Hippocampal Conjunctive Encoding, Storage and Recall: Avoiding a Tradeoff

Randall C. O'Reilly
&
James L. McClelland

Technical Report PDP.CNS.94.4
June 1994

Parallel Distributed Processing and Cognitive Neuroscience

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA

Western Psychiatric Institute and Clinic
University of Pittsburgh
Pittsburgh, PA

Neural and Behavioral Sciences
University of Southern California
Los Angeles, CA

MRC Applied Psychology Unit
Cambridge, England

Abstract

The hippocampus and related structures are thought to be capable of: 1) representing cortical activity in a way that minimizes overlap of the representations assigned to different cortical patterns (pattern separation); and 2) modifying synaptic connections so that these representations can later be reinstated from partial or noisy versions of the cortical activity pattern that was present at the time of storage (pattern completion). We point out that there is a tradeoff between pattern separation and completion, and propose that the unique anatomical and physiological properties of the hippocampus might serve to minimize this tradeoff. We use analytical methods to determine quantitative estimates of both separation and completion for specified parameterized models of the hippocampus. These estimates are then used to evaluate the role of various properties and of the hippocampus, such as the activity levels seen in different hippocampal regions, synaptic potentiation and depression, the multi-layer connectivity of the system, and the relatively focused and strong mossy fiber projections. This analysis is focused on the feedforward pathways from the Entorhinal Cortex (EC) to the Dentate Gyrus (DG) and region CA3. Among our results are the following: 1) Hebbian synaptic modification (LTP) facilitates completion but reduces separation, unless the strengths of synapses from inactive presynaptic units to active postsynaptic units are reduced (LTD). 2) Multiple layers, as in EC to DG to CA3, allow the compounding of pattern separation, but not pattern completion. 3) The variance of the input signal carried by the mossy fibers is important for separation, not the raw strength, which may explain why the mossy fiber inputs are few and relatively strong, rather than many and relatively weak like the other hippocampal pathways. 4) The EC projects to CA3 both directly and indirectly via the DG, which suggests that the two-stage pathway may dominate during pattern separation and the one-stage pathway may dominate during completion; methods the hippocampus may use to enhance this effect are discussed.

Introduction

It is well accepted that the hippocampus and related structures are critically involved in memory. However, it is not yet well understood exactly what role they play. We follow Marr (1969, 1970, 1971) and many others (Wickelgren, 1979; Teyler & Discenna, 1986; Sutherland & Rudy, 1989; Rolls, 1990; Squire, 1992; Schmajuk & DiCarlo, 1992; Gluck & Myers, 1993; Humphreys, Bain, & Pike, 1989; Damasio, 1989) in proposing that the hippocampus can be understood as part of a dual memory system consisting of cortical and hippocampal components (McClelland, McNaughton, O'Reilly, & Nadel, 1992; McClelland, McNaughton, & O'Reilly, submitted). In brief, we propose that the cortex is responsible for developing stable, efficient, and general representations of the world, while the hippocampus is responsible for storing the contents of specific episodes or events (i.e., particular states of the world). The critical distinctions between these two tasks are the temporal duration underlying the formation of the representations, and the relationship between other representations in the system. The hippocampus must form and store its representations rapidly (in order to bind together temporally coincident events), while the cortex must form and store its representations very slowly in order to capture the relevant general structure common to different samples of the environment. Representations in the hippocampus must be kept distinct, since very similar episodes often need to be distinguished (i.e., where one parked one's car today is not necessarily the same place as yesterday). In contrast, for the cortex to exploit the shared structure present in ensembles of events and experiences, it must assign similar internal representations to similar events, and to do so it must make the representations overlap.

Thus, the role in our theory of the hippocampus as a memory system can be stated quite simply: in addition to the rapid storage of similar patterns without undue interference, the hippocampus must be capable of using partial or possibly noisy cues to retrieve previously stored patterns, so that memories may be later accessed. Thus, the hippocampus must perform *pattern separation* at the time of storage, which makes the stored patterns more distinct from each other, and *pattern completion* at the time of recall in order to recover the full stored pattern from a partial retrieval cue.

The motivation for the present work comes from the realization that pattern separation and completion are at odds with each other. To the extent that the system takes similar input patterns and separates them, it will form distinct new memories. However, this will work against the completion process, which requires that an overlapping input pattern trigger the recall of an existing memory instead of the creation of a distinct new one. Thus, to be useful, a memory system like the hippocampus must have ways of dealing with this tradeoff between pattern separation and completion. Our hypothesis is that some of the unique anatomical and physiological properties of the hippocampus can be understood as ways of minimizing this tradeoff.

Our investigation follows in the tradition of what McNaughton has termed the "Hebb-Marr" model (Hebb, 1949; Marr, 1969, 1970, 1971; McNaughton & Morris, 1987; McNaughton & Nadel, 1990). This model provides a framework for associating functional properties of memory with the mechanisms of pattern separation, learning (synaptic modification), and pattern completion. Further, it relates these mechanisms to underlying anatomical and physiological properties of the hippocampal formation. Under this model, the two basic computational structures in the hippocampus are the feedforward pathway from the Entorhinal cortex (EC) to area CA3, which is important for pattern separation and pattern completion, and the recurrent connectivity within CA3, which is primarily important for pattern completion. The model relies on the sparse, random, projections in the feedforward pathway from the EC to the Dentate and CA3, coupled with strong inhibitory interactions within DG and CA3, to form sparse, random, and conjunctive repre-

sentations (i.e., each active unit reflects the influence of a conjunction of active units in the input). These representations overlap less that the EC input patterns that give rise to them — in some cases, as we shall see, this pattern separation effect can be very dramatic.

We develop a set of analytical models that build upon the principles of the feedforward component of the Hebb-Marr model and include several important and previously unexplored features of the hippocampus. Other researchers have developed analytical and simulation models that have explored some aspects of pattern separation (e.g., Torioka, 1978, 1979; Gibson, Robinson, & Bennett, 1991). The key features that these approaches share with our own are the explicit consideration of input pattern overlap as an independent variable in the evaluation of pattern separation, and the use of networks that combine the assumption of sparse, random projections with an idealization of the combined effects of feedforward and lateral inhibition called the "k-Winners-Take-All" ($kWTA$) assumption. According to this assumption, feedforward and lateral inhibition work together so that only a roughly constant number (k) of neurons in a given region which receive the strongest excitatory input become active. These features lead to pattern separation, which we give an intuitive as well as formal treatment of based on hypergeometric probability distributions.

Pattern completion occurs in both the feedforward and recurrent components of the Hebb-Marr model. In the feedforward case, completion can occur by way of a variable inhibitory threshold that depends on the total amount of activity in the input pattern. This threshold allows the full activity pattern to be active upon presentation of a partial input cue because the threshold is lower for partial input patterns (McNaughton & Nadel, 1990). While this mechanism will work perfectly for orthogonal stored patterns, it breaks down with increasing overlap, causing erroneous units to become active. Pattern completion via the recurrent connectivity occurs through a settling process, which results in the progressive cleanup of a partial cue pattern to a stored attractor pattern. As many authors have noted, this recurrent projection is probably used for auto-associating activity patterns within CA3 (e.g., McNaughton & Morris, 1987; Rolls, 1989). Many useful results have already been obtained through analytic studies of recurrent auto-associative networks (e.g., Treves & Rolls, 1991, 1992; Gibson & Robinson, in press; Amit, Gutfreund, & Sompolinsky, 1987; Hopfield, 1982, 1984).

In the context of the extensive literature on recurrent pattern completion, our strategy has been to focus on the relatively neglected feedforward pathway, and to relate the findings we obtain to relevant findings in that literature. To summarize briefly, the recurrent auto-associative completion depends on how close the probe pattern is to the stored memory; therefore, it is useful to do as much pattern completion as possible in the feedforward system, to maximize retrieval from the system as a whole. At the same time, maximal separation of different patterns is necessary to avoid spurious blending of attractors in the recurrent pathway. Thus, we conclude that it is useful for the feedforward pathway to do as much of both completion and separation as is possible, to optimize the overall performance of the pattern retrieval system.

Having a concrete computational framework for examining the influence of various anatomically and physiologically related variables on both pattern separation and pattern completion, we are able to quantitatively evaluate the ways in which the hippocampus might avoid or minimize the effects of the separation/completion tradeoff. Our results indicate that certain properties of the hippocampus make sense when viewed in terms of improving the characteristics of this tradeoff.

Hippocampal Anatomy and Physiology

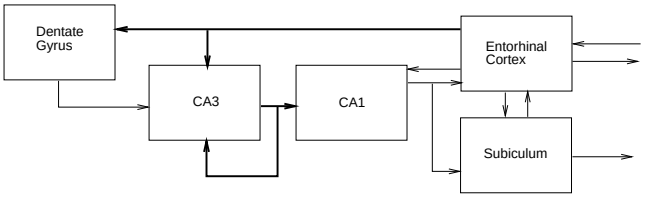

Figure 1: Schematic diagram of the regions of the hippocampus. In this paper, we focus on the feedforward pathway from the entorhinal cortex to the Dentate and the CA3.

In this section, we review aspects of the hippocampal anatomy and physiology relevant for our model. First, we present a functional account of what the different regions of the hippocampus and related structures are doing during storage and recall of memories. Then, we discuss some anatomical and physiological properties that motivate the subsequent modeling. For the functional account, we briefly review the relevant features of the hippocampus. The hippocampus proper consists of a set of interconnected regions known as the Dentate Gyrus (DG), and the fields of Ammon's Horn, which are principally the CA3 and CA1 (see figure 1). Both the DG and the CA3 receive input from the Entorhinal Cortex (EC) via the perforant path projections. CA1 also receives input from the EC, but via a different projection arising from a different layer of the EC (Tamamaki, 1991). CA3 receives mossy fiber inputs from the DG, and it also has recurrent collaterals that interconnect neurons within the CA3. CA1 receives projections from the CA3 via the Schaffer collateral pathway.

Functionally, we think of the hippocampal system as performing the task of an auto-association network that is capable of recalling from partial cues prior activity patterns over both the Entorhinal Cortex (EC) and the Subiculum. These regions provide both input into the hippocampus proper, and output from the hippocampus to the rest of the brain via extensive bi-directional connectivity with wide areas of the neocortex[1] (Van Hoesen & Pandya, 1975a; Van Hoesen, Pandya, & Butters, 1975; Van Hoesen & Pandya, 1975b; Insausti, Amaral, & Cowan, 1978; Van Hoesen, 1982). Thus, memory retrieval in the hippocampus amounts to performing pattern completion over the EC and Subiculum, which in turn trigger the reinstatement of activity patterns in the neocortex that reflect the content of the original memory. For the remainder of the paper, we consider only the EC, but assume that some of what is said applies to the Subiculum as well.

The auto-associative function of the hippocampus as a whole could theoretically have been implemented directly in the EC via recurrent collaterals similar to those in the CA3, without the need for any of the "additional" circuitry of the hippocampus proper. However, since this is not the case, it is likely that the circuitry of the hippocampus provides some advantages over a more direct auto-associator. The hypothesis that we explore in this paper is that this "additional" circuitry is needed to perform the pattern separation function. In particular, we analyze the role of the Dentate Gyrus (DG) and area CA3 in separating activity patterns coming from the EC. We consider CA3 to be the locus of storage for the pattern-separated hippocampal representations, while the DG is thought to assist in the encoding and recall of these representations.

The use of a memory representation in the CA3 that is different from that in the EC introduces a new problem: how can the CA3 representation be used to re-activate the original EC representation in order to perform pattern completion during recall? This is particularly problematic because the CA3 representation, by virtue of the pattern separation process, should have little direct correlation

[1]The Subiculum is apparently more of an output system than an input system, although it does provide inputs into area CA1

with the EC pattern that it represents. Thus, we view the CA1 as being a "translator" that forms an association between the CA3 representation and the EC representation. For this purpose, the CA1 representations are thought to be stable, relatively sparse (but not as sparse as the CA3), and most importantly, invertible.

The overall scenario for how memories are encoded in the hippocampus under this account is as follows:

- An EC representation of the patterns of activity throughout many regions of the neocortex is formed via projections from these areas.
- A distinct, pattern-separated representation of the EC activity pattern is formed in area CA3, with the help of the DG.
- Simultaneously, another representation of the EC activity pattern is formed in the CA1 via direct projections from the EC. This CA1 representation is invertible, in that it can reproduce the corresponding EC representation.
- The link between the CA3 and CA1 representations is forged via learning that occurs on the projections from CA3 to CA1. This allows the CA3 activity pattern to later be able to activate the corresponding CA1 activity pattern, which can in turn activate the original EC representation.
- Learning also occurs both in the feedforward pathways to CA3 from the EC, and in the recurrent collaterals within the CA3 itself. This learning enables partial input patterns to trigger pattern completion of the CA3 representation during recall.

There are several particular features of the hippocampus which we find to be important for understanding how it carries out its function. We describe these features in the context of the rat hippocampus.

- In addition to the principal excitatory neurons within each region of the hippocampal system, there are also inhibitory interneurons. Both of these neuron types typically receive excitatory projections from the other regions, but only the excitatory neurons project out of the region. Thus, the inhibitory neurons form local feedback circuits that probably serve to regulate activity levels in the system (McNaughton & Morris, 1987).
- There are distinctive differences in activity levels in the different regions of the hippocampus (Figure 2 shows data from Barnes, McNaughton, Mizumori, Leonard, and Lin (1990)). In particular, the DG seems to have an unusually sparse level of activity, but CA3 and CA1 are also less active than the input/output layers, EC and Subiculum.
- The perforant path is a broad, diffuse projection originating in layer II of the EC. Each DG granule cell receives roughly 5,000 synaptic inputs (Squire, Shimamura, & Amaral, 1989), and each CA3 cell receives from 3,750 to 4,500 synaptic inputs from the EC (Amaral & Claiborne, 1990; Brown & Zador, 1990). This amounts to approximately 2% of the roughly 200,000 layer II EC neurons in the rat (Squire et al., 1989).
- The DG has roughly 4-6 times the number of excitatory neurons as the other regions of the hippocampus, with roughly $1x10^6$ granule cells in the rat, (Boss, Peterson, & Cowan, 1985), as compared to an estimated 160,000 CA3 pyramidal neurons and 250,000 in CA1 (Squire et al., 1989; Boss, Turlejski, Stanfield, & Cowan, 1987).

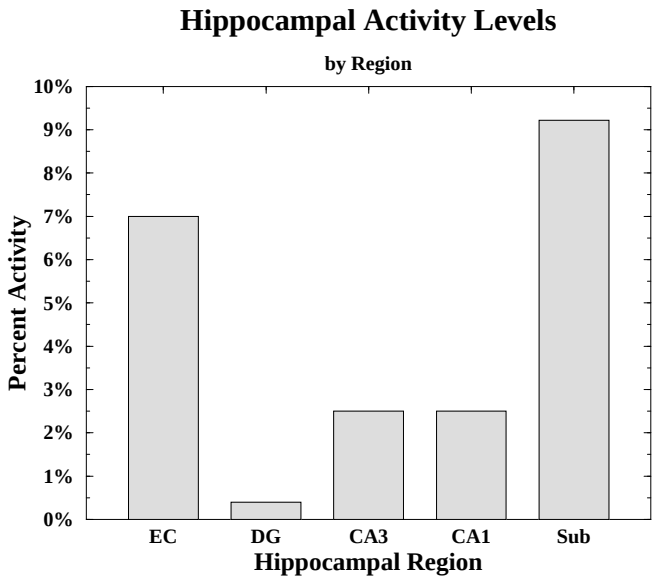

Figure 2: Activity levels in the various regions of the hippocampus, computed as mean firing rate divided by maximum firing rate, which gives percentage of neurons firing at maximum rate. In general, the inputs to the hippocampus (EC, DG) are more active, while the layers inside the hippocampus (DG, CA3, CA1) are less active. The DG figure, which is an estimate (B.L. McNaughton, personal communication), indicates a much sparser activity level than any of the other areas. Data for other regions from Barnes, McNaughton, Mizumori, Leonard, and Lin (1990)

- The projection between the DG and CA3, known as the mossy fiber pathway, is distinctive in several ways. It is sparse, focused, and topographic ("lamellar") projection (Squire et al., 1989). Each CA3 neuron receives only around 52-87 synapses from this projection (Claiborne, Amaral, & Cowan, 1986), but each synapse is widely believed to be significantly stronger than the perforant path inputs to CA3 (McNaughton & Nadel, 1990) since they terminate close to the soma, and are relatively large synapses (Brown & Johnston, 1983; Yamamoto, 1982; Brown, Wong, & Prince, 1979). However, the exact magnitude of the mossy fiber strength is not known.

- The projections from the EC to the DG and CA3 are strictly feedforward—no direct feedback from these regions to the EC are known to exist (McNaughton & Nadel, 1990).

- Associative, NMDA-dependent Long Term Potentiation (LTP) has been demonstrated in both the perforant pathway and the Schaffer collateral pathways (McNaughton, 1983; Brown, Kairiss, & Keenan, 1990). LTP has also been demonstrated in the mossy fiber pathway (Barrionuevo, Keslo, Johnston, & Brown, 1986), but it is not NMDA dependent (Harris & Cotman, 1986). It is not known if the LTP in the mossy fibers is associative (i.e., that both pre and postsynaptic activity is necessary), but many think it is not (Brown et al., 1990).

- In addition, evidence indicates that an associative Long Term Depression (LTD) phenomenon might be taking place in these pathways as well (Levy & Desmond, 1985; Levy, Colbert, & Desmond, 1990).

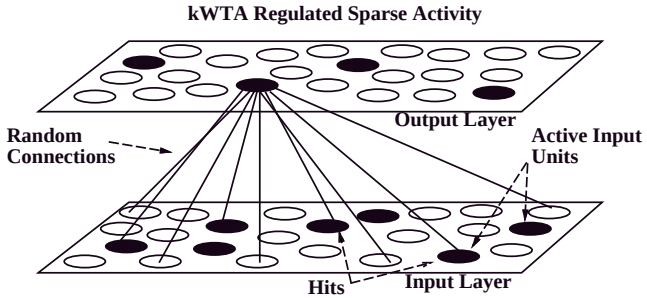

Figure 3: Elements of the two-layer analytical model with random connectivity and k-Winners-Take-All ($kWTA$) activity. Excitatory input comes into the output layer units from the active input layer units via sparse random connections. The k output layer units with the most input (i.e., greatest number of "hits," where a weight connects from an active input unit) actually become active, and their activity suppresses the less active units (e.g., via lateral inhibitory connections mediated by inhibitory interneurons, though these are not implemented in the model).

Assumptions of the Model

The present paper focuses on the feedforward perforant and mossy fiber pathways and the areas they connect, including the EC, DG, and CA3. As such, we base our analytical model primarily on an abstraction of the perforant path connectivity. The basic building blocks in our model consist of two layers of units, an input layer and an output layer, and the connectivity between them (see figure 3). Each output unit has a fixed number, F, of incoming synapses from the input layer, where F is smaller than the number of units in the input layer, N_i. Thus, each output unit is partially connected to the input layer. The pattern of connectivity is assumed to be completely random, which is an approximation to the diffuse pattern of connectivity in the perforant path [2].

The activity regulation that appears to be operative in the hippocampus, which is attributable to the action of inhibitory interneurons as described above, is captured in our analytical model by introducing a competitive k-Winners-Take-All ($kWTA$) type of activation function. If the $kWTA$ approximation is correct, then around k neurons (out of the entire output-layer population) will be active at any given point. These k neurons can be thought of as those that have a level of excitatory input exceeding the inhibitory input from the interneurons. Thus, one can think of the inhibitory input as a floating threshold for activation. This $kWTA$ approximation can be relaxed by allowing fewer than k units to become active when the input to the system is weak, resulting in a k-or-less WTA function. This corresponds to using an inhibitory threshold that has a fixed lower bound, but can float above this lower bound when the input is strong enough. The implications of this will be explored later in the paper.

The basic two-layer model is specified by four essential parameters: N_i, which is the total number of excitatory input neurons; k_i, which is the number of those input neurons that are active at any one time (we treat k_i as a constant for complete input patterns, which assumes a $kWTA$ activation function on the input layer); α_o, which is the percent activity in the output layer (equivalent to $\frac{k_o}{N_o}$, but expressed as a percentage for reasons that will become clear later); and the fan-in F. Given this parameterization, the excitatory input to an output unit will be a function of the number

[2]This approximation is essential for the analytical model, but simulation models without this constraint, and analyses of the relatively focused mossy fiber pathway, indicate that some level of ordering of the connections (i.e., connectivity density determined by a Gaussian distribution with a σ of .25 in units of a half-width of the input layer) does not affect our results significantly.

Area	α	N	F_{EC}	F_{DG}
EC	6.25%	200,000	—	—
DG	0.39%	850,000	4,006	—
CA3	2.42%	160,000	4,003	64

Table 1: Parameters used in the analytical models. F_{EC} is the fan-in from the EC, and likewise F_{DG} is the fan-in from the DG. The particular fan-in sizes reflect an attempt to achieve a consistent output activity level given the round-off errors associated with the use of an integer threshold. The estimated DG activity level is from B.L. McNaughton, personal communication, while the other values are referenced in the section on hippocampal anatomy and physiology.

(out of F) of "hits" (active units) on its input connections. For our purposes, we assume that the number of hits, denoted H_x (where x indicates which input pattern generated the hits), sufficiently determines which output units will be active in the $kWTA$ competition, so there is no need to introduce into the model an activation function that transforms raw input (H_x) into an equivalent of neural membrane potential.

Our analytical framework centers around the computation of the conditional probability of an output unit being active for a particular input pattern (e.g., pattern B) given that it was previously active for an input pattern (A) that overlapped to some degree with B. This framework is then extended to account for various conditions such as: learning that took place on the active and/or inactive input lines for the previous pattern; partial input patterns; the presence of multiple layers of processing; and variable activation thresholds. In modeling the multi-layer input to CA3 that comes via the DG, we introduce the problem that the connectivity between the DG and CA3, the mossy fiber pathway, is focused, and not diffuse. It happens that this does not affect our model significantly, for reasons which are discussed in a later section of the paper.

The patterns of activity in area CA3 are the main focus of our investigation. Even though the CA3 has two sources of input, the direct perforant path inputs from EC and the mossy fiber inputs from DG, we can begin by analyzing a simplified system involving a CA3 having only perforant path inputs and no mossy fiber inputs. Such a system is useful for determining the relative effect of the mossy fiber inputs, and the computations are much simpler and easier to understand. We refer to such a system as a "monosynaptically connected CA3" or more simply, a "monosynaptic CA3." Subsequently, we extend the model to incorporate both the indirect Dentate projection as well as the monosynaptic EC projection.

Finally, to represent the general order of the anatomical properties of the hippocampus, we have listed a set of numerical values corresponding to the activity levels, numbers of excitatory neurons, and numbers of synaptic inputs per neuron (fan-in) by hippocampal region for the rat in table 1. We refer to these numbers as the "rat-sized" hippocampus. The choice of numbers reflects a balance between values cited in the section on hippocampal anatomy and physiology and several computational constraints, including: avoiding round-off problems due to the use of an integer-valued activity threshold; avoiding future round-off problems when scaling these values down for simulations; and minimizing computation time (i.e., there was a bias towards the smaller end of the range). The effects of the parameters on the model's behavior are well understood, and are discussed in what follows where relevant.

Pattern Separation

Pattern separation is studied by evaluating the effect of input pattern overlap on output pattern overlap. This can be reduced to its simplest form by using two probe stimuli, A and B, which overlap with each other by a specified amount (i.e., they share some percentage of active input units). If A is presented to the network and the active output units are recorded (the o_a units), and then B is presented and the corresponding active units are recorded as well (the o_b units), then one can compute the output overlap as the intersection between the o_a and o_b sets of units. Pattern separation is the degree to which this output overlap, denoted Ω_o is less than the input overlap Ω_i.

It has been known since Marr (1969) that circuitry like that in the hippocampus could lead to pattern separation effects, and pattern separation has been studied with more sophisticated analytical techniques since then (Torioka, 1978, 1979; Gibson et al., 1991). Since we use a different analytical framework for exploring pattern separation than was used in previous approaches, we present our model in detail below. However, before doing so, we present the central intuition behind the pattern separation effect. In the $kWTA$ function, the active (or "winning") output units for any given pattern x are those that have a number of hits in the upper tail of the hit probability distribution ($P(H_x)$), above the inhibitory threshold (H_x^t). Note that this threshold is like the θ parameter in Marr's (1969) "codon" model. Thus, the units that have a potential to be active for both patterns must first come from the tail of the hit distribution for pattern A (i.e., $P(H_a) >= H_a^t$). For these units to be active for pattern B, they have to again be in the tail for the distribution of hits in pattern B. However, since this second distribution is derived from the tail of the first, it differs systematically from it based on the degree to which A and B overlap. We give this second distribution, $P(H_b|P(H_a) > H_a^t)$, the shorthand label $P(H_{b|a})$.

One difference between the $P(H_a)$ and $P(H_{b|a})$ distributions is that, as overlap increases, the mean of the $P(H_{b|a})$ distribution moves roughly proportionally from the mean for the $P(H_a)$ distribution (for low overlap equivalent to the activity level of the input layer) to slightly above the threshold H_a^t for the $P(H_a)$ distribution (for fully overlapping patterns)[3] This upward shift reflects the increased probability of the output unit getting hits from the units active in pattern B due to the increasing likelihood of the same units that were hits in A being hits again in B as B overlaps more with A.

In addition, the $P(H_{b|a})$ distribution becomes narrower as overlap increases. Again, this derives from the fact that the source of the $P(H_{b|a})$ distribution is the tail of the $P(H_a)$ distribution, which is obviously narrower than the entire $P(H_a)$ distribution. Increasing overlap moves $P(H_{b|a})$ from something that looks like $P(H_a)$ for low overlap to something that looks like the tail of $P(H_a)$ for high overlap, resulting in a narrowing of the distribution. Both the mean-shift and narrowing effects are illustrated in figure 4, which shows the original $P(H_a)$ distribution and three example $P(H_{b|a})$ distributions corresponding to low, medium and high overlap of B with A.

Both the mean-shift and narrowing effects contribute to the increasingly non-linear relationship between output pattern overlap and input pattern overlap. For all but very high levels of overlap, the threshold is in the tail of the $P(H_{b|a})$ distribution. Since this has a concave shape, an upward shift of this distribution increases the area above the threshold disproportionately less than the shift of the mean. Since this shift of the mean is roughly proportional to the overlap of the input patterns, the output overlap increases at a slower rate than the input overlap. Also, the $P(H_{b|a})$ distribution narrows with increasing input overlap so that even less area of the distribution is above

[3]Since the distributions for the fully overlapping input patterns are the same, the distribution of $P(H_{b|a})$ for 100% overlap is just the tail of the $P(H_a)$ distribution, and its mean is the mean of the tail.

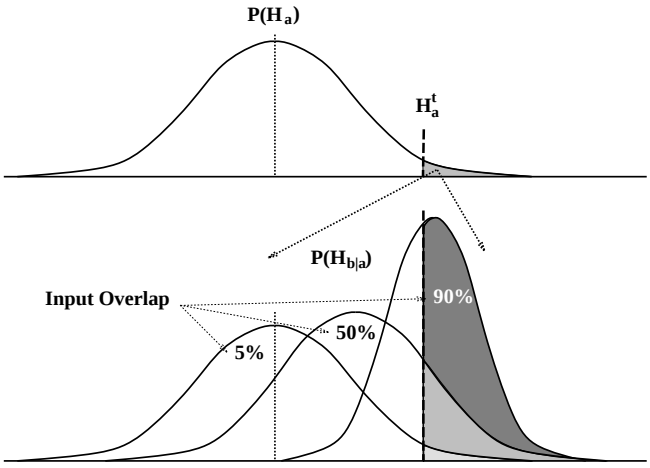

Figure 4: Representation of the effect of increasing input overlap on the probability distribution $P(H_{b|a})$, which is derived from the tail of the distribution $P(H_a)$. As input overlap increases, the distributions get narrower and the mean shifts upwards towards the threshold. These changes interact with the concave shape of the the distribution to produce a level of output overlap that is lower than the input overlap, resulting in pattern separation. Actual distributions shown are based on the hypergeometric model described in the text.

the threshold until quite high levels of input overlap are reached.

Given that the mean-shift and narrowing distribution effects are consequences of the activity threshold being in the concave tail of the hit distribution, the critical network parameters that would affect pattern separation are the level of output activity, which determines how far out in the tail of the distribution the threshold is set, and factors which determine the overall shape of the probability distribution, such as fan-in and number of units. Given that the regions of the hippocampus seem to vary systematically along the dimensions of output activity levels, numbers of units, and fan-in size, we would expect that different areas have different pattern separation properties. In particular, we would expect that the DG, with its very low output activity levels, would have very good pattern separation, relative to the CA3 or CA1, but these areas would in turn be better than the EC or the Subiculum.

Hypergeometric Model

Our analytical model of pattern separation is based on hypergeometric probability distributions, which is to say that we explicitly count the various ways of producing different numbers of input hits in order to compute their probabilities. There are two stages of calculation for the basic model of pattern overlap in a two-layer system. The first is to determine the appropriate threshold corresponding to a desired level of output activity given parameters such as the input activity level, the fan-in for each output unit, etc. The second is to compute the conditional probability that an output unit will be active (above threshold) for input pattern B given that the unit was active for input pattern A, when B overlaps with A by a specified proportion. This conditional probability is equivalent to the expected level of output pattern overlap for the specified amount of input pattern overlap.

The hypergeometric model can be illustrated with a Venn-diagram representation, as shown in Figure 5. This figure shows the space of input units with two subsets, one representing the input units active for pattern A and the other representing those input units that a given output unit

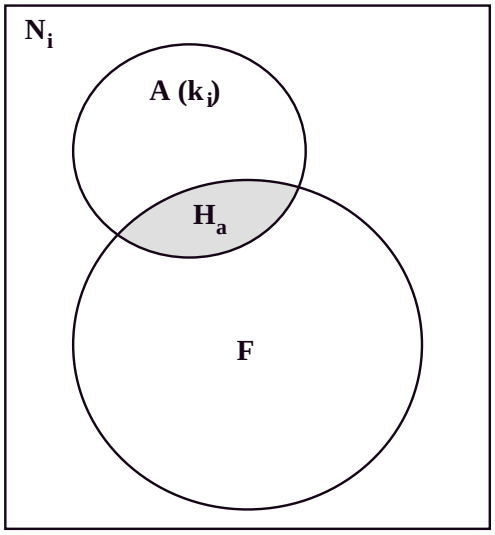

Figure 5: Venn-diagram representation of the hypergeometric model showing the units in the input layer. The overall space indicated by the rectangular box represents the set of input units, of size N_i. A subset of these units of size k_i are activated by pattern A. In addition, a given output unit with fan-in of size F is connected to a second subset of input units. The hits for the output unit are at the intersection of the fan-in and the activity subsets. The size of this intersection is given by the variable H_a. Note that this representation is not intended to be topological, as it merely represents the set-wise division of input units.

receives projections from (i.e., its fan-in). The input set is of size N_i, while the activity subset is of size k_i, and the fan-in subset is of size F. Initially, before any learning takes place, each input weight to the output unit has a value of 1. The hits for the output unit are at the intersection of the two subsets, the size of which is indicated by the variable H_a. The probability of an output unit receiving a particular number of hits is given by the hypergeometric distribution.

The hypergeometric is based on the idea of sampling (without replacement) from an environment containing two kinds of things, for example a barrel with red and blue balls. Given a certain sample size, and a specified number of red and blue balls in the barrel, the hypergeometric gives the probability of getting a specific number of red and blue balls in the sample. Thus, we can think of the fan-in F from a unit being a "sample" of the input space having both active and inactive units. There are k_i active units, and $N_i - k_i$ inactive units, and we want the probability of getting exactly H_a active units in the fan-in sample.

The logic of the expression for the hypergeometric is relatively simple, and we will be generalizing this logic to deal with more than one distinction in the environment for subsequent expressions, so we review it here. If one did happen to get H_a of the active inputs in the fan-in "sample", then there must have been $F - H_a$ of the inactive inputs in the sample. The hypergeometric computes the number of ways of getting a specific configuration of active and inactive inputs using the product of the number of ways of independently choosing H_a items from a population of size k_i, and $F - H_a$ items from a population of $N_i - k_i$. This total number of configurations that would lead to H_a hits and $F - H_a$ inactive inputs is then divided by the (larger) number of ways of picking any F items out of a population size N_i without regard to which are active and inactive. This results in

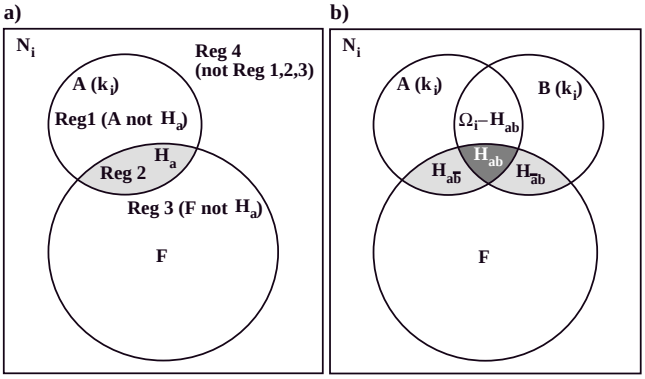

Figure 6: **a)** The four regions of the input-unit space where pattern B could intersect, with the input space defined as in the previous figure for input pattern A and a given output unit. **b)** Quantification of the numbers of units in activity pattern B from each of the four regions shown in part a. There are Ω_i units in common with A, and H_{ab} hits from the same H_a inputs that were active in A, and $H_{\bar{a}b}$ hits unique to pattern B.

an overall probability for getting H_a hits:

$$\mathcal{H}(H_a, k_i, F, N_i) = \frac{\binom{k_i}{H_a}\binom{N_i - k_i}{F - H_a}}{\binom{N_i}{F}} \tag{1}$$

where each term in the expression $\binom{N}{m}$ gives the number of ways of choosing m items from a population of N. Thus, the expression for the probability of getting H_a input hits is: $P(H_a) = \mathcal{H}(H_a, k_i, F, N_i)$.

Determining the kWTA Threshold

Using the hypergeometric, it is possible to describe the probability distribution over the range of possible number of hits for a given size network. In order to use the probability distribution to derive a threshold number of hits H_t for the $kWTA$ function, we simply sum over the upper tail of the distribution until the total probability is equal to the desired output activity level:

$$\alpha_o = \sum_{H_a = H_t^a}^{MIN(k_i, F)} P(H_a) \tag{2}$$

where α_o represents the proportion of output units active, not the number. The value of H_t^a as defined in this equation can be calculated from a given α_o by computing the sum backwards starting from the maximum number of hits (which is the smaller of k_i and F) and stopping when the sum equals or exceeds the desired α_o, with the H_a value at that point becoming H_t^a. The threshold is qualified by the pattern because its value will change when learning is introduced later.

Computing Output Pattern Overlap

Output pattern overlap is measured by presenting a pattern B that overlaps with A. Thus, we think in terms of constructing B by choosing bits of it from each of the different possible regions in the input space (shown in figure 6a) as defined by a pattern A presented previously, where the output unit in question had H_a hits on A. Any randomly selected pattern B will have a particular number (possibly 0) of units active in each of these regions. The variables used to represent these numbers are shown in figure 6b, with pattern B constrained to have Ω_i overlap with A. The regions are as follows:

Region 1: Those units which were active in pattern A, but not among those that were hits. The size of this region is $k_i - H_a$, and any pattern B will have $\Omega_i - H_{ab}$ units in this region.

Region 2: Those units which were hits for pattern A, size H_a. B will have H_{ab} units from this region.

Region 3: Those units which were in the fan-in F but not in pattern A, size $F - H_a$. B will have $H_{\bar{a}b}$ units from this region (in B but not A).

Region 4: Those units not in any of the above regions, size $N_i - (k_i - H_a) - H_a - (F - H_a)$, which simplifies to $N_i - k_i - F + H_a$. B will have the remaining $k_i - (\Omega_i - H_{ab}) - H_{ab} - H_{\bar{a}b}$, which simplifies to $k_i - \Omega_i - H_{\bar{a}b}$ units from this region.

Following the logic of the hypergeometric function as described previously, we can express the probability of obtaining any particular pattern B as the product of the number of ways of getting the specified number of units of B independently from each of the 4 regions divided by the number of ways of choosing any pattern B having Ω_i overlapping units with pattern A (which can be expressed in terms of choosing Ω_i units out of A, and $k_i - \Omega_i$ units out of $N_i - A$):

$$P_b(H_a, \Omega_i, H_{ab}, H_{\bar{a}b}) = \frac{\binom{k_i - H_a}{\Omega_i - H_{ab}} \binom{H_a}{H_{ab}} \binom{F - H_a}{H_{\bar{a}b}} \binom{N_i - k_i - F + H_a}{k_i - \Omega_i - H_{\bar{a}b}}}{\binom{k_i}{\Omega_i} \binom{N_i - k_i}{k_i - \Omega_i}} \tag{3}$$

Equation 3 can be used to compute the probability distribution for hits on pattern B by noting that these hits come from both H_{ab} and $H_{\bar{a}b}$, so that the sum of these two numbers from a particular configuration of B gives the total hits on B, H_b. This summing process can be used both to compute a threshold for pattern B, H_t^b, which may not be the same as H_t^a if learning has occurred, and to compute pattern overlap in the output layer. To compute overlap, we must restrict the summation of probabilities to those configurations of B having a level of hits on A such that $H_a >= H_t^a$. The details of this process are given in Appendix A, the result of which is an expression for output pattern overlap proportion (ω_o) as a function of input pattern overlap Ω_i.

The properties of the $P(H_{b|a})$ distribution as given by this hypergeometric formulation were illustrated in Figure 4. The distributions shown in the figure were generated using the monosynaptically-connected CA3, rat-sized parameters given in table 1. To see more clearly the pattern separation effects that result from the shape of the $P(H_{b|a})$ distribution as input pattern overlap increases, we graph the output overlap (given by equation 8 in Appendix A) as a function of input overlap using the same rat-sized parameters in figure 7. The crucial feature of this graph is that the output overlap falls well below the diagonal line that represents a linear relationship between input and output overlap. This "sub-linear" output pattern overlap with respect to input pattern overlap amounts to pattern separation.

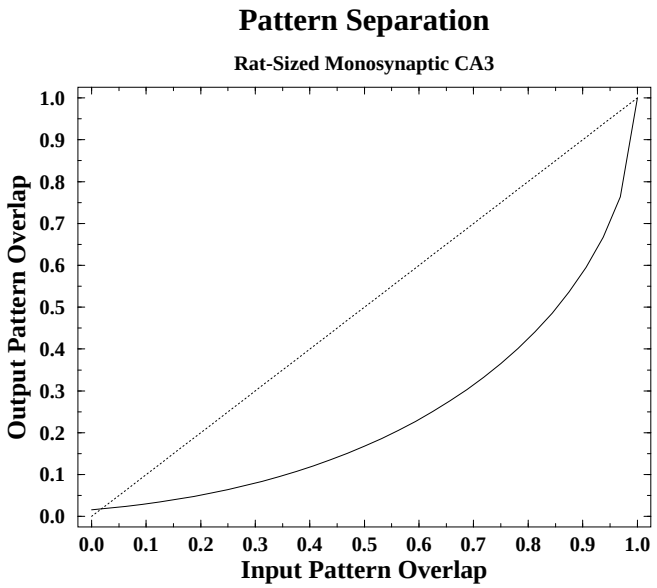

Figure 7: Pattern separation in a rat-sized monosynaptically connected CA3 (see table 1 for parameters). Pattern separation is revealed by the fact that the output pattern overlap falls well below the diagonal line that corresponds to output patterns having the same level of overlap as the input patterns. We refer to this as "sub-linear" output pattern overlap for a given level of input pattern overlap.

While figure 7 gives a sense of the approximate pattern separation capabilities of a CA3-like layer receiving only perforant path inputs, these results are of most value in their comparison with those from other parameters and network configurations, rather than for the actual magnitude of pattern separation at any given point on the curve, since it is difficult to estimate the level of input pattern overlap in the rat's EC representations. Also, it should be noted that similar results to those shown in figure 7 have been obtained by Torioka (1979), and subsequently by Gibson et al. (1991), with a model based on the binomial distribution. However, our ability to extend these results to model several important features of the hippocampus depends critically upon the hypergeometric formulation.

The Role of Output Activity & Fan-in in Pattern Separation

One principal dimension along which the different layers of the hippocampus vary is the mean level of activity (see figure 2), with the layers within the the hippocampus proper (DG, CA3, CA1) being more sparsely active than the input/output areas (EC, Subiculum), and the DG having the lowest activity levels. Using the pattern separation formalism developed above, we can evaluate the quantitative impact of output activity levels typical of the different regions of the hippocampus. Figure 8 shows, as expected, that pattern separation increases with lower output activity levels.

Aside from the activity level, the other parameter in the model that would plausibly have an important effect on pattern separation is the size of the fan-in F to an output unit. Indeed, intuition may suggest that a smaller fan-in will result in better pattern separation, due to a reduced probability of contact with the input activity pattern. However, in the hypergeometric model, F is equivalent to the "sample" size, which actually does not affect the probability distribution very

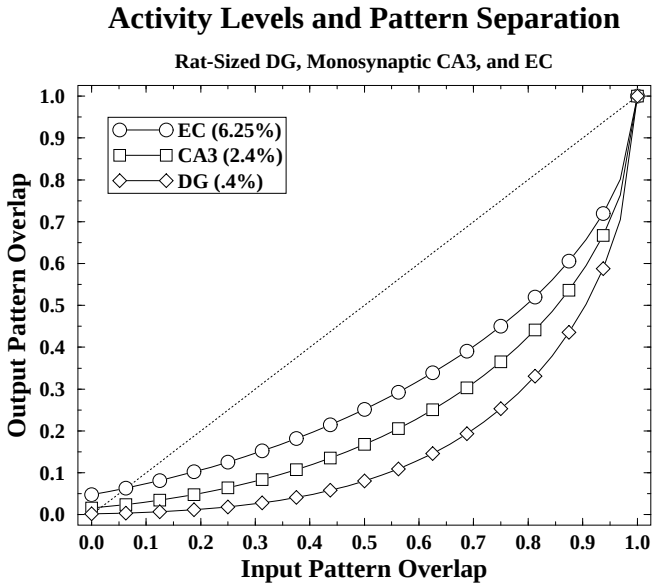

Figure 8: The effect of output layer activity levels (α_o) typical of the EC, DG, and CA3 on pattern separation. Pattern separation is enhanced for sparser activity levels (especially as as input pattern overlap increases) because the threshold is further in the tail of the hit distribution. This makes the DG especially relevant for pattern separation.

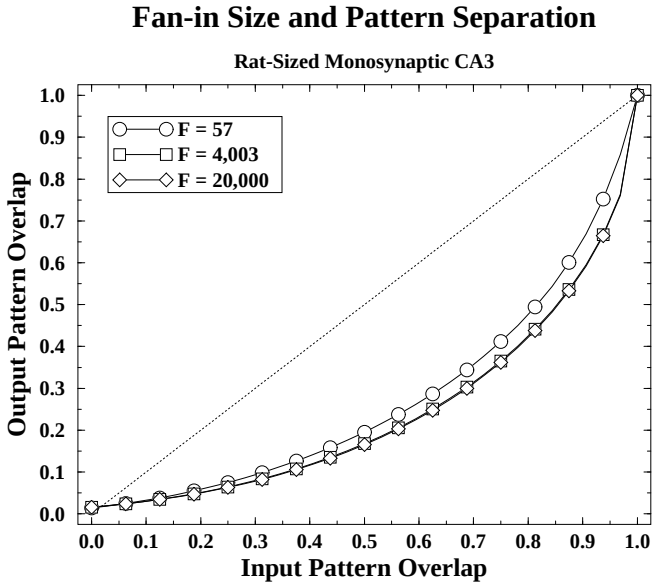

Figure 9: The effect of fan-in (F) on pattern separation, revealing reduced pattern separation for smaller F values, with not much difference between 4,000 and 20,000. The fan-in size of 57 was selected because it produced an output activation level of 2.4%, while other nearby values resulted in elevated or depressed activation levels due to integer round-off effects.

much over broad ranges. Only very small F values or very large F values (i.e., F close to N_i) would give different results. Figure 9 shows the effects of three different fan-in values on the pattern separation effects for a monosynaptically connected, rat-sized CA3 layer. While the small F case shows worse pattern separation, there is not much of a difference between 4,000 and 20,000. We interpret this as indicating that once the "sample" is sufficiently large, increasing it more yields decreasing returns. A smaller fan-in does not improve pattern separation, because the probability distribution becomes skewed and deviates from the bell-shaped distributions typical with a larger F (this will be discussed in more detail below).

It is important to remember that, despite these negative results for the effect of fan-in size on pattern separation, there are other properties of the network that the fan-in might plausibly affect, including overall memory storage capacity and the fault and noise tolerance of the representations.

Multiple Stages of Pattern Separation

Since the DG, which has a more separated representation than area CA3 by virtue of its lower activity level, feeds into CA3 via the mossy fiber pathway, there is the potential for a compounding of the pattern separation effect over multiple stages. The potential for such a mechanism is great, since each stage forms the input to the next, and our results indicate that each stage of a DG-like layer produces dramatic pattern separation. However, the most dramatic compounding effects would be seen in only a few layers. For example, an input pattern having 90% overlap would result in the following series of overlaps over succeeding DG-like layers: 50%, 8%, .45%, ...2%. Given that biological network hardware is "expensive," one could easily argue that the two layers of compounded separation evident in the disynaptic pathway from EC to DG to CA3 represents a reasonable cost/performance tradeoff.

However, what is less clear is why the second stage of the hippocampal pattern separation system has direct, monosynaptic connections from the EC input layer via the perforant pathway. The mystery of the connectivity into area CA3 is a complex issue, and we will explore several different reasons for having direct EC input into area CA3 later in the context of pattern completion. For the time being we assume that there is a good reason for the perforant path input to CA3, and focus on the effects of such input on multiple stage pattern separation. In particular, we address the following two questions: How strong does the mossy fiber input have to be in order for CA3 to exhibit the compounding effect? and Why is the mossy fiber input sparse and strong as opposed to many and weak, like the perforant path input?

In order to address these questions, pattern separation for units in area CA3 is computed using the combination of two expressions like the one used previously (equation 3). The first step is to compute the expected output pattern overlap for the DG, using the single-stage model previously described. Then, the CA3 output pattern overlap is computed by combining the EC monosynaptic hit distribution with another monosynaptic hit distribution computed using the DG pattern overlap as the input overlap probability. The F, N_i, and k_i parameters for each of these inputs correspond to the perforant path and mossy fiber pathways, respectively. In addition, a new variable M is introduced which multiplies each hit from the mossy fibers, allowing the differential strength of this pathway to be represented. A single configuration of the multi-input CA3 network thus consists of a particular input pattern A_{EC} from the EC, and another A_{DG} from the DG. Since these are independent events, the probability for this configuration is the product of the two independent probabilities, and the resulting total number of hits is the sum of the two individual numbers of hits. The same summing and thresholding techniques that were applied previously can be applied

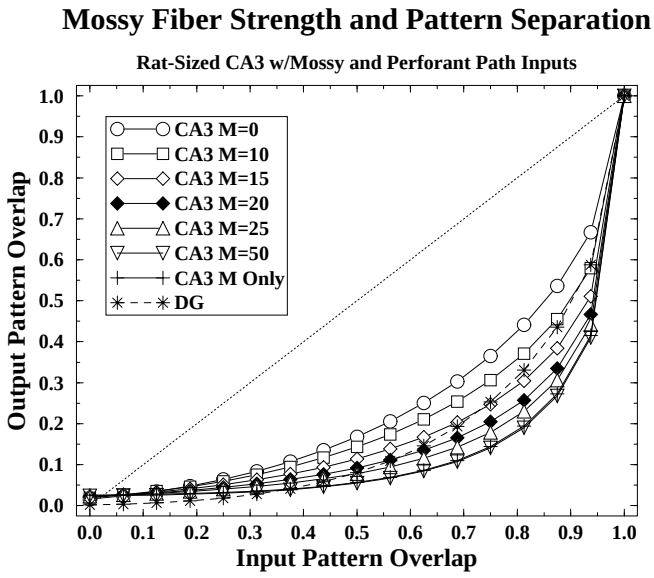

Figure 10: The effect of different strength mossy-fiber inputs into CA3 from the DG ($F_{DG} = 64$) on pattern overlap in the CA3, which is also receiving perforant path input from the EC. The strength of each mossy fiber connection relative to a perforant path connection is given by the value of M, where $M = 0$ is equivalent to the monosynaptically-connected CA3. $MOnly$ means that only mossy fiber inputs were used, which indicates the greatest amount of compounded pattern separation. The DG output overlap curve shows that when M is greater than roughly 15, a compounding of pattern separation is occurring (since these curves fall below the output overlap on the DG).

to the resulting CA3 total input probability distribution.

The Strength of DG Inputs into CA3

It is reasonable to assume that the balance between the strength of inputs from the EC and DG to CA3 neurons should determine the extent that each affects the firing properties of the CA3. The stronger the input from the DG, the more pattern separation should be observed due to the compounding effect, while weaker DG input will allow the EC inputs to dominate, resulting in the level of pattern separation shown in the previous figures for the monosynaptically connected CA3. This suggests that the DG input should be strong relative to the EC input, which is apparently the case with the mossy fiber input to CA3 as reviewed previously.

However, our results regarding the fan-in size shown in figure 9 indicate that a larger fan-in yields a greater pattern separation effect. Thus, one might expect there to be many mossy fiber inputs into each CA3 neuron, but this is not the case. On the contrary, only about 52-87 mossy fibers synapse on a given CA3 cell (Claiborne et al., 1986). Further, if we estimate the balance of the mean number of hits for this few mossy fiber inputs relative to the perforant path inputs, we find that the EC input gives over 1,000 times as many hits as the DG input (using the rat-sized parameters from table 1 for the EC input with 6.25% activity on 4,006 input weights, the mean number of hits is around 244, and for the DG input with .4% activity on 64 input weights, the mean is .23). While the mossy fibers are probably stronger than the perforant path ones, the difference is more plausibly on the order of tens of times stronger, not thousands.

Paradoxically, both the intrinsic pattern separation properties of a small F and its weakness

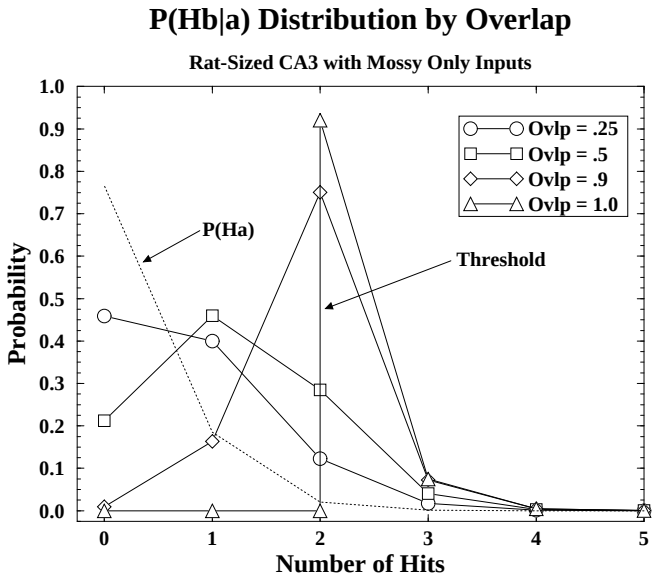

Figure 11: Distributions of $P(H_{b|a})$ for different levels of overlap in a CA3 with mossy fiber only ($MOnly$) inputs. This figure is analogous to the bottom part of figure 4, but the curves are distinctly not bell-shaped due to the low N of the hit distribution. The general effect of the small N is a relatively wide variance, and a skewed upper tail. The $P(H_a)$ distribution is shown for reference (remember that $P(H_{b|a})$ is derived from the tail of $P(H_a)$).

relative to the EC input into CA3 would seem to be working against the sparse mossy fiber inputs. However, figure 10 shows that even with a mossy strength of 10 ($M = 10$ in the figure, where M is the strength of the mossy fiber inputs relative to the perforant path), the DG input has an effect on pattern separation in CA3. Further, a mossy fiber strength of 50 gives the same degree of pattern separation as a system having only mossy fiber inputs ($MOnly$ in the figure) even though the difference in mean number of hits in this case still favors the perforant path input by a factor of 20. Finally, this level is better than the pattern separation on the DG, indicating a compounding effect.

To explain this result, we recall that the pattern separation effect comes from hits in the tails of the probability distribution, since it is the elements in the upper tail that participate in the output activity for the patterns, given the $kWTA$ competitive mechanism. It follows that the distribution that contributes the most to the upper tail will be the one which most influences the pattern separation properties of the active units. When summing two probability distributions, as in the case with the two EC and DG hit distributions, the distribution that has the greatest variance will be the one that contributes the most to the tails of the summed distribution, regardless of its mean relative to the other distribution.

Thus, we can estimate the effect of mossy fiber strength on pattern separation in area CA3 by comparing the magnitude of the *variances*, not the *means*, of the two input distributions. For a .25 input overlap level on the DG and the EC, the standard deviation of the DG hit distribution (shown in figure 11) is .76, while the EC standard deviation is 15. Thus, a mossy fiber strength of around 20 would equalize the variances of the two distributions, giving them equal influence over pattern separation on CA3. Under such conditions, one would expect to find the pattern separation curve roughly midway between the perforant-path only curve ("CA3 M=0" in figure 10), and the Mossy Only curve. This appears to be the case given the $M = 20$ curve shown in the figure.

Many and Weak vs *Few and Strong Mossy Fibers*

There is a continuum of parameters in the model that will produce the high levels of variance in the distribution of hits from the DG inputs to CA3 necessary to retain the compounding effect of pattern separation as discussed above. The relevant variables for determining the variance are the size of the sample (i.e., F), the activity level of the input layer (α_i) and the weight multiplier of each mossy input (M). The mean number of hits is approximately $F\alpha_i$, and the standard deviation (σ) of the hit distribution is:

$$\sigma \approx \frac{M}{\sqrt{F\alpha_i}} \tag{4}$$

while the mean input μ to the neuron is:

$$\mu \approx MF\alpha_i \tag{5}$$

One way to interpret the relationship between M, F and the variance and mean of the hit distribution is that, for a fixed standard deviation level σ, a small F and a big M result in less mean input μ to the unit than a big F and a small M. With .25 input overlap for example, in the case with $F = 64$ and $M = 10$, $\sigma = 7.6$, and $\mu = 7.0$, but when $F = 4,000$ and $M = 2$, $\sigma = 8.0$ and $\mu = 32$. The smaller level of excitatory input could be biologically relevant given that extensive circuitry and numbers of inhibitory interneurons are required to control the activity levels of excitatory neurons. Further, each axonal process has space and metabolic costs associated with it, so it might be more efficient for the nervous system to use the small F, big M of the mossy fiber pathway to obtain sufficient variance in the hit distribution.

A further motivation for preferring the few strong mossy fibers comes from the skew evident in the hit distribution as shown in figure 11. By comparing the standard deviation of the $P(H_a)$ distribution (.48) with that of the $P(H_{b|a})$ distribution at input overlaps of .25 and .5 (.76 and .81 respectively) one can see that this skew causes an increase in variance with increasing input overlap. In contrast, the standard deviations of the more normal distributions for the EC inputs decrease with increasing overlap as a result of the narrowing of the distribution. Thus, the relative impact of the mossy fiber pathway as measured by variance nearly doubles as input overlap increases to .5, after which point it decreases again, to a level of .53 for the .9 input overlap case.

Thus, the size and strength of the mossy fiber pathway is appropriate for passing the pattern separation benefits of the DG on to the CA3. Indeed, for mossy strengths greater than 15 in our model, there is a compounding of pattern separation so that the CA3 actually produces more separated representations than the DG. It is also interesting to note that if the direct perforant path input was not present, the issue of the balance of variance would not be relevant, and the small fan-in and large strength of the mossy fiber pathway would be more difficult to understand in such a system.

Finally, one might wonder why, if the mossy fiber strength is such that the CA3 simply has the pattern separation level of the DG, the hippocampus does not simply have a monosynaptically connected CA3 with the activity level of the DG, thereby doing away with the dentate entirely? One answer is that a sparse activity level interferes with the ability of the recurrent connections in a CA3-like layer to perform auto-associative pattern completion. Assuming that the number of synaptic connections in the CA3 recurrent collateral pathway (on the order of 10,000) represents an upper limit, then an activity level on the order of .4% would amount to an average of only 40 active synaptic inputs, compared to the roughly 250 active inputs from the EC input. It might be difficult for this few of recurrent inputs to perform the pattern completion function in the face of the

stronger feedforward input. Interestingly, with a 2.5% activity level, the CA3 recurrent collaterals deliver an average of 250 active inputs, which is at the same level as the feedforward input.

Pattern Completion

Having established in the previous section that effective pattern separation should occur in area CA3 given its connectivity with the EC and DG, we can now ask how likely it is that a partial "cue" input pattern will be able to reinstate a previously-stored CA3 output representation. In other words, we want to know how well the system will be able to remember stored memories. There are effectively two different forms of recall cues that can be used, partial cues and noisy cues. A partial cue is simply a subset of the original activity pattern, while a noisy one is a subset plus some extra noise that could come from outside the original activity pattern. As it happens, the noisy cue is exactly what we have been exploring as the B pattern above. Thus, we know that noisy cues will tend to engage the pattern separation properties of the hippocampal feedforward circuit. Indeed, we have discovered that a noisy cue having 90% signal and only 10% noise would have a less-than 50% output overlap with the original pattern!

However, if the pattern of activity on the EC during recall were to be just a subset of the original pattern, there is a possibility that this might instead engage a pattern completion process. Thus, activity level might be a factor which decides between separation and completion, allowing for a reasonable overall tradeoff between the two. This kind of partial input cue could result from activity in the EC being regulated by a *k or less* WTA function, where the threshold is relatively more fixed, so that weak inputs to the EC result in fewer active units there. We can investigate this kind of cued-recall using the analytical methods developed for pattern separation by simply composing a pattern B which has the overlapping component as defined before, but not the unique component.

Completion and the Effects of Learning

Central to the mechanism of pattern completion is the initial storage of memory representations. Clearly, a neural system which does not adapt its weights upon presentation of items to be stored will not be likely to recall these items at a later point. Figure 12 shows that without learning, a monosynaptically connected, rat-sized area CA3 has a pattern completion function which resembles its pattern separation function plus a constant offset.

Pattern completion was computed by eliminating from the expression for pattern separation the terms involving the non-overlapping portion of a pattern B. The denominator term in the resulting hypergeometric is just the number of ways the pattern of size Ω_i can be drawn from a pool of k_i units. Thus, the probability that any cue pattern B having overlap Ω_i and hits in the overlapping region H_{ab} is:

$$P_b^C(H_a, \Omega_i, H_{ab}) = \frac{\binom{k_i - H_a}{\Omega_i - H_{ab}}\binom{H_a}{H_{ab}}}{\binom{k_i}{\Omega_i}} \tag{6}$$

The tail of this distribution can be summed as before, and a conditional probability of re-activation to pattern B given activity to pattern A computed. However, a novel difficulty with this completion formula is that the input pattern B has a different number of active units depending on the overlap. This differential input activity leads to a round-off problem given that the threshold is an integer

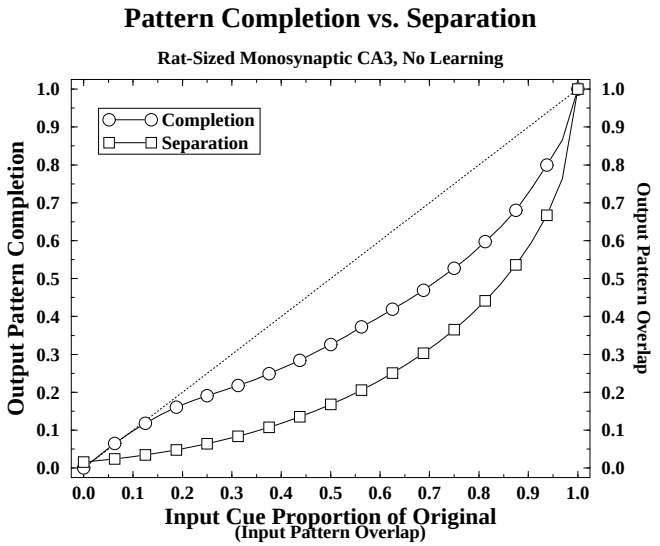

Figure 12: Pattern completion without learning resembles pattern separation plus a relatively constant offset. Note that the X-axes are comparable in the two cases, in that the completion case is the same as the separation one except that the non-overlapping parts are absent in the former case. The Y-axes are equivalent, in that overlap and completion amount to the same thing in this case. Results shown are for a monosynaptically-connected rat-sized CA3.

value. Even with the large N of the rat-sized model, the round-off problem leads to variations in actual output activity level large enough to affect the output probabilities significantly. To counteract this problem, a constraint-satisfaction based interpolation algorithm was developed (see Appendix B), and all completion graphs are of the interpolated data.

The effect of learning on pattern completion can by studied in our analytical model as long as simple learning rules are employed. Since all of the calculations involve computing the number of hits from different regions of the input space (i.e., as defined in the Venn diagrams in figures 5 and 6), it is possible to treat learning as a matter of re-weighting these hits by varying amounts depending on what region they come from. We initially consider a weight increase only (WI) learning rule that resembles associative LTP, a form of long-term synaptic modification which has been found in many regions of the hippocampus (McNaughton, 1983; Brown et al., 1990). Under this rule, an output unit which is active for pattern A will increase its weights to all input units which were also active. All other weights remain the same. In terms of the completion function just described, this weight increase will only affect those hits coming from the overlapping region, which are H_{ab} in number. Thus, in the process of tabulating the total number of hits, H_{ab} is multiplied by a learning rate factor $1 + L_{rate}$.

Increasing the value of each hit from the overlapping region during pattern completion will enhance the probability of completion. Figure 13a shows the effect of different learning rates on completion in a rat-sized CA3. Completion goes up with the learning rate, and is quite substantial even with relatively small learning rates. However, the gains made in pattern completion with increase-only learning have a detrimental effect on pattern separation, as can be seen in Figure 13b. Since the overlapping hits H_{ab} are each magnified by the learning, increasing overlap results in an increasing probability of reactivation with B for units that were active for pattern A.

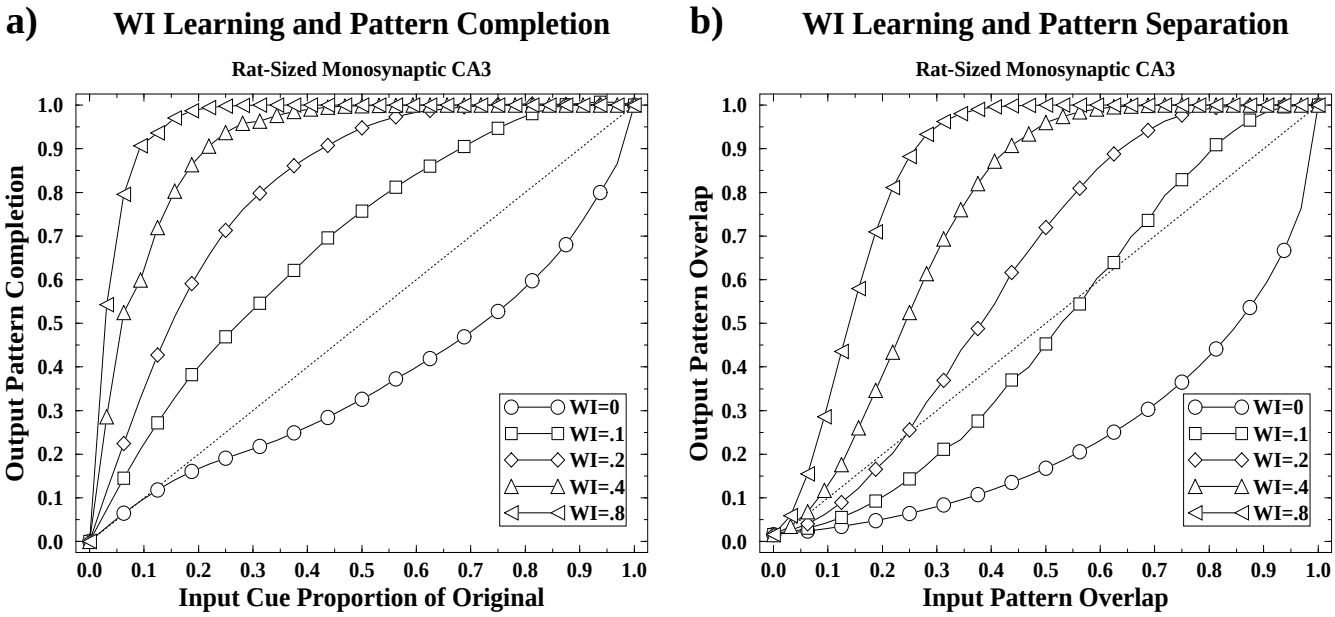

Figure 13: a) Pattern completion with weight increase-only learning (WI) shows substantial benefits with increasing learning rate, in that the curves extend well above the diagonal line that indicates the same amount of output pattern completion as is already present on the input. b) Pattern separation with WI learning shows a critical erosion of pattern separation with increasing learning rate. As in the previous figure, the separation function resembles the completion one, except now they both favor completion instead of separation as before. Both graphs are for a rat-sized monosynaptic CA3.

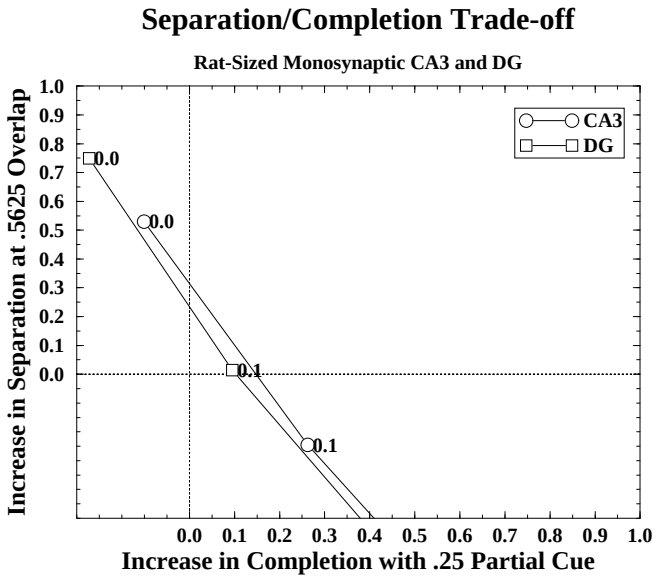

Figure 14: The trade-off between pattern separation and completion as a function of different learning rates (0, .1, with the points for .2, .3, .4, .8 being out of range for this graph) using increase-only learning. Gains in completion through learning are completely offset by losses in pattern separation. Further, activity level appears to hurt completion more than it helps separation, as is evident by the DG line being below the CA3 one. Values are plotted by computing the difference from the input value (overlap or separation) as a proportion of the maximum possible improvement. Reference points of .5625 overlap and .25 partial cue were chosen to maximally differentiate effects of learning rates.

Exploring the Trade-off Between Separation and Completion

The initial results with learning and its effects on completion and separation seem to substantiate the notion of a trade-off between the two. In order to capture this trade-off more directly, we employ graphs like figure 14, which shows the separation and completion data plotted against each other, making the trade-off clearly evident. The points with no learning show high separation but poor completion, and increasing the learning rate up through .8 trades increased completion for poorer separation (only .1 appears in the graph, with the remaining points resulting in a loss in separation performance to big to fit on the graph).

In the trade-off graph, separation and completion are evaluated at fixed points of the input variable (either .5625 overlap for the separation case, or .25 partial cue for the completion case). These points were selected to maximize the differentiation between different learning rates on completion and separation, and not for any particular *a priori* belief about the relevant input parameters of the hippocampus. Thus, they should not be interpreted as an absolute measure of performance, but should instead be used for comparing the relative effects of different manipulations on the trade-off function. The trade-off function used provides the clearest picture of the qualitative relationship between separation and completion, but the actual performance could be better or worse for particular manipulations when evaluated at different points along the individual separation and completion curves. However, rarely if ever will it be the case that one condition that looks better than another in our trade-off function is actually worse for other points along the completion or separation curves, since the ordinal relationship between curves tends to remain constant for the

different levels of overlap or sizes of partial cue.

Perhaps the most interesting effect shown in figure 14 is that the lower activity level of the DG, which results in better separation than the CA3, actually results in a worse overall trade-off due to impaired pattern completion. Thus, we can conclude that lower activity levels impair completion more than they enhance separation, at least under these conditions. Further, the higher learning rate necessary for the DG function to get to the same level of completion as the CA3 function is also a problem, since a larger learning rate leads to greater interference with previously stored representations (a point that will be elaborated below).

Increase and Decrease Learning

One approach to improving the trade-off between separation and completion is to modify the learning rule. Since completion improves by increasing the weights to overlapping input units, this aspect of the learning rule must be retained. However, completion would be unaffected if we simultaneously reduced the weights to inactive input units for the active output units. This would cause units which were active for pattern A to be less likely to become reactivated for B when the hits from B come from outside of the original A pattern (the non-overlapping region, region 3, refer to figure 6), and pattern separation would be improved.

The Weight-Increase/Decrease (WID) learning rule has the potential to improve the trade-off between separation and completion because it does not affect the benefits of learning for completion, while at the same time it enhances separation. However, it makes further assumptions about the kinds of synaptic changes taking place in the hippocampus. In particular, it requires a heterosynaptic Long-Term-Depression (LTD) phenomenon, which has been described in the hippocampus (Levy & Desmond, 1985; Levy et al., 1990), but is still the subject of some debate. Nevertheless, our results indicate that such a learning rule would be effective in avoiding the separation/completion trade-off.

In terms of the analytical model, WID learning is implemented in a similar way as WI learning, which is to multiply the number of hits coming from a particular region by the learning rate parameter. In this case, we add to the WI learning rule by also multiplying the $H_{\bar{a}b}$ hits by $1 - L_{rate}$ when tabulating the probability distribution. Thus, H_{ab} hits are increased by the same amount as the $H_{\bar{a}b}$ hits are decreased. Given that the same learning rate is used for both cases, one would expect that where most of the hits are coming from the unique region (i.e., when input overlap is less than 50%), the LTD-like weight decrease learning would make units which were active for A less likely to be active for B. When most of the hits come from the overlapping region (i.e., when input overlap is greater than 50%), the LTP-like weight increase learning will enhance output overlap. This results in a threshold-based solution to the completion/separation trade-off in terms of input overlap, so that separation occurs for patterns with overlap lower than the threshold, and completion occurs above the threshold. When both increase and decrease learning have the same rate parameter, the threshold is at around 50% input overlap, but it is possible to move this threshold up or down by using different L_{rate} parameters for weight increase and decrease.

Figure 15a shows the thresholding effect of WID learning. As expected, the threshold for pattern separation is centered around 50% input overlap, so that separation is enhanced below this level, and completion is favored above it. Plotting this data against the corresponding completion data in the trade-off format (figure 15b) shows that there are regions of the learning rate parameter for which the trade-off is more optimal than others. In addition, the pattern separation advantage for the DG is now preserved for the lower learning rates, which lends support to the idea that the CA3 should take advantage of the DG input to better its position on the trade-off function. This

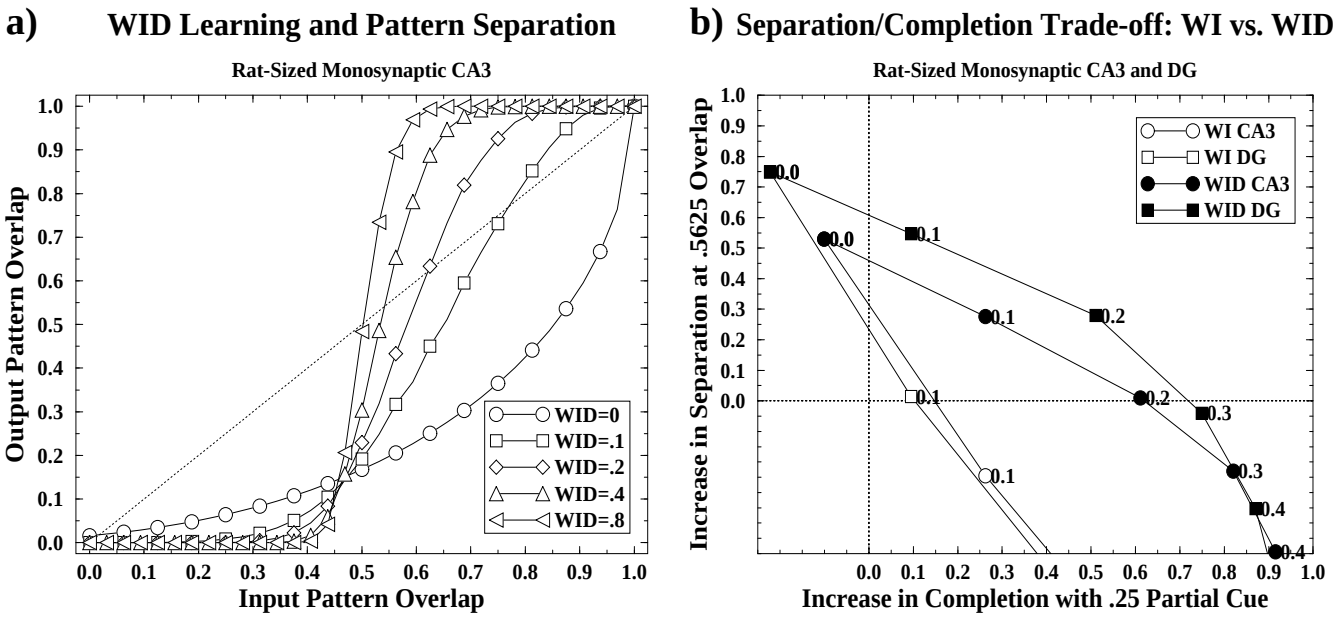

Figure 15: a) Pattern separation with weight increase and decrease learning (WID) shows a threshold for pattern separation at around 50% input overlap, which becomes more pronounced as the learning rate increases. b) The trade-off between pattern separation and completion as a function of different learning rules, weight increase-only (WI) and weight increase-decrease (WID). The WID learning rule preserves the pattern separation advantage for the DG over CA3 while retaining comparable levels of pattern completion.

possibility is explored in the next section.

Multiple Stages of Separation and Completion

For pattern separation, multiple layers of units can compound the separation effects, which is consistent with the existence of the EC to DG to CA3 connectivity in the hippocampus. One might assume that a similar kind of compounding effect would occur for pattern completion since a partially completed pattern resulting from a first layer of processing will allow a second layer to complete it further. However, this is not the case. Instead, due to the $kWTA$ competitive activation function, pattern completion in the first layer will result in a *full pattern of activity* which overlaps with the target pattern by the degree of completion, but also has the remainder of the pattern consisting of units that were not active for the original pattern. This is because the $kWTA$ function has a floating threshold, which is lower for partial input patterns that excite the output units less, thus allowing the relatively constant k_o units in the output layer to become active. Of these k_o units, only a portion of them will be the same as those active for the full pattern (the completed or overlapping portion). All of the original units will not be reactivated because the particular subsample of the original input pattern will favor some distributions of input weights more than others, and units which would not have enough hits for the complete pattern can still have enough hits with a partial input pattern to become active.

The units which become active only for pattern B are essentially noise in the pattern completion process. Thus, instead of a more complete partial cue pattern resulting from the first layer of processing, the result is a partially overlapping pattern like those used in the pattern separation analyses. As such, the pattern separation function from the first to the second layer will be highly

sensitive to the noise, resulting in an output pattern in the second layer that has *less* overlap than the pattern on the first layer.

The negative impact of multiple stages of processing on pattern completion may be the key to understanding why the hippocampus has direct inputs from the EC via the perforant path. Without this input, the gains in pattern separation from the two layers of processing could be lost in worse pattern completion performance. Again, we can view this as a trade-off function. One plausible hypothesis about the nature of this trade-off is that the best balance is struck by having a relatively strong multi-layer pathway for pattern separation, but a not insubstantial direct pathway for pattern completion. Thus, the extremes of no direct input or all multi-layer input would be worse than some middle ground between these two alternatives. In the hippocampus, this would amount to the CA3 having strong inputs from the DG but also reasonably influential inputs directly from EC as well.

Alternatively, if the hippocampus was able to regulate the relative strength of the direct vs. multi-stage (DG) inputs based on the need for completion *vs* separation, then the result could be the best of both compounded separation and single-stage completion. Indeed, given that the input pattern characteristics are different for completion and separation (i.e., completion is driven by a partial cue, while separation works on full input patterns), there is reason to believe that the hippocampus could self-regulate the balance between separation and completion. One mechanism by which this might happen is a threshold that is determined by factors other than just the $kWTA$ constraint. For example, as the overall input becomes weaker, the floating threshold will get lower, but it might do so at a slower rate than that which would preserve the full number of active units in the output layer. Thus, fewer units would become activated, which would send a weaker signal to the next layer. However, this signal might be less noisy than with a lower threshold. These issues are investigated below.

As a first step in exploring the possible implications of the direct and multi-stage inputs to area CA3, we examined the interaction between both WI and WID learning and the strength of the mossy pathway input, which was varied from 0 to a reasonably large value (50), with an extreme case where no direct perforant path inputs to CA3 existed at all. The pattern separation figures where generated as described previously, and the extension of our formalism to multi-layer pattern completion is a straightforward application of the same technique. The only difference is that there are now differences between WI and WID learning for pattern completion (these rules give identical completion results in the single-layer case). The difference derives from the fact that the pattern on the first processing layer (DG) has both overlapping and non-overlapping components, which are increased and decreased respectively in the WID rule, whereas the non-overlapping components are not decreased in the WI rule. Functionally, this will cause CA3 completion to be worse for DG completion levels under 50% in the WID case than with WI learning.

Figure 16 shows the trade-off between separation and completion in a system having either mossy fibers with a strength of 0, 15, 25, and 50, or no direct connections between the EC and the CA3 (the $MOnly$ system), with the DG figures included for reference. In apparent contrast to the predictions of the hypothesis that an intermediate mossy fiber strength would result in the best compromise between separation and completion, the mossy-only (i.e., a purely multi-stage system without direct input to CA3) case, because of its enhanced pattern separation, represents the best trade-off. Thus, the penalty for multi-stage completion is not so severe as to eliminate the advantages of multi-stage separation. This is understandable given that once the pattern overlap exceeds 50% after the fist layer of processing (e.g., in the DG), the second layer will tend to perform pattern completion given the thresholded nature of the pattern separation curves under

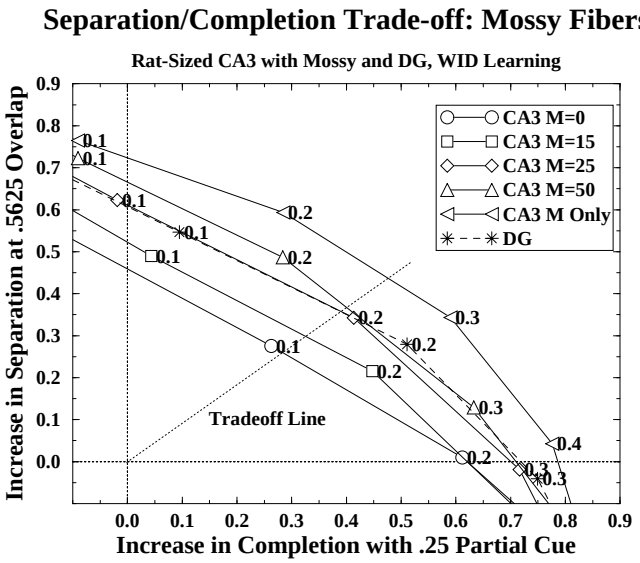

Figure 16: The trade-off between pattern separation and completion for area CA3 with different strengths of mossy inputs, as indicated by the M parameter ($MOnly$ means no direct perforant path inputs were present, mossy fibers only). Even though completion is considerably reduced for a given learning rate as mossy strength increases (i.e., individual points at the same learning rate are shifted to the left), pattern separation enhanced by the two-stage system enough to make for a better overall trade-off, even in the mossy-fiber only extreme. "Tradeoff Line" indicates a possible tradeoff choice between completion and separation, allowing the learning rate required to intersect this line to be compared across M strengths.

WID learning (see figure 15a). Therefore, the ill effects of multiple stages on pattern completion are only evident with very partial input patterns and/or low rates of learning.

One might argue that given a benefit in performance without the perforant path input to CA3, this pathway is not relevant. However, it is important to consider the learning rate necessary to achieve a given level of performance on the tradeoff curve. To compare learning rates across the different tradeoff curves, a possible "Tradeoff Line" is shown in Figure 16, which represents a particular choice about the relative importance of completion vs. separation. The learning rate associated with the point at which a particular tradeoff curve intersects this line reflects the amount of learning needed to achieve the desired balance of completion and separation. For the line shown in the figure, the learning rates increase from a low of around .1 for the $M = 0$ case to around 2.6 for the $MOnly$ case. Since learning is responsible for improving pattern completion, this result follows from the finding that completion is impaired in a multi-layer system: a larger learning rate must be used to attain a sufficient level of completion performance.

If there were no adverse consequences of employing a large learning rate, then one could argue that the mossy-only system is more optimal than a combined direct and multi-layer system. However, it is generally true that the more weights are changed in the system, the more likely it is that existing memories will experience interference, which would limit the ability of the hippocampus to retain memories over time. Since intermediate values of mossy fiber strength, $M = 15$ for example, produce a significant benefit in the trade-off curve but require a lower learning rate than the mossy-only system for equivalent completion performance, they might represent an overall better trade-off than the mossy-only system when the cost of interference is factored in.

Hybrid Systems and Variable Thresholds

As was suggested above, it is possible that the hippocampus can regulate the effective balance between multi-stage and direct inputs into area CA3 by taking advantage of differences in completion vs. separation input patterns. Also, it is possible that differential learning (or lack of learning entirely) in the direct vs. multi-stage pathways can improve the trade-off. These possibilities are referred to collectively as "hybrid systems," in that they represent a combination of two distinct input pathways optimized for completion or separation.

The first hybrid system we consider is the "Mossy For Separation Only" (MSEPO), which uses the mossy fiber inputs to CA3 only for pattern separation, and not for pattern completion. Thus, pattern completion in this system will be like that of the direct pathway system, and separation will be like that of the combined direct and multi-stage system. Obviously, one could consider a hybrid in which, in addition to the MSEPO system, the direct pathway was absent for separation, resulting in even better separation, but such a system requires further assumptions for a biological mechanism. The MSEPO system, on the other hand, has a plausible basis in the physiology of the hippocampus based on the firing properties of the DG excitatory neurons, which drive the mossy fiber inputs to CA3.

The MSEPO system relies on the notion that the low activity levels of the DG are a product of high levels of inhibition within this layer. Further, individual neurons have spiking thresholds that require them to be depolarized a certain amount above the resting potential. With strong inhibitory input resulting from the activity of inhibitory interneurons, it is possible that partial input activity patterns on the EC would be unable to activate the DG neurons above threshold. Thus, partial EC input patterns would only activate the direct path input to CA3, and not the multi-stage mossy fiber inputs because the DG neurons would not be significantly active, resulting in the MSEPO property. This is consistent with the idea mentioned earlier that the $kWTA$ threshold

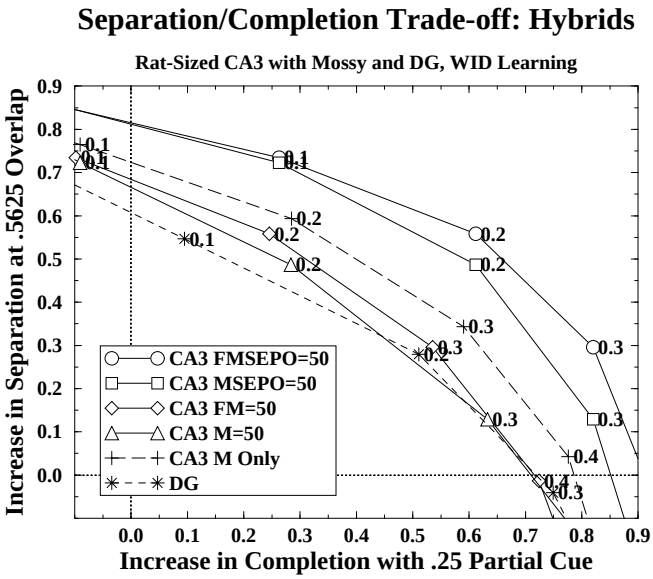

Figure 17: The trade-off between pattern separation and completion for area CA3 with two forms of hybrid systems that enhance multi-stage processing for pattern separation, and/or single-stage perforant path inputs for pattern completion. One hybrid is the "mossy-for-separation-only" case (MSEPO), where the mossy fibers are inactive for completion. The other is the "Fixed Mossy" case (FM), where the mossy fibers do not undergo learning. FMSEPO is the combination of these two, and $M = 50$ is the "control" condition with the same mossy fiber strength (50) as the others. The MSEPO hybrids give better performance at a lower learning rate.

might be determined by factors other than the k active units constraint (e.g., the neuron's leak current). While there would probably be a gradual dropoff in DG activity level with decreasing input activity, we can model the MSEPO system in an all-or-nothing way for simplicity.

The other hybrid system we consider is the "Fixed Mossy" system (FM), which has the mossy fiber weights not subject to learning. Since learning reduces pattern separation, fixing the mossy weights should have the effect of increasing pattern separation relative to a system where they are learned, and since much of the pattern completion is a result of the direct pathway inputs, this should not affect pattern completion significantly. Indeed, if the FM manipulation is combined with MSEPO, then the increase in pattern separation would not produce any effect on completion since mossy fibers are only involved in separation anyway. While LTP has been demonstrated in the mossy fiber pathway (Barrionuevo et al., 1986), it is not NMDA dependent (Harris & Cotman, 1986), and it is not known if it is associative (Brown et al., 1990). Thus, the FM condition represents an optimal extreme that is probably not fully realized in the actual mossy fiber pathway.

Figure 17 shows that the MSEPO and FM manipulations have the predicted effects, with the combined FM and MSEPO (FMSEPO) system producing the best overall trade-off. Further, the MSEPO manipulations have an advantage over the other systems because lower learning rates are required to achieve good performance on the tradeoff function. For example the MSEPO conditions produce maximal separation and completion performance with a .2 learning rate, while the mossy only condition requires a .3 learning rate. The enhanced separation effect from just fixing the mossy fibers is somewhat offset by poorer completion performance, resulting in only a slightly better trade-off. This trade-off is not evident in the MSEPO condition, because the mossy pathway is not involved in completion in this case, resulting in better performance with the fixed mossy

fibers.

The simplified, binary MSEPO hybrid, where the mossy fiber inputs are either active or not, can be generalized somewhat by exploring the effects of allowing a very small fraction of DG neurons to be activated during completion. This would happen if, for example, the threshold for activation in the DG does not float in such a way as to maintain a fixed number of active neurons but instead stays fixed or only adapts a little, so that, when a partial input arises, the number of units that exceeds threshold is less than in the case where a complete input is presented. In this case, although fewer of the DG units that were active when the whole pattern was presented will become active, the probability that those that do become active will have been part of the pattern produced by the complete pattern will increase.

To measure the effect of varying the threshold in our analytical model, we plot the overall probability that a unit active for pattern A is re-activated for partial pattern B ($P(B|A)P(A)$) (i.e., the "signal" strength) and the overall probability that a unit not active in A is activated for pattern B ($P(B|\bar{A})P(\bar{A})$) (i.e., the "noise"), and the probability that, given a unit is active for pattern B, what is the probability it came from those that were active in A ($P(A|B)$) (i.e., the proportion of "signal" in the overall activity pattern). This is shown for both 25% and 90% input partial cue size in figure 18 for a rat-sized DG layer without learning. The overall level of activity corresponds to the sum of the two individual sources of activity ("signal" and "noise"). This figure indicates that, indeed, the DG could send a clearer, though less complete, signal to CA3 if it did not lower the threshold as much for weaker input cues.

Implications of MSEPO for Mossy Fiber Strength

The use of the MSEPO condition has implications for the number and strength of the mossy fiber synapses. During the initial storage of an actvity pattern on CA3, the DG neurons participate in pattern separation, and the mossy fiber input into CA3 neurons, together with the direct perforant path input, plays a role in determining which of these neurons become active. However, if the mossy fibers are not active during recall, then only the perforant path inputs determine which CA3 neurons are active. Thus, learning taking place in the perforant path inputs has to compensate for the absence of the DG input during recall that was present during storage.

If the mossy fiber inputs constitute too much of the total input to the active CA3 neurons, then the learning in the perforant path inputs will not be capable of making up the difference during recall, and pattern completion will suffer. Thus, this is another factor in support of the idea that the mossy fiber pathway should just provide enough variance in the the tail of the CA3 hit distribution to put a subset of neurons that are already receiving a large number of perforant path hits over the activity threshold. One can think of the mossy fiber pathway as selecting a random sample from the population of CA3 neurons that is strongly excited by the perforant path input. Since this only requires a relatively weak level of additional input, learning in the perforant pathway can easily compensate for the absence of the mossy fiber input during recall.

The Consequences of a Focused Mossy Fiber Pathway

One of the assumptions of our analytical model is that the connectivity is diffuse and approximately random. While this might be a reasonable approximation for the perforant path connectivity, it is not so clear that it applies to the mossy fiber pathway, which has a more focused character. However, there are two factors that lead us to conclude that the results of our model presented so far are valid even with the mossy fiber pathway being focused and not random. One factor is that

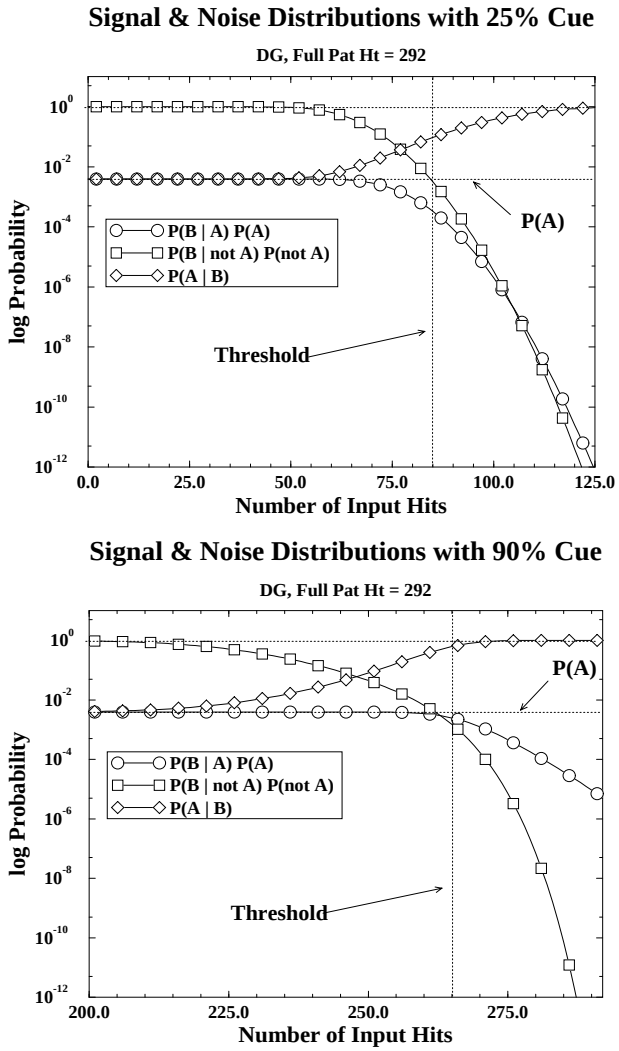

Figure 18: "Signal" ($P(B|A)P(A)$), "Noise" ($(P(B|\bar{A})P(\bar{A})$) and proportion of "Signal" ($P(A|B)$) as a function of the activation threshold in completion for the DG. Values are in log coordinates to provide better resolution. The $kWTA$ threshold is indicated. For small input cues (e.g., 25%, top), a higher threshold would send a clearer but weaker signal to CA3, since the $P(A|B)$ curve continues to get larger with an increasing threshold. This is not the case with more complete input cues as in the case of the 90% cue (bottom), where the proportion of signal is nearly maximal at the $kWTA$ threshold.

the DG activity pattern, having been generated by the competitive activation process based on the random inputs via the perforant path, can be considered to be randomly related to the original EC input pattern. Assuming this, then our model of the DG input to CA3 can treat each DG input unit as an independent random variable with a probability equal to the activity level of the layer.

In the context of a random activity pattern on DG, the effect of a focused vs. diffuse projection to CA3 is to narrow the range of possible inputs any given CA3 neuron could receive. Essentially, a narrow mossy projection amounts to each CA3 neuron sampling from a subspace of the DG smaller than the whole thing. Interpreted in this way, it is possible to determine what the effects of this "mini DG" input to CA3 would be as compared to each CA3 neuron sampling from the entire DG. Intuitively, it might seem that concentrating all of the inputs from a smaller region of the DG would have a significant impact on the probability distributions for hits, but this is not the case. The probability distributions from the entire DG do not differ significantly from those which are obtained when the CA3 has 64 input weights from a sub-space of the DG having only 2,000 units in it (roughly .2% of the DG).

When examined in the context of our analytical model, this result makes sense. Recalling the analogy for the hypergeometric probability distribution in terms of balls in a barrel, the size of the input layer in our model is analogous to the total number of balls in the barrel. However, since the activity level is the same (on average) over the entire input layer, the expected proportion of active to inactive units ("red vs. blue balls") remains the same regardless of the number of total units. While the total number of balls in the barrel is relevant when sampling with replacement, its relevance decreases as the number gets larger (e.g., > 100). Thus, given a sufficiently large space to sample from, the probabilities are determined by the proportion of items in the space more than by the size of the space itself. We assume that in the case of the mossy fiber pathway, the projection is sufficiently wide to allow each CA3 neuron to sample from at least .2% or more of the DG neurons.

There is an additional consequence deriving from the focused mossy projection having to do with the effect of the small "N" of the sample of DG units seen by each CA3 unit. When each CA3 unit samples from the entire DG, the expected level of overlap on the DG seen by each unit would be very close to the probability as we have computed it, because the variance in overlap level for the 3,400 active DG units is minimal due to the large number of units contributing (each of which has the computed, independent probability of being active in both patterns A and B). The distribution of actual levels of overlap over the entire DG when each unit has an independent probability of overlap can be modeled by the binomial distribution. The standard deviation (in terms of probability) for such a distribution is $\sqrt{n_{dg}\omega_{dg}(1-\omega_{dg})}/n_{dg}$, which is .00742. However, when computed for the case where the DG contains effectively 2,000 units, only 8 will be active, so the standard deviation for the overlap distribution becomes .153. When this rather wider distribution is convolved with the non-linearities of the pattern separation curve, the resulting pattern overlap might not compare very well with results based on the expected value of DG pattern overlap.

To test this possibility, we computed the pattern separation on CA3 when using the binomial distribution of DG pattern overlap as compared to the usual expected value computation. Since this distinction is relevant only for the mossy fiber inputs to CA3, we used the mossy-only case with several different expected DG overlap levels. The CA3 pattern separation is the convolution of the binomial distribution of DG overlap levels with their resulting CA3 pattern separation levels. Figure 19 shows that there is some difference, but it is not substantial.

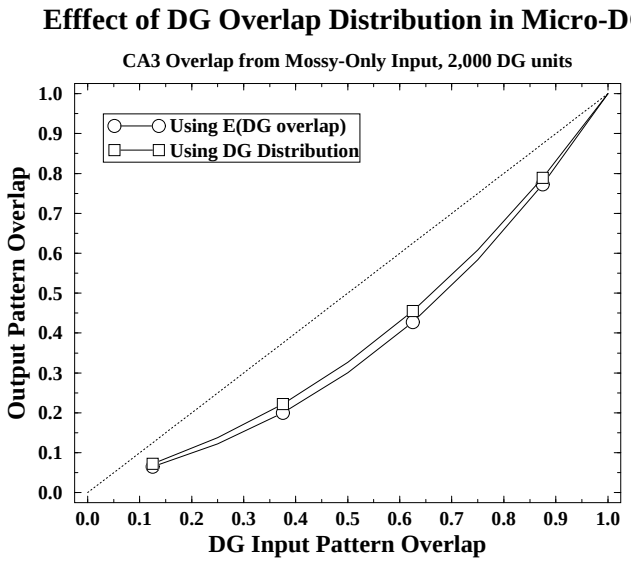

Figure 19: The effect on pattern separation of using the DG overlap distribution in the micro-DG case (2,000 DG units per CA3 unit), as compared to using the expected value of DG overlap. The expected value computation underestimates the true level of pattern overlap on CA3, but only by a small amount.

Discussion

Through analytical models, we have evaluated and compared with plausible alternatives several properties of the hippocampus in order to determine why the hippocampus is constructed as it is. The metric we have used in this evaluation is that of avoiding to the greatest extent possible a trade-off between pattern separation and pattern completion. This metric derives from a larger theoretical model about what the role of the hippocampus is for memory and behavior that is consistent with many sources of evidence, from human and animal amnesics to hippocampal neuron recording data from behaving animals. By using analytical techniques that allow our results to be generalized along the range of parameters in our models, we have been able to establish with some certainty that the parameters corresponding to those of the hippocampus are indeed effective in avoiding the separation/completion trade-off.

The strength of the mossy fiber inputs from the dentate gyrus to area CA3 is an important parameter in our model, with a strength of at least 10 times that of a perforant path synapse being necessary to result in significant improvements on the separation/completion tradeoff. However, we were able to show that the mossy fibers exert their influence by having a high level of variance in the input distribution, instead of through their raw strength in "detonating" the cell. Thus, evidence that indicates that the mossy fibers do not act as "detonator" synapses (Brown & Zador, 1990) does not necessarily mean that a more modest level of EPSP would not be effective in improving the pattern separation properties of the CA3 neurons.

One feature which we found to be highly significant for improving the trade-off was the use of an increase and decrease weight update function, as opposed to an increase-only update function. While there is strong evidence that associative LTP is taking place in the perforant pathway connections in the hippocampus (McNaughton, 1983; Brown et al., 1990), the evidence for heterosynaptic LTD, while present, is perhaps less plentiful (Levy & Desmond, 1985; Levy et al., 1990). Our model leads us to believe that LTD is very important for the functioning of the hippocampus, and that the

relative magnitude of LTP *vs* LTD would have relevance for the setting of a separation/completion threshold.

Further, there are some empirical issues regarding the balance between LTP and LTD that need to be addressed. For example, if the very simple increase/decrease learning rule explored in this paper were applied repeatedly for many different patterns, all the weights in the system would go to zero (assuming them to be bounded at a lower limit of zero) because the relatively sparse activity levels in the hippocampus would cause weights to be decreased much more frequently than they are increased. One straightforward way of dealing with this problem is to have the weights approach their upper and lower bounds using an exponentially decreasing step size, which dynamically alters the ratio of increase to decrease learning based on the size of the weight. In addition, one can use a lower limit that is above zero, so that weights that have not been potentiated recently are still functional. These modified learning rules, and other more sophisticated ways of regulating the overall size of the weights, remain to be thoroughly explored.

In addition to increase/decrease learning, we have identified several ways in which the hippocampus could further improve its performance on the separation/completion trade-off, including the inactivation of the DG neurons for partial input patterns, and the unmodifiability of the mossy fiber synaptic strength. Neither of these features is as critical to the functioning of the hippocampus as the LTD learning, but both have a degree of plausibility that merits further empirical research on these topics.

Our finding that the completion-separation tradeoff found in the feedforward pathway from EC to CA3 is minimized when the DG participates in initial representation formation (separation) but not in retrieval (completion) nicely complements the earlier findings of Treves and Rolls (1992). They showed that the recurrent collateral feedback from CA3 pyramidal cells may tend to swamp the signal arising from the EC during storage, but that strong mossy fiber input could overcome this swamping effect. They further noted that associative modification of synaptic inputs into CA3 is important if incomplete input patterns are to successfully initiate pattern completion via CA3 collaterals. Thus the two analyses both point to the possibility that the dentate gyrus may play a more important role in storage than in retrieval.

We would, however, be hesitant to suggest that the DG typically remains totally inactive during recall. One reason for this is the fact that the perforant path input to the dentate is rich in NMDA receptors, and LTP is easily induced in the perforant path input to the dentate (Bliss & Lomo, 1973). If the DG's role is only to separate patterns for encoding, this plasticity is very puzzling since as we have seen LTP enhances completion but tends to reduce pattern separation. Indeed, we showed that, while it might require larger learning rates, a system without the direct perforant path inputs to CA3 can actually perform better on both separation and completion than a system with both forms of input.

One resolution of this situation would be to exploit the idea raised in this paper that the DG may use a relatively fixed threshold, so that during completion neurons that become active are very likely to have been present during earlier storage of the complete pattern. This would then allow the plasticity in the perforant path projection to the DG to result in enhancement of pattern completion with a minimal disruption of the pattern separation effect. Obviously this matter deserves further experimental as well as theoretical exploration.

In the course of identifying those parameters which are relevant for the pattern separation and completion properties of the hippocampus, we have also identified those parameters which, despite intuitions to the contrary, are *not* relevant. These include the size of the fan-in, and the focused aspect of the mossy fiber pathway projections to CA3. In both cases, these parameters

affect variables in the probability distributions that reflect the sample size used in computing the distributions, but not the intrinsic probabilities. The fan-in size is like the sample taken of the input environment, and as long as this is sufficiently large, the statistics of the input layer will determine the probability distribution. Similarly, the focused mossy fiber projection amounts to giving each CA3 neuron a smaller subspace of the DG input layer to sample from. Again, as long as this sample is sufficiently large (e.g., above 1,000 or so), then the difference between a focused and a diffuse mossy pathway is not substantial. Of course, there are likely to be other constraints on these parameters which are not included in our model.

A central question that arises concerns the degree to which our models can be related to empirical findings. We have attempted to incorporate findings from the literature on hippocampal anatomy and physiology into our models, and evaluate their impact on pattern separation and completion. However, the actual results from our model have not been compared to results obtained by recording the activity patterns of neurons in the hippocampus. This is because the relevant data has yet to be collected. Little is known about the pattern overlap and level of activity on the entorhinal cortex in different behavioral contexts, to say nothing of the simultaneous recording of this kind of data from the EC, DG and CA3. The kind of data our model needs can best be provided by massive parallel recording of activity patterns in different regions of the hippocampus, because our predictions are about pattern level properties, not about the individual firing patterns of single neurons. The kind of recording techniques required are just being developed (Wilson & McNaughton, 1993), and we eagerly await the data they will provide.

Thus, we feel that our work makes a contribution by identifying and comparing at a very basic level several important mechanisms that are almost undoubtedly involved in how the hippocampus functions. There is ample support for the notion that activity in the hippocampus and elsewhere in the brain is regulated by a network of inhibitory interneurons, and that there are significant differences between the activity levels of different hippocampal regions. Further, the anatomical connectivity of the perforant path and the mossy fiber pathway have been extensively studied, and their properties are reasonably well established. Given that our model is based upon only the two assumptions of activity regulation and diffuse, random connectivity, its behavior is almost certainly relevant for understanding the role of the perforant path in the DG and area CA3.

Further, our model is able to relate neural mechanisms on the one hand with behaviorally relevant information processing functions on the other, establishing a critical link for understanding the role of the hippocampus in the broader context of animal and human behavior. In summary, perhaps the best interpretation of our results is that we can not say that we now know for sure what the hippocampus is doing, but we can say that if it were doing what we think it is, then it is well designed to do so.

Appendix A: Computing the Conditional Probability for H_b given H_a

Given a specific number of hits H_a on pattern A, the probability that a unit will be active for a pattern B will be the cumulative probability of getting more than H_t^b total hits from the presentation of pattern B. Since these hits will come from both H_{ab} and $H_{\bar{a}b}$, we need to perform a double-sum over all possible combinations of these numbers which will lead to having more than H_t^b total hits:

$$P_b^{tot}(H_a, \Omega_i) = \sum_{H_{ab}=MAX(0,H_a-(k_i-\Omega_i))}^{MIN(\Omega_i,H_a)} \sum_{H_{\bar{a}b}=0}^{k_i-\Omega_i} \begin{cases} P_b(H_a, \Omega_i, H_{ab}, H_{\bar{a}b}) & ; H_{ab} + H_{\bar{a}b} >= H_t^b \\ 0 & ; otherwise \end{cases} \quad (7)$$

The lower limit on the H_{ab} sum term is necessary to prevent the first coefficient in the numerator $\begin{pmatrix} k_i - H_a \\ \Omega_i - H_{ab} \end{pmatrix}$ from being undefined due to trying to select more units $(\Omega_i - H_{ab})$ than are in the region $(k_i - H_a)$. The upper limit prevents a similar problem from happening with the second coefficient, which as a total of H_a possible units in it.

Since equation 7 gives the probability for activity in B for a specified prior H_a level, the total probability for activity in B among those units which were active in A is the sum over the tail of this probability distribution where $H_a >= H_t^a$. The conditional probability is just this total divided by the prior probability for activity in A:

$$\omega_o(\Omega_i) = P(H_b >= H_t^b | H_a >= H_t^a) = \frac{\sum_{H_a >= H_t^a} P_b^{tot}(H_a, \Omega_i)}{\alpha_o} \quad (8)$$

Appendix B: Interpolation Algorithm for Pattern Completion Curves

This algorithm takes advantage of the information about what the actual output activity was for each data point, which gives an estimate of the magnitude of the error of the data point as compared to the desired output activity level. The line is fit through the data points by incorporating smoothness constraints with a data-fitting cost function that uses both the actual output completion data and the deviation in output activation level. This algorithm produced smooth lines having total squared errors (in terms of the cost function being minimized) of less than .01.

There are four terms in the cost function which is iteratively minimized along the gradient of the function with respect to the parameters: local and average slope deviation (i.e., smoothness constraints), an exponential function of the actual activation level error which is used to correct the actual data points, and the squared distance between the corrected data points and the the fitted line. The parameters of the function are two parameters of the exponential function (g = gain and o = offset), the weighting of the exponential function term, k_{exp} and the weighting of the entire error term, k_{err} and the N data point estimates, d^*_{1-N}.

The cost of the local-slope deviation is just the slope between the current estimate point and the previous one and the slope between the previous two points (it is 0 before $i = 3$):

$$C_{ls} = k_{ls} \sum_{i=3} ((d^*_i - d^*_{i-1}) - (d^*_{i-1} - d^*_{i-2}))^2 \tag{9}$$

where k_{ls} is a weighting constant. The derivative of this with respect to the estimate is:

$$\frac{\partial C_{ls}}{\partial d^*_i} = -k_{ls}(d^*_i - d^*_{i-1}) - (d^*_{i-1} - d^*_{i-2}) \tag{10}$$

The aggregate slope deviation is computed by first computing the mean local slope within a range r around a given data point:

$$s_i = \frac{1}{2r} \sum_{j=i-r}^{i+r} d^*_j - d^*_{j+1} \tag{11}$$

The cost term is then just the squared difference between this and the slope at the current point times a weighting constant (k_s):

$$C_s = k_s \sum_i ((d^*_i - d^*_{i-1}) - s_i)^2 \tag{12}$$

The derivative of this term with respect to the estimate point is:

$$\frac{\partial C_s}{\partial d^*_i} = -k_s((d^*_i - d^*_{i-1}) - s_i) \tag{13}$$

The exponential function which scales the activation-level error is specified as a function of the input overlap proportion (ω_i):

$$f_{err} = e^{g(\omega_i - o)} \tag{14}$$

which reflects the increasing influence of the activation error as overlap increases. The error term used to correct the obtained pattern overlap data points is:

$$err_i = (\alpha_o - \alpha_o^{d_i})(1 + k_{exp} f_{err}) \tag{15}$$

where $\alpha_o^{d_i}$ is the output activity level associated with the data point d_i and α_o is the desired activity level for the output layer.

Finally, the data cost is just the difference between the estimate and the obtained data (d_i) minus the error term:

$$C_d = k_d(d_i^* - (d_i - k_{err} err_i))^2 \tag{16}$$

The gradient of this cost term is taken with respect to each of the modifiable parameters. We let $\alpha_{err} = (\alpha_o - \alpha_o^{d_i})$:

$$\frac{\partial C_d}{\partial d_i^*} = -k_d(d_i^* - (d_i - err_i)) \tag{17a}$$

$$\frac{\partial C_d}{\partial k_{err}} = -k_d(d_i^* - (d_i - err_i))err_i \tag{17b}$$

$$\frac{\partial C_d}{\partial k_{exp}} = -k_d(d_i^* - (d_i - err_i))k_{err}\alpha_{err}f_{err} \tag{17c}$$

$$\frac{\partial C_d}{\partial g} = -k_d(d_i^* - (d_i - err_i))k_{err}\alpha_{err}k_{exp}(\omega_i - o)f_{err} \tag{17d}$$

$$\frac{\partial C_d}{\partial o} = k_d(d_i^* - (d_i - err_i))k_{err}\alpha_{err}k_{exp}gf_{err} \tag{17e}$$

The algorithm is started with the data point estimates, d_i^* set to the actual pattern completion data points, and reasonable initial values for the remaining parameters. Then, parameters are iteratively adjusted by some fraction ϵ of the gradient as computed above until the total cost (the sum of all the individual costs shown above) stabilizes. The initial parameters used were: $k_{err} = 1$, $k_{exp} = 1.8$, $g = 5.2$, $o = .7$, $k_{ls} = .1$, $k_s = .8$, $k_d = .1$, $\epsilon = .1$.

References

Amaral, D. G. Ishizuka, N. & Claiborne, B. (1990). Neurons, numbers and the hippocampal network. *Progress in Brain Research*, *83*, 1–11.

Amit, D., Gutfreund, H., & Sompolinsky, H. (1987). Information storage in neural networks with low levels of activity. *Physical Review A*, *35*, 2293–2303.

Barnes, C. A., McNaughton, B. L., Mizumori, S., Leonard, B. W., & Lin, L.-H. (1990). Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Progress in Brain Research*, *83*, 287–300.

Barrionuevo, G., Keslo, S., Johnston, D., & Brown, T. (1986). Conductance mechanism responsible for long-term potentiation in monosynaptic and isolated excitatory synaptic inputs to hippocampus. *Journal of Neurophysiology*, *55*, 540–550.

Bliss, T. & Lomo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology (London)*, *232*, 331–356.

Boss, B., Peterson, G., & Cowan, W. (1985). On the numbers of neurons in the dentate gyrus of the rat. *Brain Research*, *338*, 144–150.

Boss, B., Turlejski, K., Stanfield, B., & Cowan, W. (1987). On the numbers of neurons in fields CA1 and CA3 of the hippocampus of Sprague-Dawley and Wistar rats. *Brain Research*, *406*, 280–287.

Brown, T. & Johnston, D. (1983). Voltage-clamp analysis of mossy fiber synaptic input to hippocampal neurons. *Journal of Neurophysiology*, *50*, 487–507.

Brown, T., Kairiss, E., & Keenan, C. (1990). Hebbian synapses: Biophysical mechanisms and algorithms. *Annual Review of Neuroscience*, 475–511.

Brown, T., Wong, R., & Prince, D. (1979). Spontaneous miniature synaptic potentials in hippocampal neurons. *Brain Research*, *177*, 194–199.

Brown, T. H. & Zador, A. (1990). Hippocampus. In G. Shephard (Ed.), *The synaptic organization of the brain*. Oxford: Oxford University Press, Chap. 11, 346–388.

Byrne, J. H. & Berry, W. O. (Eds.) (1989). *Neural models of plasticity: Experimental and theoretical approaches*. San Diego, CA: Academic Press, Inc.

Claiborne, B., Amaral, D., & Cowan, W. (1986). A light and electron microscopic analysis of the mossy fibers of the rat dentate gyrus. *Journal of Comparative Neurology*, *246*, 435–458.

Damasio, A. R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, *1*, 123–132.

Gibson, W. & Robinson, J. (in press). Statistical analysis of the dynamics of a sparse associative memory. *Neural Networks*.

Gibson, W., Robinson, J., & Bennett, M. (1991). Probabilistic secretion of quanta in the central nervous system: Granule cell synaptic control of pattern separation and activity regulation. *Philosophical Transactions of the Royal Society (Lond.) B*, *332*, 199–220.

Gluck, M. A. & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*.

Gluck, M. A. & Rumelhart, D. E. (Eds.) (1990). *Neuroscience and connectionist theory*. Hillsdale, NJ: Lawrence Earlbaum Associates, Inc.

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

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

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

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. USA*, *81*, 3088–3092.

Humphreys, M., Bain, J., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*, 208–233.

Insausti, R., Amaral, D., & Cowan, W. (1978). The monkey entorhinal cortex. ii. cortical afferents. *Journal of Comparative Neurology*, *264*, 356–395.

Levy, W. & Desmond, N. (1985). The rules of elemental synaptic plasticity. In W. Levy, J. Anderson, & S. Lehmkuhle (Eds.), *Synaptic modification, neuron selectivity, and nervous system organization*. Hove, England: Lawrence Earlbaum Associates, Chap. 6, 105–121.

Levy, W. B., Colbert, C. M., & Desmond, N. L. (1990). Elemental adaptive processes of neurons and synapses: A statistical/computational perspective. In (Gluck & Rumelhart, 1990), Chap. 5, 187–235.

Marr, D. (1969). A theory of cerebellar cortex. *J. Physiol. (London)*, *202*, 437–470.

Marr, D. (1970). A theory for cerebral neocortex. *Proc. Royal Soc London B*, *176*, 161–234.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society (Lond.) B*, *262*, 23–81.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (submitted). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionst models of learning and memory. *The Journal of Neuroscience*.

McClelland, J. L., McNaughton, B. L., O'Reilly, R. C., & Nadel, L. (1992). Complementary roles of hippocampus and neocortex in learning and memory. *Society for Neuroscience Asbtracts*, *18*(2), 1216.

McNaughton, B. L. (1983). Activity dependent modulation of hippocampal synaptic efficacy: Some implications for memory processes. In W. Seifert (Ed.), *Neurobiology of the hippocampus*. San Diego, CA: Academic Press, Inc., Chap. 13, 233–252.

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

McNaughton, B. L. & Nadel, L. (1990). Hebb-marr networks and the neurobiological representation of action in space. In (Gluck & Rumelhart, 1990), Chap. 1, 1–63.

Rolls, E. (1990). Principles underlying the representation and storage of information in neuronal networks in the primate hippocampus and cerebral cortex. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks*. San Diego, CA: Academic Press.

Rolls, E. T. (1989). Functions of neuronal networks in the hippocampus and neocortex in memory. In (Byrne & Berry, 1989), 240–265.

Schmajuk, N. A. & DiCarlo, J. J. (1992). Stimulus configuration, classical conditioning, and hippocampal function. *Psychological Review*, *99*(2), 268–305.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*, 195–231.

Squire, L. R., Shimamura, A. P., & Amaral, D. G. (1989). Memory and the hippocampus. In (Byrne & Berry, 1989).

Sutherland, R. J. & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology*, *17*(2), 129–144.

Tamamaki, N. (1991). The organization of reciprocal connections between the subiculum, field CA1 and the entorhinal cortex in the rat. *Society for Neuroscience Asbtracts*, *17*, 134.

Teyler, T. J. & Discenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, *100*, 147.

Torioka, T. (1978). Pattern separability and the effect of the number of connections in a random neural net with inhibitory connections. *Biological Cybernetics*, *31*, 27–35.

Torioka, T. (1979). Pattern separability in a random neural net with inhibitory connections. *Biological Cybernetics*, *34*, 53–62.

Treves, A. & Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain. *Network*, *2*, 371–397.

Treves, A. & Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus*, *2*, 189–199.

Van Hoesen, G. (1982). The parahippocampal gyrus. New observations regarding its cortical connections in the monkey. *Trends In Neurosciences*, *5*, 345–350.

Van Hoesen, G. & Pandya, D. (1975a). Some connections of the entorhinal (area 28) and perirhinal (area 35) corticies of the rhesus monkey. I. Temporal lobe afferents. *Brain Research*, *95*, 1–24.

Van Hoesen, G. & Pandya, D. (1975b). Some connections of the entorhinal (area 28) and perirhinal (area 35) corticies of the rhesus monkey. III. Efferent connections. *Brain Research*, *95*, 39–59.

Van Hoesen, G., Pandya, D., & Butters, N. (1975). Some connections of the entorhinal (area 28) and perirhinal (area 35) corticies in the rhesus monkey. II. Frontal lobe afferents. *Brain Research*, *95*, 25–38.

Wickelgren, W. A. (1979). Chunking and consolidation: A theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychological Review*, *86*, 44–60.

Wilson, M. A. & McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, *261*, 1055–1058.

Yamamoto, C. (1982). Quantal analysis of excitatory postsynaptic potentials induced in hippocampal neurons by activation of granule cells. *Experimental Brain Research*, *46*, 170–176.