hippocampal function

5.1.1 How does the hippocampus work?

Anatomical specialization

Our model of hippocampal place cells, which was described in Chapter 3, provides further support for a set of ideas which were proposed initially by several researchers in different contexts, but which have been gradually integrated in recent years to form a largely coherent view of how the hippocampus operates. This view assigns a specific computational role to the major subfields of the hippocampal formation, and is generally consistent with experimental data from a wide array of paradigms.

According to this theory, hippocampal processing can be thought of as happening in multiple consecutive stages. An important development of the last decade is the realization that direct input from areas other than the immediately preceding one (and particularly the

direct input to all hippocampal subfields from entorhinal cortex) is probably essential for the proper functioning of each area. In this context, it is important to note that very little is known about the neural representations in entorhinal cortex (EC), which is the major input and output area for the hippocampus, except that it seems likely to represent highly processed, multimodal information. This lack of information about the nature of the input considerably hampers further understanding of hippocampal information processing.

Dentate gyrus. The dentate gyrus (DG) is the first stage of hippocampal processing, and receives the majority of its inputs from EC. Since the DG also happens to be the area of the hippocampal formation (perhaps save for the subiculum) with the most complex internal circuitry, theoretical approaches to the DG have mostly concentrated on its computational role, and few attempts have been made to understand how it might implement these functions. This lack of detailed understanding, compounded by the corresponding lack of physiological data, is currently one of the major stumbling blocks in proving the feasibility of several theories of hippocampal function, including ours.

On the computational side, there is an almost surprising degree of agreement regarding the role of the dentate gyrus. Most theories require the DG to perform some kind of preprocessing of the EC input to make it more suitable for further processing by later stages of the hippocampus. This transformation is thought to involve what has been referred to as pattern separation or "orthogonalization", i.e., a process through which dentate representations become less similar than the corresponding entorhinal representations. Such orthogonalization of EC patterns plays an important role in both of the models described in this thesis. This role is rather direct in the place cell model. First, it assumes that dentate granule cells have tight spatial and directional tuning. Second, it assumes that at least some of these neurons can distinguish places where the visual component of the entorhinal representation is identical, but which can still be distinguished based on the path integration information that is also represented in EC. For the consolidation model, a critical requirement is that novel patterns can be stored in the hippocampus immediately without interference with or from the patterns that are already stored. However, fast storage of new patterns and preservation of existing memory representations are normally conflicting requirements (the stability-plasticity problem). One manipulation which is known to attenuate the effect of this conflict is the orthogonalization of the patterns to be stored (in our case, in CA3) either actively, or simply by making the patterns sparser, which is known to happen in the DG.

Area CA3. That area CA3 may function as an autoassociative memory is by no means a new idea. What is a more recent development is the idea that the attractors formed in the CA3 recurrent network do not have to be discrete points in the space of ensemble activities; instead, the attractors may also comprise a continuous set of states on a multi-dimensional manifold. The main computational contribution of our place cell model is to show how the traditional framework of learning and recall in attractor networks can be extended to continuous sets of inputs, outputs, and stored representations. Beyond showing the feasibility of this general scheme, our model demonstrates that this extended version of the attractor framework, which already proved to be a useful model of memory phenomena, can also capture the characteristics of the CA3 place cell representation in rodents with considerable accuracy. For further discussion regarding the function of area CA3, refer to Chapter 3.

Area CA1. No theories of comparably general appeal have been proposed for area CA1. The physiological correlates of behavior in CA1 are relatively unexciting since, at least in the spatial domain, the CA1 representation is essentially a copy of the CA3 place cell representation (although there are some hints that they may dissociate under some circumstances). The only hint from anatomy is that CA1 is the main output area of the hippocampus (projecting to a variety of targets), and it receives input from both CA3 and entorhinal cortex. Neither of the models in this thesis explicitly addresses CA1 so far, although such an extension (especially for the place cell model) would be highly desirable. On the other hand, our models are consistent with several of the existing proposals regarding the function of CA1. In particular, the proposal of Hasselmo according to which CA1 performs a comparison between its entorhinal and CA3 inputs in order to determine the familiarity of the current stimulus, and this signal is in turn used to alter the neuromodulatory activity of the septum, constitutes a natural extension of the current form of neuromodulatory control in our model.

Subiculum. Finally, it is worth noting that the subicular complex, which interacts in complicated ways with the hippocampus, surrounding neocortex, and various subcortical regions, and which has been implicated in various behaviors, has not received much theoretical attention, and the exact functional roles of its multiple subregions are currently largely unknown.

Multiple processing modes

One idea that figures prominently in both our place cell model and the model of hippocampal-neocortical interaction is that the hippocampus can operate in multiple, more or less distinct, processing modes. In particular, the place cell model assumes the existence of two modes, a "learning mode" and a "recall mode", which differ in terms of the efficacy of different types of connections in the hippocampus, and in the degree of plasticity of these connections. Most earlier models (either explicitly or implicitly) assumed a strict separation of these two modes. Our work has shown that, at least when the set of inputs is continuous, plasticity should be modulated in a graded rather than a binary fashion. And although we have so far retained two distinct modes for modulating the efficacies of connections (particularly for rendering the CA3 recurrent connections ineffective during new learning), this modulation too may very well be graded (and perhaps tied to the modulation of plasticity), thus blurring the distinction between the different modes. Such graded modulation might also be more consistent with physiological data, since no abrupt changes have been reported in recordings of place cell activity as the animal explores the environment.

The consolidation model attributes not two, but three different modes of processing to the hippocampus (see also Hasselmo, 1999; Redish, 1999, Ch. 8), each with a well-defined computational role. Two of these functions, namely, the storage of new information and recall based on partial cues, may be identified with the two processing modes explored in the spatial model. In addition, hippocampally-dependent consolidation in our model also requires a "replay" mode, which provides neocortex with further exposure to the (approximate) conditions characterizing the original learning episode, thereby allowing neocortex to integrate the experience into its complex representation of the world. Replay might take place during "off-line" periods (such as slow wave sleep). Of course, this third mode may also be present in the context of spatial processing (actually, most of the data on hippocampal reactivation during sleep comes from rodent place cells), and modeling these off-line phenomena would be a natural extension of the place cell model. However, detailed modeling of these phenomena would require much more attention to the temporal aspects of hippocampal dynamics, since the temporal patterns of activity during replay are likely to be autonomously generated within the hippocampus rather than input-driven. On the other hand, understanding this off-line processing should also further our understanding of awake processing, especially since the mechanisms underlying replay need not be completely distinct from those supporting the other two modes, particularly recall.

5.1.2 What is the hippocampus for?

In Chapter 2, at the beginning of this thesis, we argued that almost all the theories and most of the experiments investigating the functional role of the hippocampus have centered on one of three classes of functions: memory, spatial processing, and conjunctive coding. As we stated at the outset, one of the main goals of this thesis was to integrate these apparently quite distinct functions by finding links between them at various levels of description through careful, quantitative analysis of the underlying computations. Our results indicate that essentially the same set of mechanisms can support long-term declarative memory, the formation of spatial representations, and the creation of complex, conjunctive representations.

Our account accommodates the theories as follows. First, the learning and recall processing described above, which may collectively be referred to as autoassociative operations, combined with the consolidation operations occurring during replay, together support standard accounts of the role of the hippocampus in memorial functions. Second, the orthogonalizing action of the dentate gyrus, which creates dissimilar outputs from similar inputs representing complex stimuli, is exactly what is required to support conjunctive coding. This operation results in combined codes for complicated inputs, which are subsequently stored by the CA3 autoassociator, and can then influence cortical processing through recall and consolidation. Third, when applied to the sorts of sensory inputs available during spatial behavior, the combined action of dentate pattern separation and CA3 autoassociation leads to the creation of place cells with appropriate properties.

That we think all these functions of the hippocampus can be accommodated within a single general framework does not mean that dissociations between them cannot be observed. Different classes of functions (and even performance in different tasks within a given class) may rely on the basic underlying processes to a different extent, and these basic processes themselves require different anatomical regions and physiological processes. These differences should underlie the observed behavioral dissociations following various manipulations. The work described in this thesis strongly argues that theories seeking to understand neural phenomena should take full advantage of the entire scale of available experimental data from the anatomical to the behavioral level, in addition to applying computational constraints.

Thus we would argue that the conservation of the basic structure and general physiology of the hippocampus in different species actually reveals a corresponding conservation of the fundamental computations carried out by the hippocampus. However, the substantial

differences in neocortical structure, combined with the marked differences in sensory experience and behavioral patterns in different species (such as primates and rodents) mean that the hippocampi of different animals operate in considerably different neural environments. In particular, the inputs to the hippocampus may represent different things, or the same things in different ways, in different species, so that essentially the same processing within the hippocampus may result in rather different internal representations, and might even subserve different high-level computations. Similarly, the output from the hippocampus may be interpreted in different ways depending on the neural context.

5.2 Future plans

The research described in this thesis can be extended in many ways. Our model of place cells, although it encompasses a fairly broad class of experimental data, still fails to address some important issues. In particular, the temporal aspects of processing are largely ignored, both at the fast time scale of rhythmic activity and individual action potentials, and at the slow time scale of long-term experience-dependent changes in the place cell representation. Our work on memory consolidation represents an initial attempt at constructing a well-founded general theory of hippocampal-cortical interaction, and is thereby forced to be even less complete. The fact that there exists very little data that strongly constrains this theory, but there is a wealth of data that provide weaker, indirect constraints, means that proving the plausibility of the theory involves the modeling of a wide array of different phenomena.

Detailed analyses of outstanding issues for the place cell model and the model of memory consolidation have been provided in the Discussion sections of Chapter 3 and Chapter 4, respectively. In this section, I discuss some of the possible future directions that apply to both models. In this context, the most important priority is the construction of an improved general model of the hippocampus. There are two obvious ways of extending our existing model, which was described in Chapter 3. First, it would be important to expand the anatomical scope of the model by modeling explicitly areas of the hippocampal formation other than CA3. As we saw above, detailed understanding of the dentate gyrus would be particularly important, but, ultimately, other relevant areas such as CA1, entorhinal cortex, and the subiculum will need to be included. Second, the degree of biological realism should be increased by including important features of hippocampal processing that we have so far ignored, particularly in the temporal domain.

Extending our model to deal with the temporal aspects of representation and learning presents an exciting challenge since it involves a set of novel and currently heavily investigated phenomena that interact in complex ways. First, in order to enable our model to describe the temporal structure in the activity of place cells (such as the oscillations at the theta frequency), we need to go beyond steady state solutions of the dynamical equations (e.g., by letting the inhibitory time constant be finite, and by considering the oscillatory nature of the septal modulatory input). It may also be necessary to switch from the current rate-based model to a spike-based model, which would allow us to study the fine temporal structure of neuronal correlations (such as synchronization). In addition, a spike-based model would allow us to study the effects of spike-timing-dependent synaptic plasticity, which has not been done in the context of a recurrent network storing a meaningful representation. Incorporating oscillations, spiking, and spike-based plasticity would also make it possible to study the systematic relationships between spatial representation and temporal activity patterns, most notably the phase precession effect in place cells, and whether these spatio-temporal activity patterns may be a result of learning.

In addition, as we described above, a crucial issue that ties together detailed modeling of the hippocampus and more abstract, computational modeling of hippocampal-cortical interactions is the notion of multiple processing modes. Therefore, the understanding of these different modes and their regulation, including neuromodulation and sleep, on different levels ranging from the physiological to the computational, remains a basic priority for our future research.

Appendix A

A simple, analytically solvable place cell model

This Appendix describes a variation on the place cell model which was introduced in Chapter 3. Owing to a few simplifications compared to that model, this model becomes analytically tractable. As we will see, this approach offers some additional insight into the development of continuous attractors with different tuning properties.

A.1 Description of the model

Most aspects of the model are identical to those described in Chapter 3. In particular, the basic architecture of the model is the same as what is shown in Figure 3.1. The central part of the model is the CA3 network, which consists of a set of fully interconnected excitatory neurons and a single globally connected inhibitory neuron. The excitatory neurons receive external inputs from two sources: a fixed, neuron-specific input through the mossy fibers which depends on the location and heading of the animal, and another set of inputs from neurons in entorhinal cortex (EC) through the perforant path. The perforant path connections and the CA3 recurrent connections are assumed to be modifiable in an activity-dependent manner.

In the current model, we use a simplified treatment of neuromodulatory control of hippocampal dynamics and plasticity. In particular, we assume that the hippocampal network has

two modes of operation, and switching between the two occurs in a binary fashion. When the rat is in a familiar environment, no learning takes place in any of the connections, the relative efficacy of the mossy fiber inputs with respect to the perforant path connections and CA3 recurrent synapses decreases, and the intrinsic dynamics of the recurrent network dominates activity in CA3. This we will call "recall mode". On the other hand, when the rat first encounters a new environment, learning in both the perforant path inputs to CA3 and the recurrent connections is initiated, the recurrent connections are suppressed, inhibition in CA3 is reduced, and inputs through the mossy fiber connections dominate. This state of the network will be referred to as "learning mode". We seek to understand how the pattern of weights that is set up during learning can produce the patterns of activity of place cells seen subsequently.

The input representations (i.e., the tuning of cells in EC and of the mossy fiber input to CA3) are also similar but much simplified compared to the representations used in Chapter 3. In particular, instead of modeling the ways in which constraints provided by different perceptual cues might conspire to define the spatial (and directional) tuning of the cells, we start from EC cells and mossy fiber inputs which have pre-defined spatial and directional tuning. We also make the input tuning curves of different cells to be shifted and rotated versions of each other, i.e., all tuning curves have the same variance but different preferred values in both location and heading space. As a further simplification, the tuning width in EC and the mossy fibers are the same, and, at least in the simplest formulation, the tuning curves themselves are just scaled and shifted cosines, with wrap-around boundary conditions for a single spatial and a heading dimension. The net tuning of the k'th cell in EC becomes

$$z_k(x,\phi) = I_k^0(x,\phi) = d_0\left[1-\epsilon_0+\epsilon_0\cos(\phi-\phi_k)\right]\left[1-\eta_0+\eta_0\cos\frac{(x-x_k)\pi}{l}\right], \qquad \text{(A.1)}$$

where d_0 measures the maximum value of the input, x is the location and ϕ is the head direction of the animal, x_i and ϕ_i represent the preferred values of that cell, $2l$ is the diameter of the arena, and η_0 and ϵ_0 measure the degree of spatial and directional tuning of the inputs.

Now the only part of the model that remains to be defined is the double dynamics of neuronal activities and weight changes. Neuronal dynamics in the CA3 network is described

by essentially the same equations as in Chapter 3; in particular, in recall mode,

$$\tau \dot{u}_i = -u_i + \sum_j J_{ij} g_u(u_j) - h g_v(v) + I_i^{PP} \tag{A.2}$$

where u_i is the membrane potential of the i'th pyramidal cell, τ is the membrane time constant, J_{ij} is the strength of the connection from neuron j to neuron i, h is the efficacy of inhibition, and v is the membrane potential of our global inhibitory cell. g_u and g_v are the activation functions for pyramidal cells and the inhibitory neuron, respectively, so that, for instance, $r_i = g_u(u_i)$ is the firing rate of i'th pyramidal cell. We choose the activation functions to be threshold linear, i.e., for pyramidal neurons, $g_u(u_j) = \beta(u_j - \mu)\Theta(u_j - \mu)$, where Θ is the unit step function (zero for negative arguments and one for positive ones). μ stands for the threshold and β is the slope of the activation function above the threshold. Analogously, $g_v(v) = \gamma(v - \nu)\Theta(v - \nu)$ for the inhibitory neuron. I_i^{PP} represents the net perforant path input to this CA3 cell, and dot on the left-hand-side stands for differentiation with respect to time. The activity update equation for the inhibitory neuron is just

$$\tau' \dot{v} = -v + w \sum_j g_u(u_j) \tag{A.3}$$

where w represents the strength of the excitatory connection from any one pyramidal cell onto the inhibitory cell.

The above scheme becomes significantly simpler in learning mode, where the effect of inhibition is assumed to be negligible, and the recurrent connections are also switched off. In this case, the equation for pyramidal neurons reduces to the following:

$$\tau \dot{u}_i = -u_i + I_i^{MF} \tag{A.4}$$

where $I_i^{MF} = I_i^0$ is the input for this case, which is thought to be dominated by the mossy fibers.

We assume that both the perforant path and the recurrent weights change in a Hebbian manner in learning mode, that is, the weight change is proportional to both the presynaptic activity and the degree of postsynaptic depolarization. We also assume exponential weight decay in the absence of pre- or postsynaptic activity to prevent weights from growing indefinitely. This leads to the following update equations for perforant path weights W_{ik} and

recurrent weights J_{ij}:

$$\dot{W}_{ik} = -\kappa W_{ik} + u_i z_k \tag{A.5}$$

where κ is the rate of weight decay, and

$$\dot{J}_{ij} = -\kappa J_{ij} + u_i r_j \tag{A.6}$$

In the next section, I will derive expressions for the steady state values of these weights, and then go on by finding the steady state solution for the neuronal activities in recall mode.

A.2 Results

This section summarizes the main results of the model I described above. Since all the locally observable quantities in the model (activities and weights) are in principle time dependent (though the weights are assumed to be fixed in recall mode), a full description of the behavior of the system would involve determining the value of each of these microscopic variables at every time, as a function of the parameters and initial conditions. This is an enormous task, and the results can be extremely hard to interpret. However, since many of the above equations have a relatively simple format, we can try to find special kinds of solutions. In particular, all of the above dynamical equations have solutions (at least in a certain range of the parameters) which converge to fixed values of the variables. As discussed in more detail in Chapter 3, there are reasons to believe that this may provide a good approximation to the actual behavior of the network, especially if we are interested in the average neural activity on the time scale of about a second. We also expect synaptic weights to stop changing after the environment has been fully explored (although the actual way this happens may be more complicated than the simple stochastic convergence assumed here, and an example of this is described in Section 3.2). Therefore, we look for solutions of the dynamical equations that are constant in time. We will also have to examine the stability of these solutions, that is, if the variables have initial values that are close to the fixed point solution, whether they eventually reach these stationary values or they become more and more distinct from them.

Let us first consider the learning phase. The dynamical equations governing this mode (equations (A.4) to (A.6)) are particularly simple,so we can instantly recognize their stable solutions. The membrane potential of the pyramidal cells converges to $u_i = I_i^0$, with I_i^0

given by equation (A.1), which shows how the stable value of u_i depends on the current location and head direction of the animal. Assuming that changes in the input are much slower than the membrane time constant (a very reasonable assumption), we can use these values for u_i in the equations for the weight changes.

As the rat explores the environment, the modifiable synaptic connections change continuously. However, unless weight changes are much faster than the characteristic time required to efficiently sample the environment, the weights will eventually fluctuate around some stable values. These values can be found by making the expected weight change zero: $\langle \delta W_{ik} \rangle = 0$ in equation (A.5) and $\langle \delta J_{ij} \rangle = 0$ in equation (A.6). The averaging should be carried out over all possible values of x and ϕ. Here we assume that exploration is uniform in both space and angle within the environment (which, as we have seen in Section 3.2, is a simple, but not necessarily a particularly realistic assumption). This gives us equations for the stable values of W_{ik} and J_{ij}, respectively. The equation for J_{ij} is

$$J_{ij} = \frac{1}{\kappa} \langle u_i r_j \rangle \tag{A.7}$$

If we substitute in the stable values of u_i and $r_j = g_u(u_j)$ and do the averaging, we get an expression for J_{ij}. However, since the averaging is hard to do exactly in this case, we will use a linear approximation for the activation function, which leads to the following expression:

$$J_{ij} = a \left[\cos \frac{(x_i - x_j)\pi}{l} + e \right] [\cos(\phi_i - \phi_j) + b] \tag{A.8}$$

where a, e and b are constants. This expression has some of the attributes which have been suggested as desirable characteristics of the CA3 recurrent weight structure. In particular, we expect in general that these weights will only depend on the differences between the preferred locations and preferred head directions of the cells.

Repeating the same steps for the perforant path connections, we get an interesting result. It turns out that the perforant path input to CA3 becomes, as a result of learning:

$$I_i^{PP} = d\,[1 - \epsilon + \epsilon \cos(\phi - \phi_i)] \left[1 - \eta + \eta \cos \frac{(x - x_i)\pi}{l} \right] \tag{A.9}$$

that is, it has the same functional form as the input through the dentate gyrus. The intuition for this result is the following. At a particular moment in time, when the rat is

at a certain location and faces in a certain direction, a particular collection of cells in CA3, as well as a set of cells in entorhinal cortex fire. According to our Hebbian learning rule, the connections that are strengthened most are the ones that connect the neurons that fire the most, which will be ones that represent the current location and head direction in the two areas. Therefore, any CA3 neuron will come to receive the strongest connections from neurons that represent the same location and head direction, and it will be driven the most by perforant path inputs at that location and head direction.

Thus, in the learning phase, the recurrent connections within CA3 and the direct input connections from entorhinal cortex become established. But what happens if we allow the dynamics of the recall phase to run with these parameters? Does it converge at all? What kinds of solutions can we get? Is there a stable solution which is tuned is space but is insensitive to head direction (place field)? If so, what are the circumstances under which such a solution emerges? These are the questions we try to address when we analyze the behavior of the recall mode.

The dynamics of this mode is described by equations (A.2) and (A.3), with J_{ij} given by equation (A.8) and I_i^{PP} by equation (A.9). This is a complex set of nonlinear equations, and they seem to be easier to handle in the continuous limit, where the u_i's are replaced by a continuous function $u(x, \phi, t)$ which represents the membrane potential of units in CA3 which have preferred location x and preferred head direction ϕ, and J_{ij} is replaced by $J(x, x', \phi, \phi') = J(x - x', \phi - \phi')$. The last equality holds because, as we have seen, J_{ij} only depends on the differences in the two cells' preferred parameters. Also replacing the sums by the appropriate integrals, equations (A.2) and (A.3) now become

$$\begin{aligned}\tau \frac{\partial u(x,\phi,t)}{\partial t} = & - u(x,\phi,t) + \int d\phi' \int dx' J(x - x', \phi - \phi') g_u(u(x',\phi',t)) - \\ & - h g_v(v(t)) + I\end{aligned} \tag{A.10}$$

and

$$\tau' \frac{dv}{dt} = -v(t) + w \int d\phi' \int dx' g_u(u(x',\phi',t)) \tag{A.11}$$

With all the assumptions of our model, the steady state solutions for these equations are relatively easy to find. By making the temporal derivatives zero and taking advantage of the

linear nature of the activation function so that we can write $g_v(v) = \gamma(v - \nu)$ for sufficiently high levels of overall activity, we get the following equation for the steady state firing rates:

$$r(x,\phi) = g_u(u^*(x,\phi)) = g_u(\int_{-\pi}^{\pi} d\phi' \int_{-l}^{l} dx'[J(x-x',\phi-\phi') - h\gamma w]r(x',\phi') + h\gamma\nu + I(x,\phi)) \tag{A.12}$$

Now we can plug into this equation the expressions we derived for J and I. What we arrive at is an integral equation for $r(x,\phi)$. A closed form solution to this equation can actually be obtained, but that still contains some free parameters that should be determined from a set of nonlinear consistency equations which in general have no analytical solution. However, it may be possible to get all the solutions numerically for a given set of parameters. More promising perhaps is the possibility that we can actually check whether the kind of solution we are interested in exists. For instance, we may look for a place field-like solution which is tuned in space and is independent of ϕ. These assumptions simplify the treatment sufficiently so that we can now determine the parameters in the solution.

Probably the most exciting finding from this analysis is that an activity profile across CA3 cells which is tuned in space but insensitive to head direction can emerge spontaneously from the network dynamics, even in the absence of external inputs. This is called the marginal phase of the network by Ben-Yishai et al. (1995). Such an exact solution does not exist if the input is directionally tuned, which comes as no surprise; we would not expect the network to be able to suppress completely a feature in the input that is present continuously. However, assuming that the perforant path input is weaker than the recurrent connections, we do expect the real solution to be close to that in the marginal phase (the case with no external input). In this case, the location of the bump of activity in space, which is a free parameter in the marginal phase, should be determined by the tuning of the input.

The exact form of the solution in the marginal phase is actually very simple:

$$r(x,\phi) = \begin{cases} \beta abr_{10}\left[\cos\frac{x\pi}{l} - \cos\frac{x_c\pi}{l}\right] & \text{if } |x| < x_c \\ 0 & \text{otherwise} \end{cases} \tag{A.13}$$

where x_c is the width of the spatial tuning, and can be determined from the following

equation:

$$\frac{2x_c\pi}{l} - \sin\frac{2x_c\pi}{l} = \frac{1}{\beta abl} \tag{A.14}$$

and r_{10} is a constant (actually, the amplitude of the first purely spatial two dimensional Fourier component of $r(x,\phi)$) which can also be determined numerically from the parameters of the model.

Thus, we have verified that the dynamical equations of our model actually have place field-like fixed points. The next question is whether the system actually converges to this fixed point. A necessary condition for this to happen is that the fixed point corresponding to the place field solution be stable. This already guarantees that the network converges to this solution for some initial conditions, and the only remaining question is the extent of its basin of attraction.

In order to examine the stability of the solution determined above, we have performed linear stability analysis on the dynamical system described by equations (A.2), (A.3), (A.8) and (A.9) in the vicinity of the fixed point given by equation (A.13). We find that a necessary condition for this fixed point to be stable is that the network parameters satisfy a set of five inequalities, most of which are relatively complex. However, it is very easy to check whether any given set of parameters satisfy these conditions.

Finally, we have run numerical simulations of this simple model. We found that, in a relatively large area of the parameter space, the network actually converges to the non-directional place field solution every time if it is started from activities defined by the feedforward input. In a different area of the parameter space, the emergent solution is tuned in both space and direction. In addition, the solution also depends on training experience. In particular, if training is constrained to only two opposite directions (but includes all locations), the non-directional solution never emerges. Thus, this simple model can display many of the interesting characteristics of the considerably more complex model described in Chapter 3.

Bibliography

Acsády, L., Kamondi, A., Sík, A., Freund, T., and Buzsáki, G. (1998). GABAergic cells are the major postsynaptic targets of mossy fibers in the rat hippocampus. *J Neurosci*, 18:3386–3403.

Aggleton, J. P. and Saunders, R. C. (1997). The relationships between temporal lobe and diencephalic structures implicated in anterograde amnesia. *Memory*, 5:49–71.

Aguirre, G. K., Detre, J. A., Alsop, D. C., and D'Esposito, M. (1996). The parahippocampus subserves topographical learning in man. *Cereb Cortex*, 6:823–829.

Aigner, T. G., Walker, D. L., and Mishkin, M. (1991). Comparison of the effects of scopolamine administered before and after acquisition in a test of visual recognition memory in monkeys. *Behav Neural Biol*, 55:61–67.

Alonso, A. and Kohler, C. (1982). Evidence for separate projections of hippocampal pyramidal and non- pyramidal neurons to different parts of the septum in the rat brain. *Neurosci Lett*, 31:209–214.

Alvarado, M. C. and Rudy, J. W. (1995a). A comparison of kainic acid plus colchicine and ibotenic acid-induced hippocampal formation damage on four configural tasks in rats. *Behav Neurosci*, 109:1052–1062.

Alvarado, M. C. and Rudy, J. W. (1995b). Rats with damage to the hippocampal-formation are impaired on the transverse-patterning problem but not on elemental discriminations. *Behav Neurosci*, 109:204–211.

Alvarez, P. and Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: a simple network model. *Proc Natl Acad Sci U S A*, 91:7041–7045.

Alyan, S. and McNaughton, B. L. (1999). Hippocampectomized rats are capable of homing by path integration. *Behav Neurosci*, 113:19–31.

Amaral, D. G. (1993). Emerging principles of intrinsic hippocampal organization. *Curr Opin Neurobiol*, 3:225–229.

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

Amaral, D. G. and Witter, M. P. (1989). The three-dimensional organization of the hippocampal formation: a review of anatomical data. *Neuroscience*, 31:571–591.

Ambrogi Lorenzini, C. G., Baldi, E., Bucherelli, C., Sacchetti, B., and Tassoni, G. (1999). Neural topography and chronology of memory consolidation: a review of functional inactivation findings. *Neurobiol Learn Mem*, 71:1–18.

Anagnostaras, S. G., Maren, S., and Fanselow, M. S. (1999). Temporally graded retrograde amnesia of contextual fear after hippocampal damage in rats: within-subjects examination. *J Neurosci*, 19:1106–1114.

Arleo, A. and Gerstner, W. (2000). Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. *Biol Cybern*, 83:287–299.

Auer, R. N., Jensen, M. L., and Whishaw, I. Q. (1989). Neurobehavioral deficit due to ischemic brain damage limited to half of the CA1 sector of the hippocampus. *J Neurosci*, 9:1641–1647.

August, D. A. and Levy, W. B. (1999). Temporal sequence compression by an integrate-and-fire model of hippocampal area CA3. *J Comput Neurosci*, 6:71–90.

Ault, B. and Nadler, J. V. (1982). Baclofen selectively inhibits transmission at synapses made by axons of CA3 pyramidal cells in the hippocampal slice. *J Pharmacol Exp Ther*, 223:291–297.

Bannerman, D. M., Good, M. A., Butcher, S. P., Ramsay, M., and Morris, R. G. (1995). Distinct components of spatial learning revealed by prior training and NMDA receptor blockade. *Nature*, 378:182–186.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J., Leonard, B. W., and 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.

Battaglia, F. P. and Treves, A. (1998). Attractor neural networks storing multiple space representations: A model for hippocampal place fields. *Phys Rev E*, 58:7738–7753.

Baxter, M. G. and Murray, E. A. (2001). Opposite relationship of hippocampal and rhinal cortex damage to delayed nonmatching-to-sample deficits in monkeys. *Hippocampus*, 11:61–71.

Becker, S., Moscovitch, M., Behrmann, M., and Joordens, S. (1997). Long-term semantic priming: a computational account and empirical evidence. *J Exp Psychol Learn Mem Cogn*, 23:1059–1082.

Ben-Yishai, R., Bar-Or, R. L., and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. *Proc Natl Acad Sci U S A*, 92:3844–3848.

Benardo, L. S. and Prince, D. A. (1982). Ionic mechanisms of cholinergic excitation in mammalian hippocampal pyramidal cells. *Brain Res*, 249:333–344.

Berger, T. W., Alger, B., and Thompson, R. F. (1976). Neuronal substrate of classical conditioning in the hippocampus. *Science*, 192:483–485.

Berger, T. W., Rinaldi, P. C., Weisz, D. J., and Thompson, R. F. (1983). Single-unit analysis of different hippocampal cell types during classical conditioning of rabbit nictitating membrane response. *J Neurophysiol*, 50:1197–1219.

Bi, G. Q. and Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J Neurosci*, 18:10464–10472.

Blair, H. T. and Sharp, P. E. (1995). Anticipatory head direction signals in anterior thalamus: evidence for a thalamocortical circuit that integrates angular head motion to compute head direction. *J Neurosci*, 15:6260–6270.

Bliss, T. V. and Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *J Physiol*, 232:331–356.

Blum, K. I. and Abbott, L. F. (1996). A model of spatial map formation in the hippocampus of the rat. *Neural Comput*, 8:85–93.

Bogacz, R., Brown, M. W., and Giraud-Carrier, C. (2001). Model of familiarity discrimination in the perirhinal cortex. *J Comput Neurosci*, 10:5–23.

Bohbot, V. D., Kalina, M., Stepankova, K., Spackova, N., Petrides, M., and Nadel, L. (1998). Spatial memory deficits in patients with lesions to the right hippocampus and to the right parahippocampal cortex. *Neuropsychologia*, 36:1217–1238.

Bolhuis, J. J., Stewart, C. A., and Forrest, E. M. (1994). Retrograde amnesia and memory reactivation in rats with ibotenate lesions to the hippocampus or subiculum. *Q J Exp Psychol B*, 47:129–150.

Bostock, E., Muller, R. U., and Kubie, J. L. (1991). Experience-dependent modifications of hippocampal place cell firing. *Hippocampus*, 1:193–205.

Bragin, A., Jandó, G., Nádasdy, Z., Hetke, J., Wise, K., and Buzsáki, G. (1995). Gamma (40-100 Hz) oscillation in the hippocampus of the behaving rat. *J Neurosci*, 15:47–60.

Breindl, A., Derrick, B. E., Rodriguez, S. B., and Martinez, J. L. (1994). Opioid receptor-dependent long-term potentiation at the lateral perforant path-CA3 synapse in rat hippocampus. *Brain Res Bull*, 33:17–24.

Brown, M. W. and Aggleton, J. P. (2001). Recognition memory: what are the roles of the perirhinal cortex and hippocampus? *Nat Rev Neurosci*, 2:51–61.

Brown, M. W. and Xiang, J. Z. (1998). Recognition memory: neuronal substrates of the judgement of prior occurrence. *Prog Neurobiol*, 55:149–189.

Brunel, N. and Trullier, O. (1998). Plasticity of directional place fields in a model of rodent CA3. *Hippocampus*, 8:651–665.

Buckmaster, P. S. and Schwartzkroin, P. A. (1994). Hippocampal mossy cell function: a speculative view. *Hippocampus*, 4:393–402.

Buckner, R. L., Kelley, W. M., and Petersen, S. E. (1999). Frontal cortex contributes to human memory formation. *Nat Neurosci*, 2:311–314.

Buckner, R. L., Logan, J., Donaldson, D. I., and Wheeler, M. E. (2000). Cognitive neuroscience of episodic memory encoding. *Acta Psychol (Amst)*, 105:127–139.

Buffalo, E. A., Reber, P. J., and Squire, L. R. (1998). The human perirhinal cortex and recognition memory. *Hippocampus*, 8:330–339.

Bunsey, M. and Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature*, 379:255–257.

Burgard, E. C. and Sarvey, J. M. (1990). Muscarinic receptor activation facilitates the induction of long-term potentiation (LTP) in the rat dentate gyrus. *Neurosci Lett*, 116:34–39.

Burgess, N., Donnett, J. G., Jeffery, K. J., and O'Keefe, J. (1997). Robotic and neuronal simulation of the hippocampus and rat navigation. *Philos Trans R Soc Lond B Biol Sci*, 352:1535–1543.

Burgess, N. and O'Keefe, J. (1996). Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus*, 6:749–762.

Burwell, R. D. and Amaral, D. G. (1998). Cortical afferents of the perirhinal, postrhinal, and entorhinal cortices of the rat. *J Comp Neurol*, 398:179–205.

Bussey, T. J., Clea Warburton, E., Aggleton, J. P., and Muir, J. L. (1998). Fornix lesions can facilitate acquisition of the transverse patterning task: a challenge for "configural" theories of hippocampal function. *J Neurosci*, 18:1622–1631.

Bussey, T. J., Duck, J., Muir, J. L., and Aggleton, J. P. (2000). Distinct patterns of behavioural impairments resulting from fornix transection or neurotoxic lesions of the perirhinal and postrhinal cortices in the rat. *Behav Brain Res*, 111:187–202.

Buzsáki, G. (1986). Hippocampal sharp waves: their origin and significance. *Brain Res*, 398:242–252.

Buzsáki, G. (1989). Two-stage model of memory trace formation: a role for "noisy" brain states. *Neuroscience*, 31:551–570.

Buzsáki, G. (1996). The hippocampo-neocortical dialogue. *Cereb Cortex*, 6:81–92.

Cabeza, R. and Nyberg, L. (2000a). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *J Cogn Neurosci*, 12:1–47.

Cabeza, R. and Nyberg, L. (2000b). Neural bases of learning and memory: functional neuroimaging evidence. *Curr Opin Neurol*, 13:415–421.

Caine, E. D., Weingartner, H., Ludlow, C. L., Cudahy, E. A., and Wehry, S. (1981). Qualitative analysis of scopolamine-induced amnesia. *Psychopharmacology (Berl)*, 74:74–80.

Carpenter, G. A. and Grossberg, S. (1993). Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. *Trends Neurosci*, 16:131–137.

Chance, F. S., Nelson, S. B., and Abbott, L. F. (1999). Complex cells as cortically amplified simple cells. *Nat Neurosci*, 2:277–282.

Cho, Y. H., Beracochea, D., and Jaffard, R. (1993). Extended temporal gradient for the retrograde and anterograde amnesia produced by ibotenate entorhinal cortex lesions in mice. *J Neurosci*, 13:1759–1766.

Cho, Y. H. and Kesner, R. P. (1996). Involvement of entorhinal cortex or parietal cortex in long-term spatial discrimination memory in rats: retrograde amnesia. *Behav Neurosci*, 110:436–442.

Cho, Y. H., Kesner, R. P., and Brodale, S. (1995). Retrograde and anterograde amnesia for spatial discrimination in rats: role of hippocampus, entorhinal cortex, and parietal cortex. *Psychobiology*, 23:185–194.

Chrobak, J. J. and Buzsáki, G. (1998). Operational dynamics in the hippocampal-entorhinal axis. *Neurosci Biobehav Rev*, 22:303–310.

Clark, R. E., West, A. N., Zola, S. M., and Squire, L. R. (2001). Rats with lesions of the hippocampus are impaired on the delayed nonmatching-to-sample task. *Hippocampus*, 11:176–186.

Clayton, N. S. and Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, 395:272–274.

Clayton, N. S. and Dickinson, A. (1999). Scrub jays (Aphelocoma coerulescens) remember the relative time of caching as well as the location and content of their caches. *J Comp Psychol*, 113:403–416.

Clayton, N. S., Yu, K. S., and Dickinson, A. (2001). Scrub jays (Aphelocoma coerulescens) form integrated memories of the multiple features of caching episodes. *J Exp Psychol Anim Behav Process*, 27:17–29.

Cohen, N. J. and Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system.* MIT Press, Cambridge, Massachusetts.

Collins, D. R., Lang, E. J., and Pare, D. (1999). Spontaneous activity of the perirhinal cortex in behaving cats. *Neuroscience*, 89:1025–1039.

Corkin, S. (1984). Lasting consequences of bilateral medial temporal lobectomy: Clinical course and experimental findings in H.M. *Semin Neurol*, 4:249–259.

Corkin, S., Amaral, D. G., Gonzalez, R. G., Johnson, K. A., and Hyman, B. T. (1997). H. M.'s medial temporal lobe lesion: findings from magnetic resonance imaging. *J Neurosci*, 17:3964–3979.

Crow, T. J. and Grove-White, I. G. (1973). An analysis of the learning deficit following hyoscine administration to man. *Br J Pharmacol*, 49:322–327.

Davidson, T. L., McKernan, M. G., and Jarrard, L. E. (1993). Hippocampal lesions do not impair negative patterning: a challenge to configural association theory. *Behav Neurosci*, 107:227–234.

De Renzi, E., Liotti, M., and Nichelli, P. (1987). Semantic amnesia with preservation of autobiographic memory. A case report. *Cortex*, 23:575–597.

Deadwyler, S. A., Bunn, T., and Hampson, R. E. (1996). Hippocampal ensemble activity during spatial delayed-nonmatch-to-sample performance in rats. *J Neurosci*, 16:354–372.

Debanne, D., Gahwiler, B. H., and Thompson, S. M. (1998). Long-term synaptic plasticity between pairs of individual CA3 pyramidal cells in rat hippocampal slice cultures. *J Physiol*, 507 (Pt 1):237–247.

Desgranges, B., Baron, J. C., and Eustache, F. (1998). The functional neuroanatomy of episodic memory: the role of the frontal lobes, the hippocampal formation, and other areas. *Neuroimage*, 8:198–213.

Doboli, S., Minai, A. A., and Best, P. J. (2000). Latent attractors: a model for context-dependent place representations in the hippocampus. *Neural Comput*, 12:1009–1043.

Dudek, S. M. and Bear, M. F. (1993). Bidirectional long-term modification of synaptic effectiveness in the adult and immature hippocampus. *J Neurosci*, 13:2910–2918.

Dunnett, S. B. (1985). Comparative effects of cholinergic drugs and lesions of nucleus basalis or fimbria-fornix on delayed matching in rats. *Psychopharmacology (Berl)*, 87:357–363.

Dusek, J. A. and Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proc Natl Acad Sci U S A*, 94:7109–7114.

Dusek, J. A. and Eichenbaum, H. (1998). The hippocampus and transverse patterning guided by olfactory cues. *Behav Neurosci*, 112:762–771.

Duva, C. A., Floresco, S. B., Wunderlich, G. R., Lao, T. L., Pinel, J. P., and Phillips, A. G. (1997). Disruption of spatial but not object-recognition memory by neurotoxic lesions of the dorsal hippocampus in rats. *Behav Neurosci*, 111:1184–1196.

Eckerman, D. A., Gordon, W. A., Edwards, J. D., MacPhail, R. C., and Gage, M. I. (1980). Effects of scopolamine, pentobarbital, and amphetamine on radial arm maze performance in the rat. *Pharmacol Biochem Behav*, 12:595–602.

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nat Rev Neurosci*, 1:41–50.

Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., and Tanila, H. (1999). The hippocampus, memory, and place cells: is it spatial memory or a memory space? *Neuron*, 23:209–226.

Eichenbaum, H., Stewart, C., and Morris, R. G. (1990). Hippocampal representation in place learning. *J Neurosci*, 10:3531–3542.

Eldridge, L. L., Knowlton, B. J., Furmanski, C. S., Bookheimer, S. Y., and Engel, S. A. (2000). Remembering episodes: a selective role for the hippocampus during retrieval. *Nat Neurosci*, 3:1149–1152.

Elgersma, Y. and Silva, A. J. (1999). Molecular mechanisms of synaptic plasticity and memory. *Curr Opin Neurobiol*, 9:209–213.

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392:598–601.

Fanselow, M. S., Kim, J. J., Yipp, J., and De Oca, B. (1994). Differential effects of the N-methyl-D-aspartate antagonist DL-2-amino- 5-phosphonovalerate on acquisition of fear of auditory and contextual cues. *Behav Neurosci*, 108:235–240.

Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex*, 1:1–47.

Foster, D. J. (1999). *A computational inquiry into navigation with particular reference to the hippocampus*. PhD thesis, University of Edinburgh.

Foster, D. J., Morris, R. G., and Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, 10:1–16.

Frank, L. M., Brown, E. N., and Wilson, M. (2000). Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron*, 27:169–178.

French, R. M. (1997). Pseudo-recurrent connectionist networks: an approach to the 'sensitivity-stability' dilemma. *Connection Science*, 9(4):353–379.

Freund, T. F. and Buzsáki, G. (1996). Interneurons of the hippocampus. *Hippocampus*, 6:347–470.

Frith, C. D., Richardson, J. T., Samuel, M., Crow, T. J., and McKenna, P. J. (1984). The effects of intravenous diazepam and hyoscine upon human memory. *Q J Exp Psychol A*, 36:133–144.

Gadian, D. G., Aicardi, J., Watkins, K. E., Porter, D. A., Mishkin, M., and Vargha-Khadem, F. (2000). Developmental amnesia associated with early hypoxic-ischaemic injury. *Brain*, 123 Pt 3:499–507.

Gaffan, D. (1993). Additive effects of forgetting and fornix transfection in the temporal gradient of retrograde amnesia. *Neuropsychologia*, 31:1055–1066.

Gais, S., Plihal, W., Wagner, U., and Born, J. (2000). Early sleep triggers memory for early visual discrimination skills. *Nat Neurosci*, 3:1335–1339.

Gallagher, M. and Holland, P. C. (1992). Preserved configural learning and spatial learning impairment in rats with hippocampal damage. *Hippocampus*, 2:81–88.

Gallistel, C. R. (1990). *The organization of learning*. MIT Press, Cambridge, MA.

Georges-Francois, P., Rolls, E. T., and Robertson, R. G. (1999). Spatial view cells in the primate hippocampus: allocentric view not head direction or eye position or place. *Cereb Cortex*, 9:197–212.

Ghoneim, M. M. and Mewaldt, S. P. (1975). Effects of diazepam and scopolamine on storage, retrieval and organizational processes in memory. *Psychopharmacologia*, 44:257–262.

Gluck, M. A., Ermita, B. R., Oliver, L. M., and Myers, C. E. (1997). Extending models of hippocampal function in animal conditioning to human amnesia. *Memory*, 5:179–212.

Gluck, M. A. and Myers, C. E. (1993). Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus*, 3:491–516.

Good, M. and Honey, R. C. (1991). Conditioning and contextual retrieval in hippocampal rats. *Behav Neurosci*, 105:499–509.

Grossi, D., Trojano, L., Grasso, A., and Orsini, A. (1988). Selective "semantic amnesia" after closed-head injury. A case report. *Cortex*, 24:457–464.

Haist, F., Shimamura, A. P., and Squire, L. R. (1992). On the relationship between recall and recognition memory. *J Exp Psychol Learn Mem Cogn*, 18:691–702.

Hamann, S. B. and Squire, L. R. (1995). On the acquisition of new declarative knowledge in amnesia. *Behav Neurosci*, 109:1027–1044.

Harris, E. W. and Cotman, C. W. (1986). Long-term potentiation of guinea pig mossy fiber responses is not blocked by N-methyl D-aspartate antagonists. *Neurosci Lett*, 70:132–137.

Hartley, T., Burgess, N., Lever, C., Cacucci, F., and O'Keefe, J. (2000). Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus*, 10:369–379.

Hasselmo, M. E. (1999). Neuromodulation: acetylcholine and memory consolidation. *Trends Cogn Sci*, 3:351–359.

Hasselmo, M. E., Schnell, E., and Barkai, E. (1995). Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *J Neurosci*, 15:5249–5262.

Hasselmo, M. E. and Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function. *Behav Brain Res*, 89:1–34.

Hasselmo, M. E., Wyble, B. P., and Wallenstein, G. V. (1996). Encoding and retrieval of episodic memories: role of cholinergic and GABAergic modulation in the hippocampus. *Hippocampus*, 6:693–708.

Hennevin, E., Hars, B., Maho, C., and Bloch, V. (1995). Processing of learned information in paradoxical sleep: relevance for memory. *Behav Brain Res*, 69:125–135.

Hetherington, P. A., Austin, K. B., and Shapiro, M. L. (1994). Ipsilateral associational pathway in the dentate gyrus: an excitatory feedback system that supports N-methyl-D-aspartate-dependent long-term potentiation. *Hippocampus*, 4:422–438.

Hill, A. J. (1978). First occurrence of hippocampal spatial firing in a new environment. *Exp Neurol*, 62:282–297.

Hinton, G. (2000). Training Products of Experts by Minimizing Contrastive Divergence. Technical report, Gatsby Unit, UCL.

Hinton, G. and Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations.* MIT Press, Cambridge, MA.

Hinton, G. and Sejnowski, T. J., editors (1999). *Unsupervised learning.* MIT Press, Cambridge, MA.

Hinton, G. E. and McClelland, J. L. (1988). Learning representations by recirculation. In Anderson, D. Z., editor, *Neural Information Processing Systems, 1987*, pages 358–366. American Institute of Physics, New York.

Hirsh, R., Leber, B., and Gillman, K. (1978). Fornix fibers and motivational states as controllers of behavior: a study stimulated by the contextual retrieval theory. *Behav Biol*, 22:463–478.

Hirst, W., Johnson, M. K., Kim, J. K., Phelps, E. A., Risse, G., and Volpe, B. T. (1986). Recognition and recall in amnesics. *J Exp Psychol Learn Mem Cogn*, 12:445–451.

Hirst, W., Johnson, M. K., Phelps, E. A., and Volpe, B. T. (1988). More on recognition and recall in amnesics. *J Exp Psychol Learn Mem Cogn*, 14:758–762.

Hoh, T., Beiko, J., Boon, F., Weiss, S., and Cain, D. P. (1999). Complex behavioral strategy and reversal learning in the water maze without NMDA receptor-dependent long-term potentiation. *J Neurosci*, 19:RC2.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A*, 79:2554–2558.

Huerta, P. T. and Lisman, J. E. (1993). Heightened synaptic plasticity of hippocampal CA1 neurons during a cholinergically induced rhythmic state. *Nature*, 364:723–725.

Irle, E. and Markowitsch, H. J. (1982). Widespread cortical projections of the hippocampal formation in the cat. *Neuroscience*, 7:2637–2647.

Ishizuka, N., Weber, J., and Amaral, D. G. (1990). Organization of intrahippocampal projections originating from CA3 pyramidal cells in the rat. *J Comp Neurol*, 295:580–623.

Jaffe, D. and Johnston, D. (1990). Induction of long-term potentiation at hippocampal mossy-fiber synapses follows a Hebbian rule. *J Neurophysiol*, 64:948–960.

Jarrard, L. E. (1993). On the role of the hippocampus in learning and memory in the rat. *Behav Neural Biol*, 60:9–26.

Jarrard, L. E. and Davidson, T. L. (1991). On the hippocampus and learned conditional responding: effects of aspiration versus ibotenate lesions. *Hippocampus*, 1:107–117.

Jarrard, L. E. and Meldrum, B. S. (1993). Selective excitotoxic pathology in the rat hippocampus. *Neuropathol Appl Neurobiol*, 19:381–389.

Jeffery, K. J. (1998). Learning of landmark stability and instability by hippocampal place cells. *Neuropharmacology*, 37:677–687.

Jeffery, K. J. and O'Keefe, J. M. (1999). Learned interaction of visual and idiothetic cues in the control of place field orientation. *Exp Brain Res*, 127:151–161.

Jetter, W., Poser, U., Freeman Jr, R. B., and Markowitsch, H. J. (1986). A verbal long term memory deficit in frontal lobe damaged patients. *Cortex*, 22:229–242.

Jung, M. W. and McNaughton, B. L. (1993). Spatial selectivity of unit activity in the hippocampal granular layer. *Hippocampus*, 3:165–182.

Káli, S. and Dayan, P. (1998). The formation of direction independent place fields in area CA3 of the rodent hippocampus using Hebbian plasticity in a recurrent network. *Soc Neurosci Abstr*, 24:931.

Káli, S. and Dayan, P. (2000). The involvement of recurrent connections in area CA3 in establishing the properties of place fields: a model. *J Neurosci*, 20:7463–7477.

Kapur, N. (1993). Focal retrograde amnesia in neurological disease: a critical review. *Cortex*, 29:217–234.

Karni, A. and Sagi, D. (1993). The time course of learning a visual skill. *Nature*, 365:250–252.

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J., and Sagi, D. (1994). Dependence on REM sleep of overnight improvement of a perceptual skill. *Science*, 265:679–682.

Kim, J. J., Clark, R. E., and Thompson, R. F. (1995). Hippocampectomy impairs the memory of recently, but not remotely, acquired trace eyeblink conditioned responses. *Behav Neurosci*, 109:195–203.

Kim, J. J. and Fanselow, M. S. (1992). Modality-specific retrograde amnesia of fear. *Science*, 256:675–677.

Knierim, J. J., Kudrimoti, H. S., and McNaughton, B. L. (1995). Place cells, head direction cells, and the learning of landmark stability. *J Neurosci*, 15:1648–1659.

Knierim, J. J., Kudrimoti, H. S., and McNaughton, B. L. (1998). Interactions between idiothetic cues and external landmarks in the control of place cells and head direction cells. *J Neurophysiol*, 80:425–446.

Kubie, J. L., Muller, R. U., and Bostock, E. (1990). Spatial firing properties of hippocampal theta cells. *J Neurosci*, 10:1110–1123.

Kubie, J. L., Sutherland, R. J., and Muller, R. U. (1999). Hippocampal lesions produce a temporally graded retrograde amnesia on a dry version of the Morris swimming task. *Psychobiology*, 27:313–330.

Kudrimoti, H. S., Barnes, C. A., and McNaughton, B. L. (1999). Reactivation of hippocampal cell assemblies: effects of behavioral state, experience, and EEG dynamics. *J Neurosci*, 19:4090–4101.

Land, C., Bunsey, M., and Riccio, D. C. (2000). Anomalous properties of hippocampal lesion-induced retrograde amnesia. *Psychobiology*, 28:476–485.

Lavenex, P. and Amaral, D. G. (2000). Hippocampal-neocortical interaction: a hierarchy of associativity. *Hippocampus*, 10:420–430.

Lepage, M., Habib, R., and Tulving, E. (1998). Hippocampal PET activations of memory encoding and retrieval: the HIPER model. *Hippocampus*, 8:313–322.

Leung, L. S. (1998). Generation of theta and gamma rhythms in the hippocampus. *Neurosci Biobehav Rev*, 22:275–290.

Levy, W. B. (1996). A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus*, 6:579–590.

Levy, W. B. and Steward, O. (1979). Synapses as associative memory elements in the hippocampal formation. *Brain Res*, 175:233–245.

Levy, W. B. and Steward, O. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, 8:791–797.

Li, X. G., Somogyi, P., Ylinen, A., and Buzsáki, G. (1994). The hippocampal CA3 network: an in vivo intracellular labeling study. *J Comp Neurol*, 339:181–208.

Li, Z. and Dayan, P. (1999). Computational differences between asymmetrical and symmetrical networks. *Network*, 10:59–77.

Lisman, J. E. (1999). Relating hippocampal circuitry to function: recall of memory sequences by reciprocal dentate-CA3 interactions. *Neuron*, 22:233–242.

Louie, K. and Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, 29:145–156.

Lynch, G. S., Dunwiddie, T., and Gribkoff, V. (1977). Heterosynaptic depression: a postsynaptic correlate of long-term potentiation. *Nature*, 266:737–739.

Maaswinkel, H., Jarrard, L. E., and Whishaw, I. Q. (1999). Hippocampectomized rats are impaired in homing by path integration. *Hippocampus*, 9:553–561.

Magee, J. C. and Johnston, D. (1997). A synaptically controlled, associative signal for Hebbian plasticity in hippocampal neurons. *Science*, 275:209–213.

Maguire, E. A., Burgess, N., Donnett, J. G., Frackowiak, R. S., Frith, C. D., and O'Keefe, J. (1998a). Knowing where and getting there: a human navigation network. *Science*, 280:921–924.

Maguire, E. A., Frith, C. D., Burgess, N., Donnett, J. G., and O'Keefe, J. (1998b). Knowing where things are parahippocampal involvement in encoding object locations in virtual large-scale space. *J Cogn Neurosci*, 10:61–76.

Maguire, E. A., Frith, C. D., and Cipolotti, L. (2001). Distinct neural systems for the encoding and recognition of topography and faces. *Neuroimage*, 13:743–750.

Maguire, E. A., Mummery, C. J., and Buchel, C. (2000). Patterns of hippocampal-cortical interaction dissociate temporal lobe memory subsystems. *Hippocampus*, 10:475–482.

Malenka, R. C. and Nicoll, R. A. (1999). Long-term potentiation–a decade of progress? *Science*, 285:1870–1874.

Manabe, T. (1997). Two forms of hippocampal long-term depression, the counterpart of long- term potentiation. *Rev Neurosci*, 8:179–193.

Maren, S. and Fanselow, M. S. (1997). Electrolytic lesions of the fimbria/fornix, dorsal hippocampus, or entorhinal cortex produce anterograde deficits in contextual fear conditioning in rats. *Neurobiol Learn Mem*, 67:142–149.

Markowitsch, H. J., Calabrese, P., Haupts, M., Durwen, H. F., Liess, J., and Gehlen, W. (1993). Searching for the anatomical basis of retrograde amnesia. *J Clin Exp Neuropsychol*, 15:947–967.

Markram, H., Lubke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275:213–215.

Markus, E. J., Qin, Y. L., Leonard, B., Skaggs, W. E., McNaughton, B. L., and Barnes, C. A. (1995). Interactions between location and task affect the spatial and directional firing of hippocampal neurons. *J Neurosci*, 15:7079–7094.

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

Marr, D. (1982). *Vision*. W. H. Freeman and Co., New York.

Martin, S. J., Grimwood, P. D., and Morris, R. G. (2000). Synaptic plasticity and memory: an evaluation of the hypothesis. *Annu Rev Neurosci*, 23:649–711.

Mayes, A. R. and Downes, J. J. (1997). What do theories of the functional deficit(s) underlying amnesia have to explain? *Memory*, 5:3–36.

Mayford, M., Bach, M. E., Huang, Y. Y., Wang, L., Hawkins, R. D., and Kandel, E. R. (1996). Control of memory formation through regulated expression of a CaMKII transgene. *Science*, 274:1678–1683.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev*, 102:419–457.

McClelland, J. L. and Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *J Exp Psychol Gen*, 114:159–197.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In Bower, G., editor, *The psychology of learning and motivation, vol 24*, pages 109–165. Academic Press, New York.

McDonald, R. J., Murphy, R. A., Guarraci, F. A., Gortler, J. R., White, N. M., and Baker, A. G. (1997). Systematic comparison of the effects of hippocampal and fornix-fimbria lesions on acquisition of three configural discriminations. *Hippocampus*, 7:371–388.

McEchron, M. D. and Disterhoft, J. F. (1997). Sequence of single neuron changes in CA1 hippocampus of rabbits during acquisition of trace eyeblink conditioned responses. *J Neurophysiol*, 78:1030–1044.

McHugh, T. J., Blum, K. I., Tsien, J. Z., Tonegawa, S., and Wilson, M. A. (1996). Impaired hippocampal representation of space in CA1-specific NMDAR1 knockout mice. *Cell*, 87:1339–1349.

McLeod, P., Plunkett, K., and Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes.* Oxford University Press, Oxford, UK.

McNaughton, B. L., Barnes, C. A., Meltzer, J., and Sutherland, R. J. (1989). Hippocampal granule cells are necessary for normal spatial learning but not for spatially-selective pyramidal cell discharge. *Exp Brain Res*, 76:485–496.

McNaughton, B. L., Barnes, C. A., and O'Keefe, J. (1983). The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. *Exp Brain Res*, 52:41–49.

McNaughton, B. L. and Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends Neurosci*, 10:408–415.

Milner, P. M. (1989). A cell assembly theory of hippocampal amnesia. *Neuropsychologia*, 27:23–30.

Mishkin, M. (1982). A memory system in the monkey. *Philos Trans R Soc Lond B Biol Sci*, 298:83–95.

Mizumori, S. J., Ragozzino, K. E., and Cooper, B. G. (2000). Location and head direction representation in the dorsal striatum of rats. *Psychobiology*, 28:441–462.

Mizumori, S. J., Ward, K. E., and Lavoie, A. M. (1992). Medial septal modulation of entorhinal single unit activity in anesthetized and freely moving rats. *Brain Res*, 570:188–197.

Morris, R. G., Anderson, E., Lynch, G. S., and Baudry, M. (1986). Selective impairment of learning and blockade of long-term potentiation by an N-methyl-D-aspartate receptor antagonist, AP5. *Nature*, 319:774–776.

Morris, R. G., Garrud, P., Rawlins, J. N., and O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, 297:681–683.

Morris, R. G. M. (1981). Spatial localization does not require the presence of local cues. *Learn Motiv*, 12:239–260.

Morris, R. G. M., Schenk, F., Tweedie, F., and Jarrard, L. E. (1990). Ibotenate lesions of hippocampus and/or subiculum: Dissociating components of allocentric spatial learning. *Eur J Neurosci*, 2:1016–1028.

Moser, E. I., Krobert, K. A., Moser, M. B., and Morris, R. G. (1998). Impaired spatial learning after saturation of long-term potentiation. *Science*, 281:2038–2042.

Muller, R. (1996). A quarter of a century of place cells. *Neuron*, 17:813–822.

Muller, R. U., Bostock, E., Taube, J. S., and Kubie, J. L. (1994). On the directional firing properties of hippocampal place cells. *J Neurosci*, 14:7235–7251.

Muller, R. U. and Kubie, J. L. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J Neurosci*, 7:1951–1968.

Muller, R. U., Kubie, J. L., Bostock, E. M., Taube, J. S., and Quirk, G. J. (1991a). Spatial firing correlates of neurons in the hippocampal formation of freely moving rats. In Paillard, J., editor, *Brain and Space*, chapter 17, pages 296–333. Oxford University Press, New York.

Muller, R. U., Kubie, J. L., and Ranck, J. B. (1987). Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. *J Neurosci*, 7:1935–1950.

Muller, R. U., Kubie, J. L., and Saypoff, R. (1991b). The hippocampus as a cognitive graph (abridged version). *Hippocampus*, 1:243–246.

Muller, R. U., Ranck, J. B., and Taube, J. S. (1996a). Head direction cells: properties and functional significance. *Curr Opin Neurobiol*, 6:196–206.

Muller, R. U., Stead, M., and Pach, J. (1996b). The hippocampus as a cognitive graph. *J Gen Physiol*, 107:663–694.

Mumby, D. G., Astur, R. S., Weisend, M. P., and Sutherland, R. J. (1999). Retrograde amnesia and selective damage to the hippocampal formation: memory for places and object discriminations. *Behav Brain Res*, 106:97–107.

Mumby, D. G., Wood, E. R., Duva, C. A., Kornecook, T. J., Pinel, J. P., and Phillips, A. G. (1996). Ischemia-induced object-recognition deficits in rats are attenuated by hippocampal ablation before or soon after ischemia. *Behav Neurosci*, 110:266–281.

Murray, E. A. and Bussey, T. J. (2001). Consolidation and the medial temporal lobe revisited: methodological considerations. *Hippocampus*, 11:1–7.

Murre, J. M. (1997). Implicit and explicit memory in amnesia: some explanations and predictions by the TraceLink model. *Memory*, 5:213–232.

Myers, C. E., Gluck, M. A., and Granger, R. (1995). Dissociation of hippocampal and entorhinal function in associative learning: A computational approach. *Psychobiology*, 23:116–138.

Nadel, L. and Bohbot, V. (2001). Consolidation of memory. *Hippocampus*, 11:56–60.

Nadel, L. and Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr Opin Neurobiol*, 7:217–227.

Nadel, L., Samsonovich, A., Ryan, L., and Moscovitch, M. (2000). Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus*, 10:352–368.

Nicoll, R. A. and Malenka, R. C. (1995). Contrasting properties of two forms of long-term potentiation in the hippocampus. *Nature*, 377:115–118.

Ochsner, K. N., Chiu, C. Y., and Schacter, D. L. (1994). Varieties of priming. *Curr Opin Neurobiol*, 4:189–194.

O'Keefe, J. (1976). Place units in the hippocampus of the freely moving rat. *Exp Neurol*, 51:78–109.

O'Keefe, J. (1999). Do hippocampal pyramidal cells signal non-spatial as well as spatial information? *Hippocampus*, 9:352–364.

O'Keefe, J. and Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, 381:425–428.

O'Keefe, J. and Conway, D. H. (1978). Hippocampal place units in the freely moving rat: why they fire where they fire. *Exp Brain Res*, 31:573–590.

O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res*, 34:171–175.

O'Keefe, J. and Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press, Oxford.

O'Keefe, J. and Recce, M. L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus*, 3:317–330.

Olton, D. S., Becker, J. T., and Handelmann, G. E. (1979). Hippocampus, space, and memory. *Behav Brain Sci*, 2:313–322.

Olton, D. S. and Samuelson, R. J. (1976). Remembrance of places passed: Spatial memory in rats. *J Exp Psychol Anim Behav Process*, 2:97–116.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, 8(5):895–938.

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

O'Reilly, R. C. and Rudy, J. W. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychol Rev.* In press.

Otto, T. and Eichenbaum, H. (1992a). Complementary roles of the orbital prefrontal cortex and the perirhinal- entorhinal cortices in an odor-guided delayed-nonmatching-to-sample task. *Behav Neurosci*, 106:762–775.

Otto, T. and Eichenbaum, H. (1992b). Neuronal activity in the hippocampus during delayed non-match to sample performance in rats: evidence for hippocampal processing in recognition memory. *Hippocampus*, 2:323–334.

Pavlides, C. and Winson, J. (1989). Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *J Neurosci*, 9:2907–2918.

Penick, S. and Solomon, P. R. (1991). Hippocampus, context, and conditioning. *Behav Neurosci*, 105:611–617.

Phillips, R. G. and LeDoux, J. E. (1992). Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behav Neurosci*, 106:274–285.

Phillips, R. G. and LeDoux, J. E. (1994). Lesions of the dorsal hippocampal formation interfere with background but not foreground contextual fear conditioning. *Learn Mem*, 1:34–44.

Pitler, T. A. and Alger, B. E. (1992). Cholinergic excitation of GABAergic interneurons in the rat hippocampal slice. *J Physiol*, 450:127–142.

Plihal, W. and Born, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *J Cogn Neurosci*, 9:534–547.

Pouget, A., Zhang, K., Deneve, S., and Latham, P. E. (1998). Statistically efficient estimation using population coding. *Neural Comput*, 10:373–401.

Qin, Y. L., McNaughton, B. L., Skaggs, W. E., and Barnes, C. A. (1997). Memory reprocessing in corticocortical and hippocampocortical neuronal ensembles. *Philos Trans R Soc Lond B Biol Sci*, 352:1525–1533.

Quirk, G. J., Muller, R. U., and Kubie, J. L. (1990). The firing of hippocampal place cells in the dark depends on the rat's recent experience. *J Neurosci*, 10:2008–2017.

Quirk, G. J., Muller, R. U., Kubie, J. L., and Ranck, J. B. (1992). The positional firing properties of medial entorhinal neurons: description and comparison with hippocampal place cells. *J Neurosci*, 12:1945–1963.

Rasmussen, M., Barnes, C. A., and McNaughton, B. L. (1989). A systematic test of cognitive mapping, working-memory, and temporal discontiguity theories of hippocampal function. *Psychobiology*, 17:335–348.

Redish, A. D. (1999). *Beyond the cognitive map: from place cells to episodic memory.* MIT Press, Cambridge, Massachusetts.

Redish, A. D. and Touretzky, D. S. (1997). Cognitive maps beyond the hippocampus. *Hippocampus*, 7:15–35.

Reed, J. M. and Squire, L. R. (1998). Retrograde amnesia for facts and events: findings from four new cases. *J Neurosci*, 18:3943–3954.

Rempel-Clower, N. L., Zola, S. M., Squire, L. R., and Amaral, D. G. (1996). Three cases of enduring memory impairment after bilateral damage limited to the hippocampal formation. *J Neurosci*, 16:5233–5255.

Riedel, G., Micheau, J., Lam, A. G., v. Roloff, E., Martin, S. J., Bridge, H., d. Hoz, L., Poeschel, B., McCulloch, J., and Morris, R. G. (1999). Reversible neural inactivation reveals hippocampal participation in several memory processes. *Nat Neurosci*, 2:898–905.

Robertson, R. G., Rolls, E. T., and Georges-Francois, P. (1998). Spatial view cells in the primate hippocampus: effects of removal of view details. *J Neurophysiol*, 79:1145–1156.

Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

Robins, A. (1996). Consolidation in neural networks and in the sleeping brain. *Connection Science*, 8(2):259–275.

Rolls, E. T. (1996). A theory of hippocampal function in memory. *Hippocampus*, 6:601–620.

Rolls, E. T. and O'Mara, S. M. (1995). View-responsive neurons in the primate hippocampal complex. *Hippocampus*, 5:409–424.

Rosenbaum, R. S., Priselac, S., Kohler, S., Black, S. E., Gao, F., Nadel, L., and Moscovitch, M. (2000). Remote spatial memory in an amnesic person with extensive bilateral hippocampal lesions. *Nat Neurosci*, 3:1044–1048.

Ross, R. T., Orr, W. B., Holland, P. C., and Berger, T. W. (1984). Hippocampectomy disrupts acquisition and retention of learned conditional responding. *Behav Neurosci*, 98:211–225.

Rotenberg, A., Mayford, M., Hawkins, R. D., Kandel, E. R., and Muller, R. U. (1996). Mice expressing activated CaMKII lack low frequency LTP and do not form stable place cells in the CA1 region of the hippocampus. *Cell*, 87:1351–1361.

Rudy, J. W. and Sutherland, R. J. (1989). The hippocampal formation is necessary for rats to learn and remember configural discriminations. *Behav Brain Res*, 34:97–109.

Rudy, J. W. and Sutherland, R. J. (1995). Configural association theory and the hippocampal formation: an appraisal and reconfiguration. *Hippocampus*, 5:375–389.

Rusconi, M. L., Zago, S., and Basso, A. (1997). Semantic amnesia without dementia: documentation of a case. *Ital J Neurol Sci*, 18:167–171.

Sakurai, Y. (1996). Hippocampal and neocortical cell assemblies encode memory processes for different types of stimuli in the rat. *J Neurosci*, 16:2809–2819.

Salmon, D. P., Zola-Morgan, S., and Squire, L. R. (1987). Retrograde amnesia following combined hippocampus-amygdala lesions in monkeys. *Psychobiology*, 15:37–47.

Samsonovich, A. (1997). *Attractor map theory of the hippocampal representation of space.* PhD thesis, University of Arizona.

Samsonovich, A. and McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *J Neurosci*, 17:5900–5920.

Samsonovich, A., McNaughton, B. L., and Nadel, L. (1998). Hierarchical multichart model of the hippocampal cognitive map. *Soc Neurosci Abstr*, 24:931.

Sara, S. J. (1981). Memory deficits in rats with hippocampal or cortical lesions: retrograde effects. *Behav Neural Biol*, 32:504–509.

Saucier, D. and Cain, D. P. (1995). Spatial learning without NMDA receptor-dependent long-term potentiation. *Nature*, 378:186–189.

Schacter, D. L. and Wagner, A. D. (1999). Medial temporal lobe activations in fMRI and PET studies of episodic encoding and retrieval. *Hippocampus*, 9:7–24.

Scoville, W. and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *J Neurol Neurosurg Psychiatry*, 20:11–21.

Seress, L. (1988). Interspecies comparison of the hippocampal formation shows increased emphasis on the regio superior in the Ammon's horn of the human brain. *J Hirnforsch*, 29:335–340.

Sharp, E. P. (1991). Computer simulation of hippocampal place cells. *Psychobiology*, 19:103–115.

Sharp, P. E. (1997). Subicular cells generate similar spatial firing patterns in two geometrically and visually distinctive environments: comparison with hippocampal place cells. *Behav Brain Res*, 85:71–92.

Sharp, P. E., Blair, H. T., Etkin, D., and Tzanetos, D. B. (1995). Influences of vestibular and visual motion information on the spatial firing patterns of hippocampal place cells. *J Neurosci*, 15:173–189.

Sharp, P. E. and Green, C. (1994). Spatial correlates of firing patterns of single cells in the subiculum of the freely moving rat. *J Neurosci*, 14:2339–2356.

Shaw, C. and Aggleton, J. P. (1993). The effects of fornix and medial prefrontal lesions on delayed non- matching-to-sample by rats. *Behav Brain Res*, 54:91–102.

Shen, B. and McNaughton, B. L. (1996). Modeling the spontaneous reactivation of experience-specific hippocampal cell assembles during sleep. *Hippocampus*, 6:685–692.

Shimamura, A. P. and Squire, L. R. (1986). Korsakoff's syndrome: a study of the relation between anterograde amnesia and remote memory impairment. *Behav Neurosci*, 100:165–170.

Shimizu, E., Tang, Y. P., Rampon, C., and Tsien, J. Z. (2000). NMDA receptor-dependent synaptic reinforcement as a crucial process for memory consolidation. *Science*, 290:1170–1174.

Shors, T. J. and Matzel, L. D. (1997). Long-term potentiation: what's learning got to do with it? *Behav Brain Sci*, 20:597–614; discussion 614–55.

Siapas, A. G. and Wilson, M. A. (1998). Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. *Neuron*, 21:1123–1128.

Skaggs, W. E. and McNaughton, B. L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science*, 271:1870–1873.

Skaggs, W. E. and McNaughton, B. L. (1998). Spatial firing properties of hippocampal CA1 populations in an environment containing two visually identical regions. *J Neurosci*, 18:8455–8466.

Skaggs, W. E., McNaughton, B. L., Wilson, M. A., and Barnes, C. A. (1996). Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus*, 6:149–172.

Smith, C. (1995). Sleep states and memory processes. *Behav Brain Res*, 69:137–145.

Sohal, V. S. and Hasselmo, M. E. (1998). GABA(B) modulation improves sequence disambiguation in computational models of hippocampal region CA3. *Hippocampus*, 8:171–193.

Somers, D. C., Nelson, S. B., and Sur, M. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. *J Neurosci*, 15:5448–5465.

Squire, L. R. (1992). Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychol Rev*, 99:195–231.

Squire, L. R. and Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Curr Opin Neurobiol*, 5:169–177.

Squire, L. R., Clark, R. E., and Knowlton, B. J. (2001). Retrograde amnesia. *Hippocampus*, 11:50–55.

Squire, L. R., Cohen, N. J., and Nadel, L. (1984). The medial temporal region and memory consolidation: A new hypothesis. In Weingartner, H. and Parker, E., editors, *Memory consolidation*, pages 185–210. Erlbaum, Hillsdale, NJ.

Squire, L. R., Knowlton, B., and Musen, G. (1993). The structure and organization of memory. *Annu Rev Psychol*, 44:453–495.

Squire, L. R. and Zola, S. M. (1998). Episodic memory, semantic memory, and amnesia. *Hippocampus*, 8:205–211.

Stackman, R. W. and Taube, J. S. (1998). Firing properties of rat lateral mammillary single units: head direction, head pitch, and angular head velocity. *J Neurosci*, 18:9020–9037.

Stark, C. E. and McClelland, J. L. (2000). Repetition priming of words, pseudowords, and nonwords. *J Exp Psychol Learn Mem Cogn*, 26:945–972.

Stark, C. E. and Squire, L. R. (2000). Recognition memory and familiarity judgments in severe amnesia: no evidence for a contribution of repetition priming. *Behav Neurosci*, 114:459–467.

Steele, R. J. and Morris, R. G. (1999). Delay-dependent impairment of a matching-to-place task with chronic and intrahippocampal infusion of the NMDA-antagonist D-AP5. *Hippocampus*, 9:118–136.

Stickgold, R. (1998). Sleep: off-line memory reprocessing. *Trends Cogn Sci*, 2(12):484–492.

Stickgold, R., Whidbee, D., Schirmer, B., Patel, V., and Hobson, J. A. (2000). Visual discrimination task improvement: A multi-step process occurring during sleep. *J Cogn Neurosci*, 12:246–254.

Sutherland, R. J., Weisend, M. P., Mumby, D., Astur, R. S., Hanlon, F. M., Koerner, A., Thomas, M. J., Wu, Y., Moses, S. N., Cole, C., Hamilton, D. A., and Hoesing, J. M. (2001). Retrograde amnesia after hippocampal damage: recent vs. remote memories in two tasks. *Hippocampus*, 11:27–42.

Sutherland, R. J., Whishaw, I. Q., and Regehr, J. C. (1982). Cholinergic receptor blockade impairs spatial localization by use of distal cues in the rat. *J Comp Physiol Psychol*, 96:563–573.

Suzuki, W. A. and Eichenbaum, H. (2000). The neurophysiology of memory. *Ann N Y Acad Sci*, 911:175–191.

Taube, J. S. (1995a). Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *J Neurosci*, 15:70–86.

Taube, J. S. (1995b). Place cells recorded in the parasubiculum of freely moving rats. *Hippocampus*, 5:569–583.

Taube, J. S. (1998). Head direction cells and the neurophysiological basis for a sense of direction. *Prog Neurobiol*, 55:225–256.

Taube, J. S., Muller, R. U., and Ranck, J. B. (1990a). Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J Neurosci*, 10:420–435.

Taube, J. S., Muller, R. U., and Ranck, J. B. (1990b). Head-direction cells recorded from the postsubiculum in freely moving rats. II. Effects of environmental manipulations. *J Neurosci*, 10:436–447.

Teng, E. and Squire, L. R. (1999). Memory for places learned long ago is intact after hippocampal damage. *Nature*, 400:675–677.

Teyler, T. J. and DiScenna, P. (1986). The hippocampal memory indexing theory. *Behav Neurosci*, 100:147–154.

Thiels, E., Xie, X., Yeckel, M. F., Barrionuevo, G., and Berger, T. W. (1996). NMDA receptor-dependent LTD in different subfields of hippocampus in vivo and in vitro. *Hippocampus*, 6:43–51.

Thierry, A. M., Gioanni, Y., Degenetais, E., and Glowinski, J. (2000). Hippocampo-prefrontal cortex pathway: anatomical and electrophysiological characteristics. *Hippocampus*, 10:411–419.

Thompson, L. T. and Best, P. J. (1990). Long-term stability of the place-field activity of single units recorded from the dorsal hippocampus of freely behaving rats. *Brain Res*, 509:299–308.

Thornton, J. A., Rothblat, L. A., and Murray, E. A. (1997). Rhinal cortex removal produces amnesia for preoperatively learned discrimination problems but fails to disrupt postoperative acquisition and retention in rhesus monkeys. *J Neurosci*, 17:8536–8549.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol Rev*, 55:189–208.

Tonkiss, J. and Rawlins, J. N. (1991). The competitive NMDA antagonist AP5, but not the non-competitive antagonist MK801, induces a delay-related impairment in spatial working memory in rats. *Exp Brain Res*, 85:349–358.

Touretzky, D. S. and Redish, A. D. (1996). Theory of rodent navigation based on interacting representations of space. *Hippocampus*, 6:247–270.

Treves, A. and Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus*, 2:189–199.

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

Trullier, O., Wiener, S. I., Berthoz, A., and Meyer, J. A. (1997). Biologically based artificial navigation systems: review and prospects. *Prog Neurobiol*, 51:483–544.

Tsien, J. Z., Chen, D. F., Gerber, D., Tom, C., Mercer, E. H., Anderson, D. J., Mayford, M., Kandel, E. R., and Tonegawa, S. (1996). Subregion- and cell type-restricted gene knockout in mouse brain. *Cell*, 87:1317–1326.

Tulving, E. (1972). Episodic and semantic memory. In Tulving, E. and Donaldson, W., editors, *Organization of memory*, pages 381–403. Academic Press, New York.

Tulving, E. (1995). Organization of memory: Quo vadis? In Gazzaniga, M. S., editor, *The cognitive neurosciences*, pages 839–847. MIT Press, Cambridge, MA.

Tulving, E., Habib, R., Nyberg, L., Lepage, M., and McIntosh, A. R. (1999). Positron emission tomography correlations in and beyond medial temporal lobes. *Hippocampus*, 9:71–82.

Tulving, E., Hayman, C. A., and Macdonald, C. A. (1991). Long-lasting perceptual priming and semantic learning in amnesia: a case experiment. *J Exp Psychol Learn Mem Cogn*, 17:595–617.

Tulving, E. and Markowitsch, H. J. (1998). Episodic and declarative memory: role of the hippocampus. *Hippocampus*, 8:198–204.

Uretsky, E. and McCleary, R. A. (1969). Effect of hippocampal isolation on retention. *J Comp Physiol Psychol*, 68:1–8.

Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., and Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science*, 277:376–380.

Verfaellie, M. and Cermak, L. S. (1994). Acquisition of generic memory in amnesia. *Cortex*, 30:293–303.

Verfaellie, M., Koseff, P., and Alexander, M. P. (2000). Acquisition of novel semantic information in amnesia: effects of lesion location. *Neuropsychologia*, 38:484–492.

Verfaellie, M., Reiss, L., and Roth, H. L. (1995). Knowledge of New English vocabulary in amnesia: an examination of premorbidly acquired semantic memory. *J Int Neuropsychol Soc*, 1:443–453.

Vizi, E. S. and Kiss, J. P. (1998). Neurochemistry and pharmacology of the major hippocampal transmitter systems: synaptic and nonsynaptic interactions. *Hippocampus*, 8:566–607.

Wagner, A. D., Stebbins, G. T., Masciari, F., Fleischman, D. A., and Gabrieli, J. D. (1998). Neuropsychological dissociation between recognition familiarity and perceptual priming in visual long-term memory. *Cortex*, 34:493–511.

Warrington, E. K. and McCarthy, R. A. (1988). The fractionation of retrograde amnesia. *Brain Cogn*, 7:184–200.

Whishaw, I. Q. (1989). Dissociating performance and learning deficits on spatial navigation tasks in rats subjected to cholinergic muscarinic blockade. *Brain Res Bull*, 23:347–358.

Whishaw, I. Q., Cassel, J. C., and Jarrard, L. E. (1995). Rats with fimbria-fornix lesions display a place response in a swimming pool: a dissociation between getting there and knowing where. *J Neurosci*, 15:5779–5788.

Whishaw, I. Q. and Jarrard, L. E. (1996). Evidence for extrahippocampal involvement in place learning and hippocampal involvement in path integration. *Hippocampus*, 6:513–524.

Wible, C. G., Findling, R. L., Shapiro, M., Lang, E. J., Crane, S., and Olton, D. S. (1986). Mnemonic correlates of unit activity in the hippocampus. *Brain Res*, 399:97–110.

Wickelgren, W. A. (1979). Chunking and consolidation: a theoretical synthesis of semantic networks, configuring in conditioning, S–R versus congenitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychol Rev*, 86:44–60.

Wiener, S. I. (1993). Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task. *J Neurosci*, 13:3802–3817.

Wiener, S. I., Paul, C. A., and Eichenbaum, H. (1989). Spatial and behavioral correlates of hippocampal neuronal activity. *J Neurosci*, 9:2737–2763.

Wiig, K. A., Cooper, L. N., and Bear, M. F. (1996). Temporally graded retrograde amnesia following separate and combined lesions of the perirhinal cortex and fornix in the rat. *Learn Mem*, 3:313–325.

Wilson, H. R. and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys J*, 12:1–24.

Wilson, M. A. and McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, 261:1055–1058.

Wilson, M. A. and McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265:676–679.

Winocur, G. (1990). Anterograde and retrograde amnesia in rats with dorsal hippocampal or dorsomedial thalamic lesions. *Behav Brain Res*, 38:145–154.

Winocur, G., McDonald, R. M., and Moscovitch, M. (2001). Anterograde and retrograde amnesia in rats with large hippocampal lesions. *Hippocampus*, 11:18–26.

Wirsching, B. A., Beninger, R. J., Jhamandas, K., Boegman, R. J., and El-Defrawy, S. R. (1984). Differential effects of scopolamine on working and reference memory of rats in the radial maze. *Pharmacol Biochem Behav*, 20:659–662.

Wood, E. R., Dudchenko, P. A., and Eichenbaum, H. (1999). The global record of memory in hippocampal neuronal activity. *Nature*, 397:613–616.

Wyss, J. M. and Van Groen, T. (1992). Connections between the retrosplenial cortex and the hippocampal formation in the rat: a review. *Hippocampus*, 2:1–11.

Xu, L., Anwyl, R., and Rowan, M. J. (1998). Spatial exploration induces a persistent reversal of long-term potentiation in rat hippocampus. *Nature*, 394:891–894.

Yamamoto, C. (1982). Quantal analysis of excitatory postsynaptic potentials induced in hippocampal neurons by activation of granule cells. *Exp Brain Res*, 46:170–176.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *J Exp Psychol Learn Mem Cogn*, 20:1341–1354.

Young, B. J., Otto, T., Fox, G. D., and Eichenbaum, H. (1997). Memory representation within the parahippocampal region. *J Neurosci*, 17:5183–5195.

Zalutsky, R. A. and Nicoll, R. A. (1990). Comparison of two forms of long-term potentiation in single hippocampal neurons. *Science*, 248:1619–1624.

Zhang, K. (1996). Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J Neurosci*, 16:2112–2126.

Zipser, D. (1985). A computational model of hippocampal place fields. *Behav Neurosci*, 99:1006–1018.

Zola, S. M., Squire, L. R., Teng, E., Stefanacci, L., Buffalo, E. A., and Clark, R. E. (2000). Impaired recognition memory in monkeys after damage limited to the hippocampal region. *J Neurosci*, 20:451–463.

Zola-Morgan, S., Squire, L. R., and Amaral, D. G. (1986). Human amnesia and the medial temporal region: enduring memory impairment following a bilateral lesion limited to field CA1 of the hippocampus. *J Neurosci*, 6:2950–2967.

Zola-Morgan, S., Squire, L. R., and Amaral, D. G. (1989a). Lesions of the hippocampal formation but not lesions of the fornix or the mammillary nuclei produce long-lasting memory impairment in monkeys. *J Neurosci*, 9:898–913.

Zola-Morgan, S., Squire, L. R., Amaral, D. G., and Suzuki, W. A. (1989b). Lesions of perirhinal and parahippocampal cortex that spare the amygdala and hippocampal formation produce severe memory impairment. *J Neurosci*, 9:4355–4370.

Zola-Morgan, S., Squire, L. R., Rempel, N. L., Clower, R. P., and Amaral, D. G. (1992). Enduring memory impairment in monkeys after ischemic damage to the hippocampus. *J Neurosci*, 12:2582–2596.

Zola-Morgan, S. M. and Squire, L. R. (1990). The primate hippocampal formation: evidence for a time-limited role in memory storage. *Science*, 250:288–290.