

Computational Models of Learning in Simple Neural Systems. (R. D. Hawkins & G. H. Bower, eds.) [The Psychology of Learning and Motivation, vol. 23] San Diego: Academic Press, 1989, 243-305.

A COMPUTATIONAL APPROACH TO HIPPOCAMPAL FUNCTION

William B Levy

I. Introduction

This article presents the early, formative stages of a theory of hippocampal function. This theory, while stimulated by the psychological observations indicating a role for the hippocampus in short-term working memory and spatial behavior, develops mainly through consideration of computational issues. These computational issues are related to the psychological viewpoint through physiological and anatomical observations. The critical anatomy which dominates our thinking about the hippocampus is shown in Fig. 1.

In the theory presented here, the hippocampus participates in the prediction of future representations based on past and present representations. All three classes of representations are derived from a multiplicity of sensory modalities, such as auditory, visual, and olfactory signals from neo- and piriform cortices. This fusion of sensory modalities requires recoding because of computational complexity problems.

Figure 2 relates the functional groupings used to explain the theory of hippocampal function to the cells of the hippocampus and entorhinal cortex.

The CA1 region of the hippocampus is postulated to be a prediction-generating layer or tier. This region produces a prediction based on its input from hippocampal region CA3.

The combined hippocampal dentate gyrus/CA3 (DG/CA3) system is

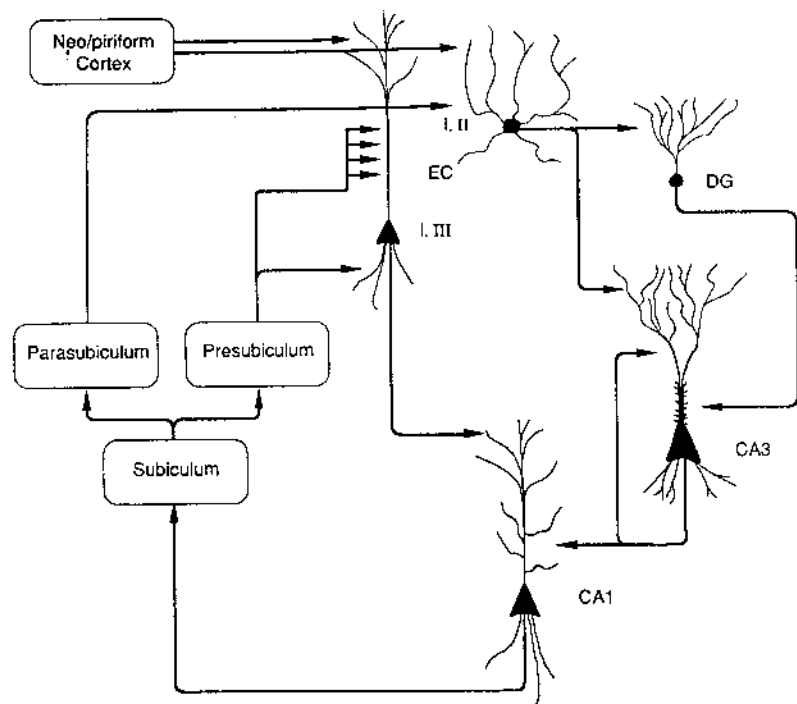


Fig. 1. Limbic system inspiration for the networks: CA1 and CA3 are regions of Ammon's horn (cornu Ammonis) of the hippocampus, DG is the hippocampal dentate gyrus, EC is entorhinal cortex, and I.II and I.III are layers II and III of EC. Copyright © 1989 by William B Levy.

postulated to be a preprocessor serving the CA1 prediction layer. This preprocessor is the focus of the chapter. Computational complexity considerations imply the utility of such recoding as part of signal mixing. That is, from a computational perspective, this preprocessor decreases the statistical dependency of individual representations and increases the similarities between successive representations. These changes allow the CA1 prediction layer to create more accurate predictive representations and predictive representations which predict further into the future.

The computational complexity problems arise from the combinatorial explosion of possible representations resulting when the hippocampus and supporting limbic structures mix representations from multiple sensory modalities. These problems are particularly difficult because both the mixed and unmixed representations occur as sequences which contain information beyond that found within the individual representations.

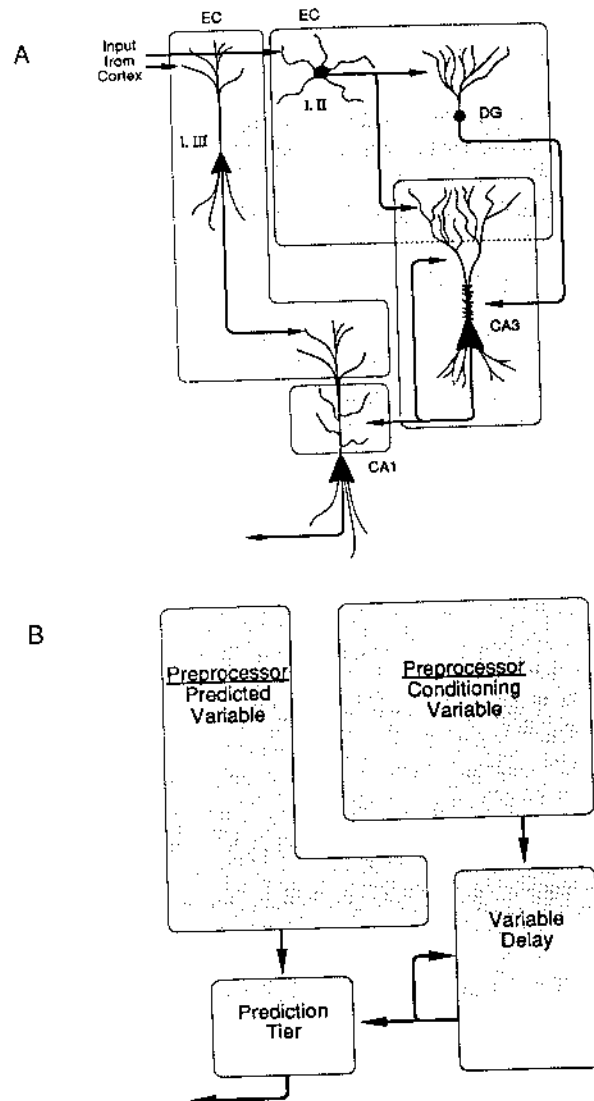


Fig. 2. A juxtaposition of the proposed computational model, B, with its anatomical basis, A. Note that the major subdivisions of the hippocampus and entorhinal cortex depicted in A do not always correspond precisely to the functional units of the model in B. One reason for this imprecise correspondence is that neurons are multicompartmented based on their segregation of afferent termination zones. Layer II of the entorhinal cortex, the dentate gyrus, and the distal CA3 dendritic region correspond to the preprocessor of the conditioning variable. Layer III of the entorhinal cortex and the distal dendritic region of the CA1 pyramids correspond to the preprocessor for the predicted variable. The variable delay resides within the CA3 region. The prediction tier corresponds to the CA1 pyramids. The theory presented here leads up to and discusses preprocessing of the conditioning variable but does not explore preprocessing of the predicted variable. Copyright © 1989 by William B Levy.

A major problem is explaining how the hippocampus computes predictions about these sequences in the face of this combinatorial explosion. The solution advanced here is the generation of an optimal approximate prediction by CA1. The form of this computation is specified by an axiomatic theory of inference called minimum relative entropy (MRE) and by the nature of synaptic modification at the CA3-CA1 synapses. With the specification of a CA1 computation, a desirable preprocessing computation can then be inferred.

The preprocessing computation, a hypothesized function of the DG/CA3 region, produces two results: (1) the reduction of the intrinsic statistical dependency of the CA3 representations, and (2) a variable time shifting of representations which allows predictions to be generated before the event being predicted.

Thus, two problems—time-shifting and complexity—are solved with one preprocessing network. This proposed dual functionality combines with the CA1 prediction function to contribute a theory of hippocampal function in small animals which relates to the theory of hippocampal function in humans. This is true because the network under consideration dynamically represents the present, recent past, and the future. The ability to represent all three time frames seems to be a functional requirement for a working short-term memory. As a result, this computational theory, based on anatomy and physiology, can exist along with psychological theories of hippocampal function for both rats and humans.

The chapter is divided into six sections. Section II provides background for bridging the gap between rat and human hippocampal theories and points out that evolutionary pressure for efficient spatial behavior is also pressure for sensory mixing to form representations and sequences of representations and pressure for accurate prediction of future representations in these sequences as well. Section III sets forth the computational issues in more detail, and Section IV summarizes these issues in the context of Section V. Section V, while giving more details of the model, discusses the anatomical and physiological observations which motivated the specific model. Section VI sketches an algorithm which is the hypothesized computation performed by the DG/CA3 region.

II. Anatomy, Evolution, and Psychology Form a Background to the Computational Theory

At this time there seem to be two, apparently differing, theories of hippocampal function. According to these theories, either the hippocampus provides a distraction-protected, short-term (Squire, 1987) working mem-

ory (Goldman-Rakic, 1987; Murray & Mishkin, 1987; Olton, 1978; Raffaele & Olton, 1988), or it is a locus contributing to, and necessary for, competent spatial behavior (O'Keefe & Nadel, 1978). [There are, however, several papers bridging these two ideas (e.g., Breese, Hampson, & Deadwyler, 1989; Eichenbaum & Cohen, 1988; Foster, Christian, Hampson, Campbell, & Deadwyler, 1987; Olton, 1985).] These two formulations are not mutually exclusive and are best related through computational considerations. As a prelude, this section presents a viewpoint which interrelates spatial function and hippocampal anatomy, setting the stage for the computational arguments in the process. Here we advance the idea that efficient signal mixing for sequence prediction, which facilitates efficient spatial behavior, was the original fundamental computation for which the hippocampus evolved.

A. ANATOMY

It is obvious merely from anatomical considerations that the hippocampus and parahippocampal limbic regions bring together, or fuse, diverse sensory signals arising from association areas of neocortex. This sensory signal fusion function is common to all the mammals which have been studied, including primates as well as rodents (Jones & Powell, 1970; Pandya & Kuypers, 1969; see Amaral, 1987; Swanson, Köhler, & Björklund, 1987, for reviews). Therefore, on purely anatomical grounds, it is sensible to infer that one function of the hippocampus and associated parahippocampal cortices is to mix sensory modalities.

With the exception of olfactory inputs, most sensory information entering the hippocampus actually comes from association neocortex via the entorhinal cortex (Insausti, Amaral, & Cowan, 1987; Jones & Powell, 1970; Pandya & Kuypers, 1969; Saper, 1982; Sorensen, 1985; Van Hoesen & Pandya, 1975; Van Hoesen, Pandya, & Butters, 1972, 1975). Most mixing of signals occurs prior to their reaching the hippocampus, within the neocortex and entorhinal cortex (but see Fig. 3). The hippocampus itself receives its multimodal signals from entorhinal cortex, with help from the adjacent subicular cortices, and in turn feeds its signals back to entorhinal cortex (Kosel, Van Hoesen, & Rosene, 1982; Sorensen & Shipley, 1979; Swanson & Cowan, 1977; see Amaral, 1987; Rosene & Van Hoesen, 1987, for reviews). Thus the entorhinal cortex appears to be the ultimate site of fusion of the sensory signals from association cortices while the hippocampus can be seen as a structure supporting the signal mixing performed in the entorhinal cortex. As a support structure, the hippocampus remixes the multimodal signals coming from entorhinal cortex. This remixing, as shown in Fig. 3, occurs primarily through diver-

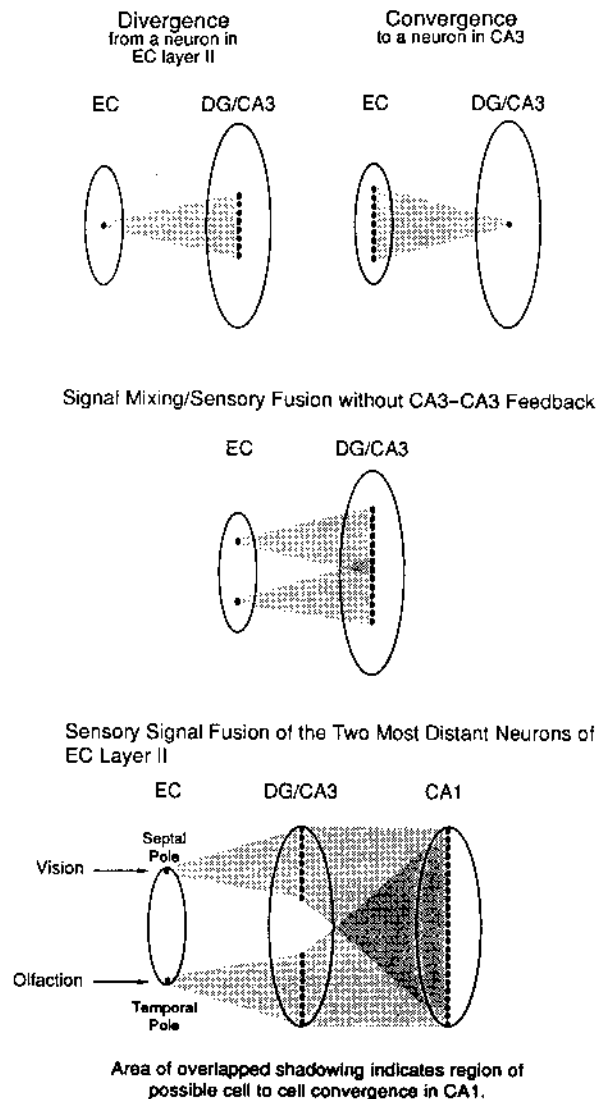


Fig. 3. Schematic illustration of divergence and convergence of signals and their contribution to sensory signal fusion. The entorhinal cortex (EC) may receive signals that are spatially separated across its septo-temporal axis. However, even without CA3-CA3 feedback, it is highly probable that the hippocampal representations of these spatially separated EC inputs will include neurons which are active due to the divergence and convergence of axonal projections. Note, in the bottom panel, the extent of convergence in CA1 even when very spatially separated EC inputs are considered. Such an association of diverse input signals should eventually lead to decreased signal complexity and better prediction. Copyright © 1989 by William B Levy.

gence-convergence of the entorhinal cortical-DG/CA3 projections, the CA3-CA1 projections, and the feedback of region CA3 upon itself (not shown in Fig. 3).

The question to address now is, How are signal mixing and remixing related to efficient spatial behavior?

B. A MATTER OF BACKGROUND: WHOSE HIPPOCAMPUS IS IT, ANYWAY?

Only mammals have a well-defined hippocampus and all mammals have this structure. Thus, it is probable that the stem mammal, a small shrewlike creature, had a hippocampus though its reptilian relatives did not. But why did the hippocampus develop? What purpose did it serve? If we can answer these questions we might be able to imagine how the human and rat hippocampi, although divergent in function, are related.

One outstanding distinction between mammals and reptiles is that mammals exhibit nesting behavior and suckle their young while reptiles do not. Consequently, the mammal with a nest full of young must return home after foraging for food. Thus the prototypical mammal had to find its way upon the earth's surface, constrained by objects which towered over it to form both visual and physical barriers. We should picture this world not from the viewpoint of erect hominids but from the lowly perspective of a small shrewlike animal scurrying through the brush and forests. A small rock or a tuft of grass can tower above a small animal to form wall-like structures. In this case, perception of spatial relations is based not on an all-encompassing, maplike view from above but on sensory sequences involving all senses.

From the above considerations come two main points. The first is that learning-dependent spatial behavior in a maze, as studied in rats and mice today, seems as relevant to the prototypical mammal as to a rat or mouse in the wild. The second point is that studies of spatial behavior essentially require animals to learn and predict sensory sequences.

It was and is a tremendous advantage for a small mammal to get from the nest to a feeding area and back to the nest again as efficiently as possible. This problem of scurrying efficiently to and from the nest is a problem of prediction which requires mixing multimodal sensory signals as these signals arrive sequentially over time. That is, an animal will return home most quickly if it can use sights, smells, sounds and other sensations as they arrive over time in order to predict correctly from moment to moment which way to turn or when to go straight. Note that such prediction is based on sensory input and, in an immediate sense, is just prediction of the set of sensory signals which will follow the set of signals being received at a given time. These sets of signals represent the envi-

ronmental stimuli bombarding the animal as it makes its way home. To predict the next set of signals in the sequence is effectively to anticipate what is "around the corner."

Maneuvering in a maze or in the wild produces a sequence of sensory inputs involving all sensory modalities: sight, smell, hearing, acceleration, touch, and perhaps even taste. Obviously, the visual input changes as an animal moves about. However, for a small animal with a diminished range of vision, the sense of smell seems at least equally important. Even so, vision and olfaction are not the only senses useful in a maze. Experimenters have found that a rat can use sound cues, such as the sound made by an exhaust fan, to locate itself relative to its surroundings. It is plausible then that our shrewlike animal would be able to use the sound of a babbling brook to orient itself. In regard to touch, we know that we as humans can use the feel of carpet versus linoleum underfoot to locate ourselves. The vestibular sense tells an animal which way it faces as it turns. It is even conceivable that a small animal might lick its surroundings and detect taste differences. Thus we can picture each of the senses solving the same problem: prediction of future multisensory perceptions.

In sum, then, it was the earliest mammal, as distinct from its ancestors, which had the most pressing need to locate itself relative to its present and future surroundings as it moved about. Additionally, in performing this relative localization, all senses are potentially useful.

The evolutionary connection between sensory signal mixing and efficient spatial behavior is further elaborated by using and extending the ideas of Jerison (1973). At the time the first mammals evolved, dinosaurs and other reptiles dominated the land by day, but at night, when it became too cold for these poikilotherms, animals that could regulate their body temperature were free to roam unmenaced by predatory reptiles. In this scenario we picture our prototypical mammal scurrying through a partially darkened maze in which the visual system is much less efficient (though still useful). Thus, the demands of nocturnal life favored the evolutionary development of superior olfactory and sound-localizing abilities to at least match a visual ability already evolved. Development of these other senses and a need to return home after each foray would have fostered the development of a new brain structure that allowed the different modalities to relate to each other.

If we consider the sensory signal fusion problem in more detail, however, our attention is drawn to a very important computational issue. Mere development of multiple, but individual, sensory perceptual abilities is not nearly as useful as an effective combination of these same abilities. It is not efficient to have one sensory system predict that home lies

to the left and another sensory system predict that it lies to the right. Efficient functioning requires the combination of multimodal signals to produce a harmonious single prediction pointing the way home. However, this sensory signal mixing is a problem of such computational complexity that it baffles engineers even today.

To illustrate the problem with respect to vision and olfaction, suppose one engineer built the best visual pattern recognition system in the world and another engineer built the best gas chromatograph-mass spectrometer for chemical identification. How should the coded representations of reflected photons and molecules be combined? The answer is far from simple since there is no obvious physical relationship between the physics of light as reflected from macroscopic objects and the reactions or structures of different molecules.

Consistent, predictable relationships between the states of neurons which code sights and the states of neurons which code smells may only exist as higher-order statistics. Such higher-order statistics would be infrequently sampled and quite variable on an evolutionary time scale. This situation contrasts with more peripheral, unimodal coding, which can rely on lower-order statistics that are constant across generations. The higher-order correlations of the multisensory problem are difficult to find because there are so many possible correlations compared with the relatively few important correlations which actually exist. Consequently particular experiences of the protomammal's ancestors would provide little help for constructing a nervous system tied to the specifics of the multisensory relationships they encountered.

Yet even though the problem of multisensory representations is a tremendously difficult one, the hippocampus and associated limbic cortices have managed to meld olfactory and visual signals together into useful multisensory representations. Actually these structures work on an even more difficult problem, that of multisensory sequences.

As mentioned above, the particular sensory signal fusion problem which occurs when traversing a maze is not a static problem involving the "mere" creation of a single multisensory representation which will eventually be recognized or categorized in some other brain region. Rather, the maze problem involves changing multisensory representations which arrive over time as the animal moves both in time and in space. Temporally ordered sets of these changing representations are called sequences. An animal which exhibits efficient spatial behavior by accurately predicting what lies around the corner is just predicting one, or several, successive representations in the current sequence of sensory representations. (E.g., if I turn right at this rock I'll pass by the lilacs

and then just down the hill and to the right is the water hole.) From a computational perspective this dynamic problem is even more complex than the already intractable static problem of sensory signal mixing.

Thus we are motivated to study hippocampal function from a dynamic perspective in which this structure predicts computationally complex, multisensory sequences.

In sum, anatomical observations indicate that the hippocampus participates in sensory signal fusion by helping to mix and, in particular, to remix the most highly processed and diverse sensory representations. Evolutionary pressure for new and improved sensory signal fusion capabilities applicable to spatial behavior would have arisen because the protomammal was nocturnal and had a home to which it returned to care for its young. This sensory signal fusion capability facilitates navigation, especially when these fused sensory representations are used for accurate sequence prediction.

The next section gives more precision to the two computational issues introduced here: (1) sensory signal fusion, which lowers signal complexity without losing information and (2) representations of the future, which act as predictions.

III. Computational Issues

The influence of the computational approach becomes profound when we consider a neural network in terms of two issues: complexity and optimization.

The exponential explosion (complexity) of possible configurations of the environment and, of more relevance, of the activity states of the neurons in the brain implies that exact computations which examine and compare each individual neural representation, one by one or in parallel, are not possible. However, there are methods for maintaining the individuality of representations while simplifying the computations made. Such methods are necessarily approximations. What we have in mind is to describe neurally computable transformations and algorithms which are, in some yet to be defined sense, optimal approximate solutions for the prediction problem.

A. MOTIVATING THE COMPUTATIONAL APPROACH

There are many reasons for using a computational approach to formulate theories of brain function. To acquaint the reader with the perspectives leading to and inherent in this approach, I will set forth what seem

to me to be some of its most important advantages, with the caveat that these thoughts are merely intuitions and lack the support of an airtight logical argument.

1. There have been so many partial models of hippocampal function briefly in vogue over the last 30 years that I have despaired of finding any simple, lasting description of hippocampal function that is specific in its predictions. The computational approach benefits from a precise mathematical language which avoids vague descriptions of how the animal "thinks" by confining itself to questions of signal transformation and statistics. In more than one way, the computational approach advocated here is an extension of the abstract approach advocated by Thompson and Spencer (1966) for the study of psychological processes.

2. Since almost every textbook says that the primary function of the brain is "information processing" an explanation of this phrase would seem to be in order. A computational perspective seems to have a good chance of providing a satisfactory definition of information processing.

3. There should be a way to study only one anatomical region of the brain, such as the basal ganglia, the CA1 region of the hippocampus, or the cerebellum, at a time. If we have to study the totality of all brain areas at once, the job of understanding and explaining the functions of these areas seems hopeless. However, if brain regions can be studied in pieces, and not just from the periphery inwards, then many scientists can work in parallel toward understanding the brain, some concentrating on the piriform cortex, others concentrating on the tectum, and so on.

To carry out such a program successfully we will need a language and a perspective which can encompass all brain regions. An abstract computational approach promises to satisfy these needs.

4. The basic computational issue of prediction is fundamental and easily understood. Rapid, accurate prediction is a continual necessity in our everyday lives. It is the process by which we put past experiences to work in order to anticipate the future and act on our environment. In fact our very survival depends upon it. Inquiry into this realm is aided by the fact that prediction is also a rather well-studied subdiscipline in mathematics and statistical theory.

Good prediction can improve the execution of almost any behavior. It enables the experimental rat to make the correct response and receive its food pellet, the engineer to design a structurally sound bridge (based on the knowledge of stress gained from past bridge building experiences), and the baseball player to accommodate his swing when he recognizes the pitcher's curve ball. [My own thoughts on prediction as a general problem in life were most influenced by the writings of Young (1970) and

Dawkins (1976). In writings on the theory of hippocampal function, prediction is often mentioned by Gray (1982) and Vinogradova (as cited in Gray, 1982.)

When dealing with a complex, multitiered structure such as the brain, we can even consider the issue of good prediction in the context of choosing good transformations, that is, prediction of the utility of a transformation in helping to make a subsequent prediction by the network. In fact this leads to the potentially infinite regress of predicting a good signal transformation in order to help predict a good signal transformation . . . in order to predict about the environment. Happily, such complications are unnecessary because the abstract, computational approach allows us to divide the processing sequence into small pieces. Using such subdivisions we may in effect consider predictions locally, in isolation from subsequent predictions.

At this local or regional level, prediction does not explicitly concern the environment external to the organism, that is, the "real world." Rather, we can and should study prediction in cases in which the environment is the set of neurons afferent to the brain region of interest. When dealing with the hippocampus we consider the environment to be essentially the set of neurons in layers II and III of the entorhinal cortex which are afferent to the hippocampus. Even though the hippocampus receives inputs from many other brain regions, we focus here on the entorhinal cortical afferents because of their high information capacity. We base this interpretation on their numerical dominance and high-frequency firing capabilities relative to other hippocampal inputs.

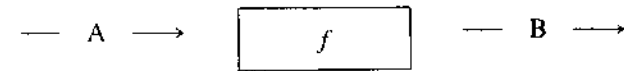
So we already have a payoff from the abstract, computational approach: Prediction is an important, identifiable issue which we may study in a limited portion of the brain by using an abstract definition of the environment.

B. COMPUTATIONAL QUESTIONS

Information- and computation-theoretic considerations will allow us to bypass the precise psychological meaning of neuronal signals and still to have viable research questions because these considerations produce their own fundamental and general meaning of neuronal signals. These theoretic considerations provide both research questions and quantitative measures for the study of hippocampal function or, for that matter, the function of any brain region.

One obvious goal of an abstract signal processing approach is to discover the transformation which converts inputs to outputs. For example,

suppose the hippocampus performs the general computation f ; our goal then is to describe the computation, or transformation, $f: A \rightarrow B$.



Here we think of A as a multidimensional input pattern made up of zeros and ones. More technically, $A(t)$ is a sequence of binary vectors over time t where the dimension of these vectors is the number of axons afferent to the hippocampus. Likewise B , with time implicit, is a corresponding sequence of vectors over the efferent axons.

As interesting and challenging as it will be to understand transformation f , this challenge is not enough. We also want to tie this transformation to other questions. On the simplest level, we can ask, Why is it necessary to transform representations such as A into representations such as B ?

For the signal mixing and sequence prediction problems, we can outline two general reasons for performing transformations on neuronal signals.

1. The computational problem of prediction of the future based on the past and the present necessitates that representations reflecting the past and present be transformed into representations reflecting the future.
2. The computations which might generate predictions concern problems of large dimension, making them intractable. That is, timely and exact predictions of representations which concern more than 50 neurons are of overwhelming computational complexity. Instead of exact solutions, approximate solutions must be computed. Therefore, signal transformations should be performed to improve these approximations.

Before concluding this introductory material, let us consider two powerful constraints which will influence the evolution and analysis of neural networks.

C. OVERRIDING COMPUTATIONAL CONCERNS: INTRACTABILITY AND OPTIMIZATION

1. Computational Intractability

Computational intractability occurs when the number of possible configurations grows exponentially. Such exponential growth is a very old story (Gamow, 1961). Most of us have heard about the not-so-wise mathematician who asked for payment by having one grain of wheat placed in

the first square of a chess board, two grains in the second, four in the third, and so on to end up with 2^{63} grains in the 64th square. As a result of this exponential growth ($2^0, 2^1, 2^2, \dots, 2^{63}$), the man was owed more wheat than will be produced in 2000 years at current production levels. Not surprisingly, he was put to death by the royal debtor.

When exponential growth characterizes a problem requiring examination of all (or most) possible outcomes, computational intractability results. For example, it is calculated that no computer (Shannon, 1950; Winston, 1977) will ever be able to find the provably unbeatable opening move of a chess game (it is posited that there is such a move, assuming that White, the computer, continues to respond optimally from then on). Finding this move and the subsequent correct responses would require a computer to examine the outcome of every possible chess game, estimated to be more than 2.5×10^{154} games. Thus, even if the computer plays very fast, as fast as one game every femtosecond, it will require 10^{135} years to complete the computation, many times the longevity of the universe.

It is well known that computational constraints, such as the large number of possible decisions, the relatively small number of processors, and a limited memory capacity, are serious concerns in high-dimensional multivariable analysis. (Here we use the term *high-dimension* to mean that the number of variables exceeds 200.) The phrases "curse of dimensionality" (Bellman, 1961) and "combinatorial explosion" (Karp, 1975) are often used to describe the difficulties which arise in high-dimensional problems. A well-known problem which suffers from this explosion is that of minimizing the distance covered by a traveling salesman who must visit a specified set of cities before returning home.

There is an extensive literature on the subject of computational intractability. An early article which emphasizes the intractability of decision-making problems faced by neurallike networks is Ashby (1956). Minsky and Papert (1969) pick up this theme and use it for their own purpose. Barlow (1959 and many other references; see, e.g., 1961a, 1961b), although nontechnical in his discourse, is clearly aware of the importance of this problem and that it must be solved by various brain regions. Recent references to the intractability problem from a computational perspective include Garey and Johnson (1979). Zucker (1981) writes lucidly about how computational constraints limit what can be done in pattern recognition problems. The results of Kirkpatrick, Gelatt, and Vecchi (1983) and Hopfield (1984; Hopfield & Tank, 1985) show that good approximations can be achieved in high dimensions. The 1983 paper by Kirkpatrick and colleagues is particularly influential in advocating the idea that, for a decision-making problem requiring optimization, good ap-

proximations in high-dimensional systems are, on average, nearly as good as the set of best possible decisions.

In the case of neural networks, the number of variables present (e.g., processing elements such as neurons) in a simple network can easily number in the thousands. Thus, the sheer number of neurons in the mammalian brain leads to computational intractability there. Figure 4 pictures the exponential growth of the number of possible representations with increasing numbers of neurons. If we consider a neuron as an on-off device, then 1000 neurons can represent 2^{1000} different configurations. For the representations mediated by the approximately 250,000 binary-valued CA1 cells of one rat hippocampus, $2^{250,000}$ different representations could occur. These are unimaginably large numbers (for comparison, consider that there are less than 2^{300} protons and neutrons in the universe). Such a large number of variables and the resulting combinatorial explosion make computations which guarantee exact solutions intractable. Since prediction generation in which a network would have to compare all the alternative possibilities is a computationally intractable problem, a network in the brain will not seek an exact solution to the prediction problem described here. Rather, it is sensible for a network to construct optimal approximations subject to constraints such as available time, number of available computational elements, the computational characteristics of these elements, and the minimally required level of accuracy.

Due to the combinatorial explosion of possible representations, a neural network trying to generate good predictions comes up against at least three insurmountable problems when the perfect, deterministic solution is desired:

1. Time: There is not enough time to compute exact solutions.
2. Processing elements: There are not enough neurons or synapses to store all the statistics needed to describe all possible representations.
3. Samples: There cannot be enough sampling to learn all the probabilities of every representation in such a way that these sample-based estimates, which are just the relative frequency of a representation, converge anywhere near true probabilities.

There are essentially two computational strategies for overcoming intractability. One strategy is to throw out information by decreasing the number of possible configurations which are recognizably different. Pattern recognition, in the sense of categorization and classification, is an example of such a procedure. It is much easier to say "I see a bird" than to give a detailed description of the bird you actually see. Discarding information is obviously risky, however, and the efficacy of this method

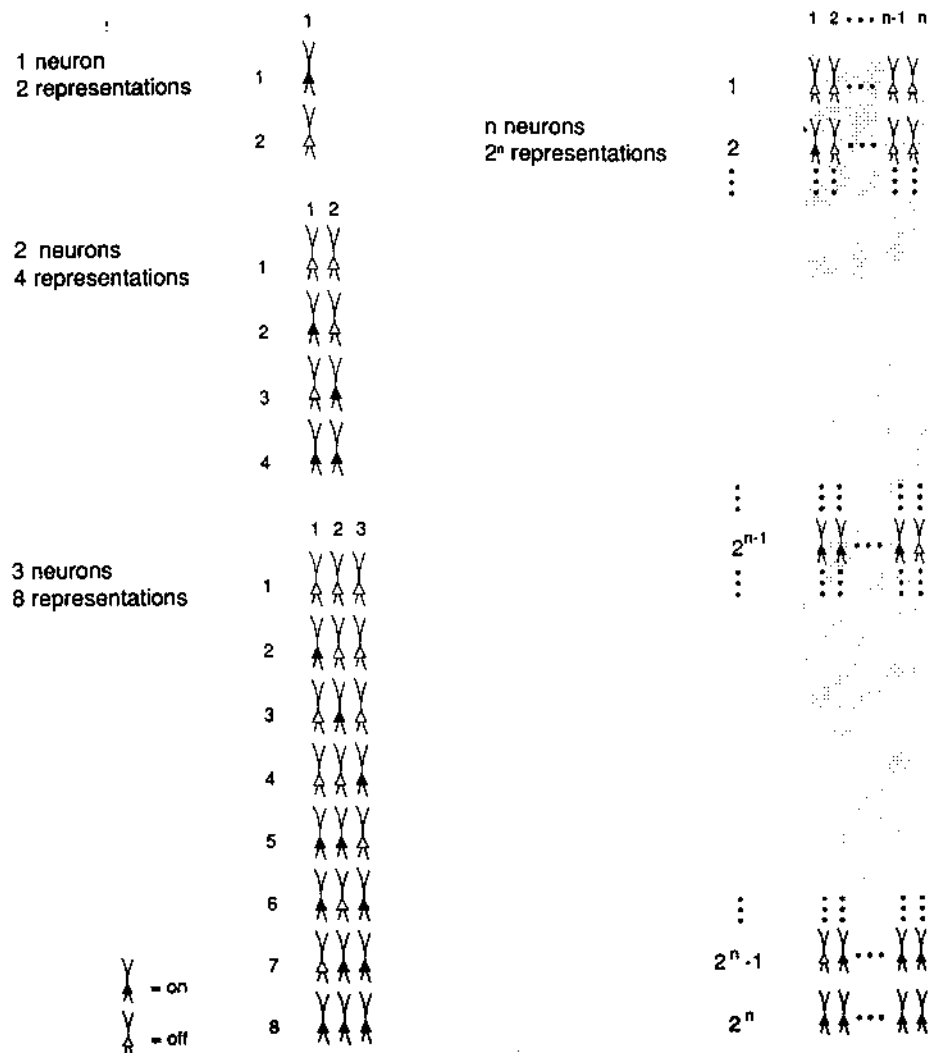


Fig. 4. Exponential growth of possible representations. As the number of neurons increases linearly to n , the number of representations possible increases exponentially as 2^n . The problem of calculating a probability for each of the 2^n representations in a network can, however, best be solved by treating each neuron as if statistically independent and approximating each probability. Copyright © 1989 by William B Levy.

depends on the amount of information lost and its importance. When the importance of any discarded information is unknown and a network or organism is performing poorly, as under mismatch conditions which activate the hippocampus, this strategy seems particularly bad.

The second strategy, which does not throw out information, is to assume statistical independence of converging neuronal activity. If each cell in the network's prediction tier computes a prediction as if statistical independence of its inputs exists, then the computational intractabilities are resolved. However, the usefulness of this approximation is a function of how much statistical dependence exists in the problem under consideration.

If this second strategy is tried and fails to yield acceptable performance, all is not lost as might be if the first strategy is applied. The computation of a prediction can be repeated on a new representation after a suitable transformation. In this case a suitable transformation would be invertible and would yield different statistics than the transformations previously tried. The statistics in question are defined in Section III,C,3, as we discuss a solution to the problem of sparse sampling.

If we are going to argue that a particular approximation is a good procedure, we need some standard for measuring the goodness of an approximation, since there are any number of possible approximation techniques. Thus, at the very least, we need to consider optimal approximations as benchmarks.

2. Optimization and Approximation

In the theory of neural networks, optimization is a pervasive but occasionally misunderstood issue. Engineers working on neurallike networks have an easy problem; they can study optimization from the standpoint of building their own optimal networks. It is more difficult for biologists (see, e.g., Staddon & Hinson, 1983), who must explain the networks found in nature.

We will develop optimization ideas to produce a context for comparing the performance of a particular network in a particular environment to the performance of the best possible network in that same environment. The idea to bear in mind is that a computation which is optimal in one environment may easily fail to be optimal in another environment. Thus, our task as biologists and neuroscientists is to understand when a brain region will produce a good approximation and when it will produce a poor one. To put this perspective to use, recall our local definition of environment for each brain region (i.e., its inputs) and consider the description of an environment as some sequence of states, or even a probability-generating function. Then we can ask, Which probability-generating func-

tions, that is, environments, are expeditiously handled by which particular anatomies and physiologies? If we can answer this question it is natural that we interpret our results under the postulate that the brain is well built. We mean well built in the sense that the appropriate probability-generating functions (environments) are analyzed by those brain regions which do a good job for that class of generators. In other words, we hypothesize not much more than that the visual system of the thalamus and cerebral cortex evolved for analyzing the visual world and that the auditory system of the thalamus and cerebral cortex evolved for analyzing the auditory world. While it is undoubtedly possible for the auditory system to analyze the visual world and vice versa with some good results, it is almost certainly an inferior solution to the computational problems of seeing and hearing. In the present context, then, the argument is *not* that all sequence prediction problems are solved by the hippocampus but that the hippocampus is used when it is better than other brain regions for solving multisensory prediction problems. If the language areas, for example, are better for predicting long linguistic sequences, then the hippocampus will not be used.

The hypothesized statistical environment follows from the problem of sensory signal fusion detailed in Section II. Specifically, the problem is to find information which is in the higher-order relationships, that is, the higher-order statistical moments, or correlations, of the input environment. These information-rich moments are hard to find because they are relatively few in number and are scattered about in an exponentially large space. Furthermore, specific higher-order relationships do not hold constant in the environment from generation to generation, so prewired computational systems cannot evolve to handle them, as they have evolved for the peripheral stages of sensory processing. This hypothesis fits well with the "plain vanilla" morphology of the hippocampus as compared with structures like the retina, the olfactory bulb, and the neocortex which, by virtue of their relative complexity, appear to have inherited more computational biases than the hippocampus.

3. Margination

Because the three resources—time, processors, and samples—are in short supply relative to the pervasive and unavoidable complexity problems and because there is a single approximation technique which solves all three scarcities, we use this approximation technique as a working hypothesis for describing a function which occurs in the hippocampus. The process is called *margination*.

Let us just consider the third difficulty associated with the combinato-

rial explosion which occurs in multidimensional networks: the problem of sparse sampling. Suppose that $X = (X_1, \dots, X_n)$ is an n -dimensional random variable, where each component X_i is binary-valued. The values of X_1, \dots, X_n could represent the zero-one or true-false output of n neurons in a simple network; hence the range space of X consists of 2^n configurations. In the case of small regions of a mammalian brain, n ranges between 1000 and 10,000,000. Even in the case of a moderately large n the presence of noise guarantees that no configuration is sampled, on average, more than once. Furthermore, most configurations will never be sampled because the number of configurations greatly exceeds the number of samples. Thus even in small brain regions there exists a problem of sparse sampling.

A well-known solution to this problem is to reduce the number of components of X by summation so as to obtain meaningful sample sizes. This process is known as margination [see Good (1963) for ideas about generalized margination]. Margination, in contrast to removing configurations from the sample space, is quite conservative since it never allows the network to experience infinite surprise (see below).

Margination can be related to the spatial integration performed by a neuron whose output is either "fire" or "not fire." Consider the set of all subsets of activity patterns which can fire this neuron. This set defines the summation over which margination occurs to create the probability of this neuron's firing.

At this time it is not yet clear which methods are best for selecting the subsets which are the margination process. However, even though mathematicians have not yet offered a tractable, optimal solution to finding these subsets, such methods are a primary concern of the computational approach advocated here.

D. REFINING THE APPROACH: DEFINITIONS AND MEASURES

This section gives some specificity to the theoretical development by defining some terminology and then by introducing some information measures.

1. Current Representations

A neural network can briefly retain information about the current state of the environment (see Fig. 5). This information is represented by the states of the processing elements. If a processing element is a neuron, as opposed to a portion of a neuron, then the state of a neuron means a specific scalar representation, such as (1) the voltage at the axonal initial segment, or (2) whether the neuron fires once within an absolute refrac-

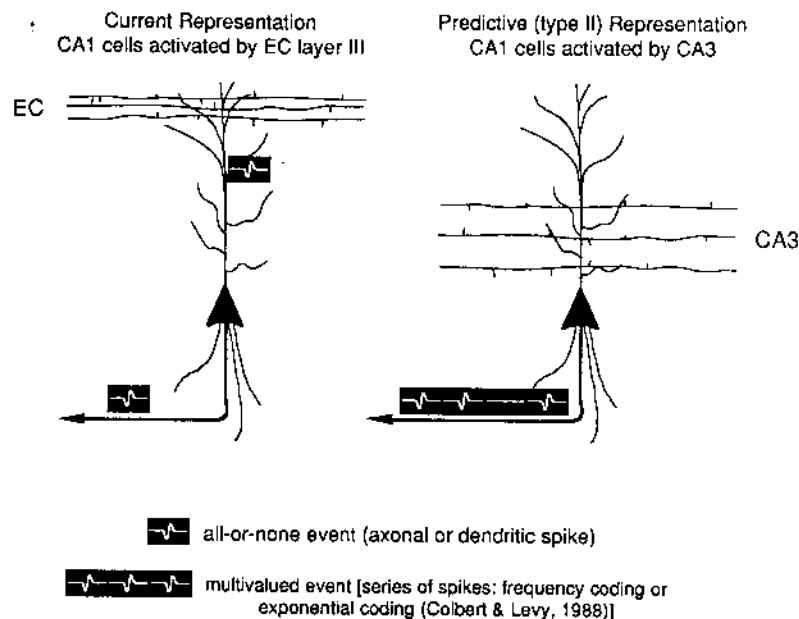


Fig. 5. A single CA1 cell participates in two types of representations: a current representation and a predictive representation. The current representation is monosynaptically activated by the entorhinal cortical (EC) input. The CA1 output of a current representation is pictured as an all-or-none event. In contrast, the predictive representation is monosynaptically activated by the hippocampal CA3 input. A type II predictive representation is illustrated by the multivalued CA1 output; the theory developed in this chapter presumes a type II predictive representation. However, there is no experimental evidence to favor this type of predictive representation over a type I predictive representation (cf. Fig. 6). Copyright © 1989 by William B Levy.

tory period, or (3) the number of firings within some larger, specified interval.

The most general definition of a representation is the state of all neurons in the brain over some very small period of time (e.g., one absolute refractory period). This global representation can be simplified because any subset of a representation is also a representation. For example, the state of all the related neurons of a single structure, such as the state of all CA1 pyramids, is also a representation.

A *current representation* or, more simply, a *representation*, of the environment, means a vector of scalar states of a specified set of processing elements. We will refer to a set of all possible values of such a vector, or the set of neurons which makes this vector, as a *representation space*.

2. Predictive Representations

Since we are interested in networks which generate predictions, or predictive representations (see Figs. 5 and 6), and since communication in these networks is limited by synaptic interactions and the computational

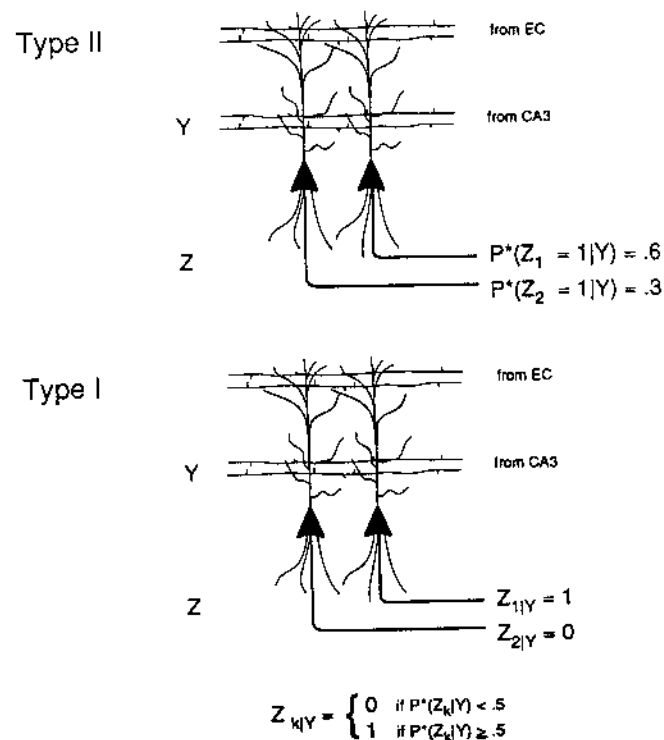


Fig. 6. CA1 cells might mediate either a type I or a type II predictive representation as their output. A type I representation (lower panel) is in the form of a binary-state neuronal output (i.e., either fire or not fire). An optimal type I predictive representation would use a CA1 cell's polarization to represent a probability (or a function of this probability). Each cell's output would then be appropriately thresholded to fire or not fire depending on this polarization state. A type II representation (upper panel) requires an output which can be interpreted as a multivalued output by the next set of neurons.

Y is the state of the CA3 input to CA1. Z is the state of the CA1 output. When Y generates the output Z , a type I or II predictive representation results. The type I output is indicated by $Z_{k|Y}$ (the most likely configuration for Z_k given Y) for each CA1 cell k . The type II output is indicated by $P^*(Z_k = 1 | Y)$, an inferred probability distribution of Z_k given Y , for each CA1 cell k . Note that the equation at the bottom indicates that the type I predictive representation is generated by thresholding a type II representation. Thus, a type II representation can also be identified with dendritic or somatic depolarization of a cell producing a type I predictive representation. Copyright © 1989 by William B Levy.

constraints inherent in the neurons themselves, we must give serious thought to just what form a prediction can take in a neural network. (However, if we identify a brain region, such as CA1, as producing predictions in conformity with the definition of a predictive representation, these predictions may be quite abstract, with no obvious correspondence between cell firing and an animal's behavior. We would be particularly lucky if, for instance, the abstract predictions coincided with what is identified with a prediction of behavioral relevance such as the head direction-predicting responses identified by Ranck, 1985, in the presubiculum.)

Apparently, only three requirements are needed for a definition of a predictive representation which enables us to identify predictions in isolated brain regions.

1. The prediction occurs before the event being predicted.
2. The prediction is specified such that the network can take advantage of the prediction (i.e., there is a mapping, available to the network, between predictions and later current representations).
3. A predictive representation is not confused with the current representation (i.e., predictive and current representations are distinguishable).

Statements (1) and (2) are natural requirements for a prediction. Note that (1) eliminates retrodictions such as the pattern recognition problem. Requirement (2) ultimately provides for the situation in which a larger network that contains a prediction-generating subnetwork can (a) evaluate the quality of the predictions and (b) use these predictions as a basis for action upon the environment (e.g., to modify the external world to prevent an undesirable event from occurring). Requirement (3) does not seem an obvious necessity, but if the predictive representation space and the current representation space are indistinguishable, then the network cannot evaluate its predictions nor learn from experience, because it cannot separate predictive from current representations. Note, however, that these two spaces must be distinguishable and still satisfy (2). Requirement (2) and the local principle, which prevents memory access by direct addressing, necessitate a mapping between each predictive representation and each current representation by the method of shared representation space. (The neurons which produce the predictive representation must be contained in the set of neurons which produce the current representation: e.g., the CA1 spiny pyramids illustrated in Fig. 5.)

Requirements (1), (2), and (3), taken together, define a predictive representation. Thus, a set of neurons connected within a network in conformity with this definition is a predictive representation space. It is a repre-

sentation space because of the mapping to a current representation space, and it is predictive because it anticipates a future event.

The definition given above does not attribute accuracy to or require accuracy from a predictive representation, so the quality of the prediction does not affect its identifiability as a predictive representation. Of course, we are really interested in the generation of accurate predictive representations. To assess the quality of the predictions we will need to define some measures. First, however, we will be a little more concrete and describe a subcategory of predictive representations which are of particular interest to us.

There appear to be two types of predictive representations worth distinguishing (see Fig. 6). The most obvious prediction, which we will call type I, is a representation of the one most probable configuration about to occur. Thus, if a current representation can be defined in a multivariate binary space, $\{0,1\}^n$, then a type I prediction occurs in the same multivariate binary space, $\{0,1\}^n$.

While the neurons generating a type I prediction produce a single, multivariate, binary event, the neurons generating a type II prediction produce a probability distribution of all possible such events. Though the type I prediction may seem the most natural form of a predictive representation, it is not the most general form, so it is the type II predictive representation which we emphasize in our theories. The type II prediction is a representation, ideally, on the interval $[0,1]^n$ (although in practice physical realities force the nervous system to make do with a discrete approximation of these n continuous intervals).

This second type of predictive representation attracts our interest for three reasons. First, the type I, most likely configuration, can be directly inferred from the type II prediction, but in general the reverse is not possible. Second, the type II prediction is a necessary intermediate step which the network must derive in order to measure the surprise suffered when an event in the predictive representation space occurs (i.e., the type II prediction is needed to calculate entropies on the relevant space, where a space means a specified set of neurons). Third, although the type I prediction is ultimately necessary for decision making and action on the environment, the type I decision should always be postponed as long as possible in a highly complex multilayered network to avoid the information loss which accompanies such a representation.

The above considerations suggest the following definition: A type II predictive representation is, perhaps implicitly, a vector of conditional probabilities; the conditioned or predicted variable is some future representation, and the probabilities of a type II predictive representation are conditional on the current representation.

3. Information Theory and Prediction

For the discussion which follows we will consider a neuron as a binary device. $P(\cdot)$ is an unconditioned probability distribution. $P(\cdot | \cdot)$ is a conditional probability distribution. $P^*(\cdot)$ is an inferred probability distribution as opposed to the true distribution. $P^*(A | B = b)$ can be read as "the inferred probability of A given B takes on value b ," and is the probability distribution over the finite set of events implied by A given that the particular event b in B has occurred. The variable A is the conditioned or predicted variable; B is the conditioning variable.

The choice of probability as the way to quantify predictions is not arbitrary. The set of relationships called probability theory is the only consistent method using a scale from zero to one for quantifying, manipulating, updating, and predicting events (Cox, 1961, 1978; Jaynes, 1978). However, probability is not the only measure we need to consider.

We, and ultimately the network itself, also need a tool for measuring the quality of a network's predictions. This measure turns out to be equivalent to quantifying the information in a prediction. Instead of using the somewhat ambiguous term "information" for this measure, we prefer Hamming's (1980) suggestion of the term "surprise." Surprise, a nonnegative scalar, occurs with each current representation. In the prediction problem, the network is trying to avoid, or minimize, surprise so that surprise is a loss function. The more probable the event which occurs, the less surprise the network suffers. If the network places a probability of one on an event which then occurs, the surprise is zero. If the network places a probability of zero on an event which then occurs, the surprise is infinite. We would like surprise to be monotonic and continuously decreasing in probability, and we would like independent events to have additive surprise. Although a more rigorous justification is possible (see Mathai & Rathie, 1975), in this presentation it is sufficient to define surprise as $-\log P^*(Z = z | X = x)$ where P^* is the probability held by the network that representation z in the space of the Z neurons will occur given that representation x in the space of the X neurons has occurred. We identify the Z space with CA1 and the X space with inputs to the preprocessor that is the DG/CA3 region (see Section V).

With the occurrence of each individual sequence pair (Z, X) , the information-theoretic loss due to the occurrence of z preceded by x is $-\log P^*(Z = z | X = x)$. Over time, then, the performance of the network as a prediction device is the average of this measure:

$$H^*(Z | X) = -\sum_{zx} P(Z, X) \log P^*(Z | X). \quad (1)$$

In an important sense it is the mathematical properties and associated theorems which justify the use of this quantity, average surprise, to characterize the performance of a network. However, we postpone its rigorous mathematical characterization to another time. It is important to note only that we are limiting the theory here to the pure prediction problem, which does not allow a feedback relationship between the network and the environment.

If no prediction is made by the network and we allow a prior unconditioned probability to be built into the network (see, e.g., rule 2 in Levy & Desmond, 1985a), the average surprise of a current representation is Shannon's entropy $H(Z)$ (Shannon & Weaver, 1949).

$$H(Z) = E[-\log P(Z = z)] = -\sum_{z \in Z} P(Z = z) \log P(Z = z), \quad (2)$$

where $E[\cdot]$ denotes expectation. Therefore, $H(Z)$ is the average, naive surprise or, just as well, $H(Z)$ is the average information in a current representation space Z .

Although the network does not need to measure the information loss of preprocessing, the theoretician does. To measure the average information loss of a transformation which takes X into Y , we use Shannon's conditional entropy $H(X | Y)$, where X is the set of all possible input signals and Y is the set of all corresponding output signals.

$$H(X | Y) = -\sum_{xy} P(X, Y) \log \frac{P(X, Y)}{P(Y)} \quad (3)$$

$$= H(X, Y) - H(Y). \quad (4)$$

This conditional entropy has several interesting properties, although for now we note only two. First, when Y is created by an invertible transformation f on X , then $H(X | Y) = 0$, the smallest possible value. Invertibly formed representations must have zero information loss, since inverting a transformation, $f^{-1}(f(X))$, to obtain what we started with, X , proves that the transformation f lost no information. Second, when Y is generated without any relationship to X by a uniform random process, then $H(X | Y) = H(X)$, the largest possible value. Thus no more information can be lost than was present to begin with.

As an aside, we note that these measures and the associated theorems defining their properties begin to explain what is meant by "information processing." Because of these multiple measures, there is more than one kind of information in a neural network: representation information and

predictive information. Thus if someone were to ask how to calculate the information stored at a synapse, we could not answer the question without specifying the kind of information of interest. In either case the calculations show that the information is relative. In the case of a predictive representation, we would produce the difference between $E[-\log P^*(Z|X)]$ and $E[-\log P'^*(Z|X)]$, where P'^* is calculated with the synapse in question removed. In the case of representation information, we would subtract $H'^*(A|B)$ from $H^*(A|B)$, where H'^* is calculated for the network with the synapse removed.

Because of the way we hypothesize a network generates type II predictive probabilities, we are interested in the statistical dependence of representation spaces, in particular the representation spaces of CA1 and CA3. The measure of statistical dependence that we use (Watanabe, 1969) quantifies how much all individual neurons Y_i in Y predict about all of Y .

$$\sum_Y P(Y) \log \frac{P(Y)}{\prod_j P(Y_j)} = \sum_j H(Y_j) - H(Y) \quad (5)$$

(This quantity is the obvious extension of Shannon's mutual information, the measure of dependence of a bivariate quantity.) Note that statistical dependence is a nonnegative quantity which is zero when full independence obtains. A deeper appreciation of the properties of this quantity comes from the fact that it is a divergence (Csiszár & Körner, 1981).

Now we put some of these ideas to use.

E. THE FORM OF OPTIMAL APPROXIMATE INFERENCE

1. Should a Network Preprocess?

Suppose multivariate binary inputs Z and X are given to any neurallike network as a sequence of inputs, first X and then Z . Suppose the network is to predict about Z using X . If we adhere blindly to conventional information theory, we might say the problem is to minimize $H(Z|X)$. However $H(Z,X)$ is fixed by some process outside the network and so is $H(X)$, so it seems that there is little the network can do. In fact, anything the network might do—for example, a transformation $f: X \rightarrow Y$ in order to minimize $H(Z|Y)$ instead of the original minimization—risks destroying the predictive information that exists between X and Z , and such a transformation certainly cannot add new information. [In fact, it is relatively easy to produce an analog of information theory's data processing lemma (Csiszár & Körner, 1981) for neurallike networks in this regard.] Yet the

brain makes many such transformations. So where has our thinking gone awry?

The problem is that we have forgotten about the combinatorial explosion which forces sparse sampling and approximate computations. As a consequence the network can never know or use the true probabilities required for calculating the entropies of Shannon's information theory. The computational intractability problems discussed previously (see Section III, C) force the network to use an approximation approach. In Section V we hypothesize that this approximation ends up being equivalent to an independence assumption conditioned on each individual neuron whose outputs constitute the predictive representation space.

Since the computation is equivalent to an independence assumption, the closer the statistics of each representation space approach independence, the smaller is the error from the approximation. The trick, then, is to minimize dependence, that is, to approach independence as closely as possible (Levy, 1985). This is done by first preprocessing representations with minimal information loss $H(X|Y)$. Then the dimensions (neurons) of these transformed representations are used for prediction in a manner that appears to presume independence.

We now present the final theoretical reason which brings us to accept the computational form which is equivalent to an independence assumption.

2. Optimal Inference of a Probability Distribution from Averages

This section justifies our network's method of forming posterior probability distributions from moment constraints (e.g., adaptively stored means and correlations of the Z and X variables) which are locally available as the synaptic weights (or strengths). In fact, there is essentially only one correct method for generating probability distributions from averages. This method of inference is called minimum relative entropy (MRE) inference (Johnson & Shore, 1983; Shore & Johnson, 1980). [MRE inference is equivalent to maximum entropy inference (Jaynes, 1978) under a uniform prior.] Since we are interested in justifying an optimization procedure, that is, an optimal approximate computation for prediction in the neural network, the results of Shore and Johnson (1980; Johnson & Shore, 1983) are most relevant. They have proven that minimum relative entropy is essentially the only correct optimization criterion to produce a probability distribution when we start with a specified, supporting state space, a prior distribution, and moment constraints. Optimal procedures and results, such as the above claim, depend on the assumptions which are made. We assume that the implied structure and computation of an

optimal network is of most interest when it arises from a minimum of assumptions; that is, the optimal results follow from first principles.

For the discrete case, the required axioms are

1. **Uniqueness:** The optimal inference procedure working from a set of moment constraints (e.g., correlations, variances) must produce a single distribution.
2. **Idempotence:** Repeated applications of the optimal inference procedure with the same moments do not change the resulting distribution from the first application.
3. **System independence:** If two sample spaces are disjoint and there are correspondingly disjoint moment constraints, then the optimal inference procedure must produce the same distribution regardless of whether the full joint distribution is formed before or after applying the inference procedure to the moment constraints.
4. **Subset independence:** If the state space can be decomposed into disjoint marginal subspaces and the constraints can be decomposed similarly, then the probability distribution inferred should be the same regardless of whether the optimization procedure is applied to the full space before or after marginalization.
5. **Invariance:** If f is any invertible transformation, and $Y = f(X)$ so that Y is a lossless representation of X , then the optimal inference procedure must produce the same probability distribution whether we work directly with X or work with Y and then apply the inverse transformation $f^{-1}(Y) = X$ to the probability distribution inferred to Y .

These axioms lead to the conclusion that the form of a probability distribution inferred from averages is multiplicative. That is, suppose we have a set of functions, g_i , of the random variable X and their expectations, $\{E\{g_i(X)\}\}$. MRE then says that the optimal probability distribution is

$$P^*(X = x) = e^{-\lambda_0} \prod_i e^{-\lambda_i g_i(x)} \quad (6)$$

where the λ values are parameters which must be determined. Thus, if synaptic modification leads to something that is a good approximation of an average, as we and others posit, the optimal distributional form is of this multiplicative form.

From the computational viewpoint, and particularly for a neurallike network, determination of λ_0 appears to be a computationally intractable problem for arbitrary sets of expectations over a multivariate space because of memory requirements that can grow exponentially with the num-

ber of dimensions of the space. However, when the expectations are limited to a particular class of expectations the combinatorial problems disappear. Specifically, the set, or a subset, of the lowest-order marginals leads to a very simple form. For the case of interest to us, in which the expectations are of the form $E[X_i | Z_k = 1]$ rather than the unconditioned form of Eq. (6) and the X_i values are binary-valued variables, the probability of the event $X = x$ is

$$P^*(X = x | Z_k = 1) = \prod_i E[X_i | Z_k = 1]^{x_i} \cdot (1 - E[X_i | Z_k = 1])^{1-x_i}. \quad (7)$$

Here the marginals are summed over half the possible states of the X space. Thus, these are the lowest-order marginals and so avoid the combinatorial explosion. That is, these marginals produce reasonable sample sizes, and they avoid an apparently intractable calculation of λ_0 .

It will be obvious to those knowledgeable in statistics that MRE inference applied to such marginals generates a probability distribution which has the independent form. This motivates two comments. First, MRE justifies the independence assumption which is often invoked for computational convenience. Second, it might now appear that we have limited the computational options so severely that a neural network can do nothing to help prediction. However, this is not the case. MRE gives no prescription that would tell us, or a network, which set of lowest order marginals to use. In fact, when we consider the possibility of permuting the X space, the available choices are on the order of 2^{2^n} different sets of lowest-order marginals. Thus, rather than limiting our options, we may have, at best, an embarrassment of riches or, at worst, another case of computational intractability.

Because our hypothesized computation in CA1 is the MRE-inferred independent form, preprocessing should produce a representation with minimal statistical dependence. In Section VI we outline an algorithm which is hypothesized to be the transformation performed by the DG/CA3 preprocessor. This algorithm minimizes the statistical dependence of the conditioning variable by creating an optimal set of lowest order marginals.

IV. Summary of the Computational Issues Relevant to the Model

We now give a brief summary of the computational theory of the hippocampus in question-and-answer form.

1. *What problem is the hippocampus solving?* The hippocampus solves a prediction problem in which the present environment, here the

signals entering the entorhinal cortex, and the stored statistics of past environments are used to predict a future event. To solve the problem we require the network to generate a predictive representation.

2. *How is the present represented?* The present is represented by the activity state of any subset of neurons, in our case the CA1 spiny pyramids.

3. *How is the past represented?* The past is usually represented in the adaptively modified synaptic weights. In addition, there may be an internal excitability parameter associated with each neuron that is adaptively adjusted by each neuron's activation history.

4. *How is a prediction represented?* A prediction, or predictive representation, precedes a current representation and is the activity vector of a subset of the same neurons that are part of the space used to represent the present (the subset is CA1 itself). However, a predictive representation is directly activated by a different set of inputs, the Schaffer–commissural efferents of CA3, than those inputs which directly activate the representation of the present, the layer III pyramids of the entorhinal cortex. Since the two types of representations share the same neurons, it is possible to map and evaluate the quality of a prediction.

Two types of predictive representations can be envisioned (see Fig. 6). A type I predictive representation is the representation predicted to be the most likely pattern of activity to follow. The type II predictive representation is more sophisticated; it implies a probability distribution over the next set of possible representations. The more sophisticated predictive representation can easily be converted to the maximum likelihood type I representation by a threshold process, $\{0,1\}$, in each neuron. More importantly, a type II predictive representation is converted to a surprise measure when the current representation being predicted finally occurs.

5. *Are there any special physiological characteristics of cells that mediate prediction?* The synapses which evoke the predictive representations should modify as a reinforced system with a special timing requirement for synaptic potentiation: The permissive postsynaptic event can follow but not precede the synaptic activity which it reinforces. If the associative coactivity requirement were not time-ordered, the prediction could get reversed, and the network might end up predicting the past.

6. *How does a neuron calculate a probability?* The physiological input–output function of a neuron produces the multiplications (actually it adds logarithms) required by MRE and by Bayes's equation for inverting (reversing) the conditional probabilities implicitly stored at each synapse. (That is, the synaptic strengths are proportional to the conditional expectations which imply conditional probabilities via MRE.)

7. *What are the characteristics of a predictive representation?* (i) It occurs before the "actual" representation. (ii) It requires circuitry that

performs time-shifting in order to take advantage of associative synaptic modification within a limited time window. (iii) A predictive representation is useful if it is constructed from statistics (i.e., past observations) based on the laws of inference, such as $P(A|B) = P(B|A)P(A)P(B)^{-1}$. (iv) It can generate outputs almost as if the actual representation has occurred. (v) It can be compared with the actual representation to evaluate the quality of the prediction. Characteristics (iv) and (v) require that there be a one-to-one mapping from the neurons which form the predictive representations to a subset of the neurons which form the actual representations. Furthermore, there is a requirement for a mechanism which keeps the two types of representations distinctive.

8. *What is a preprocessor?* A preprocessor (see Fig. 8) is a system that recodes (transforms) signals in an essentially unsupervised manner. Its synaptic modifications are self-supervised and are not dependent on the specific information content of the next stage in processing.

9. *Why preprocess representations?* There are at least two reasons: to shift a representation in time and to change the form of a representation to make it more compatible with the succeeding computation.

V. The Model

Figure 7 depicts the model we have been developing (Levy, 1988) using abstract blocks which correspond to specific parts of the hippocampus. Except for the prediction evaluation block, the blocks are identified with particular regions of the hippocampus (see Fig. 2A). However, we do not mean to disqualify the possibility that the prediction evaluation generator resides within the limbic system.

The centerpiece of the model is the prediction tier, which consists of the principal neurons of CA1, the spiny pyramids. This group of cells produces two types of representations: a current representation and a predictive representation. When generated as a monosynaptic response to entorhinal layer III pyramidal cell activity, the output of CA1 is a representation of the current environment. When generated as a monosynaptic response to Schaffer–commissural CA3 activity, the output of CA1 is a predictive representation.

Because we are dealing with an abstract idea of prediction, the model works just as well whether we consider the axonal output of CA1 or a dendritic event in the CA1 pyramids as the predicted variable. (In keeping with mathematical usage and the idea that CA1 generates a conditional probability, such as $P(A|B)$, the term predicted variable, A in the example, will refer to current representations in CA1; the term conditioning variable, B in the example, will refer to the CA3 Schaffer–commissural

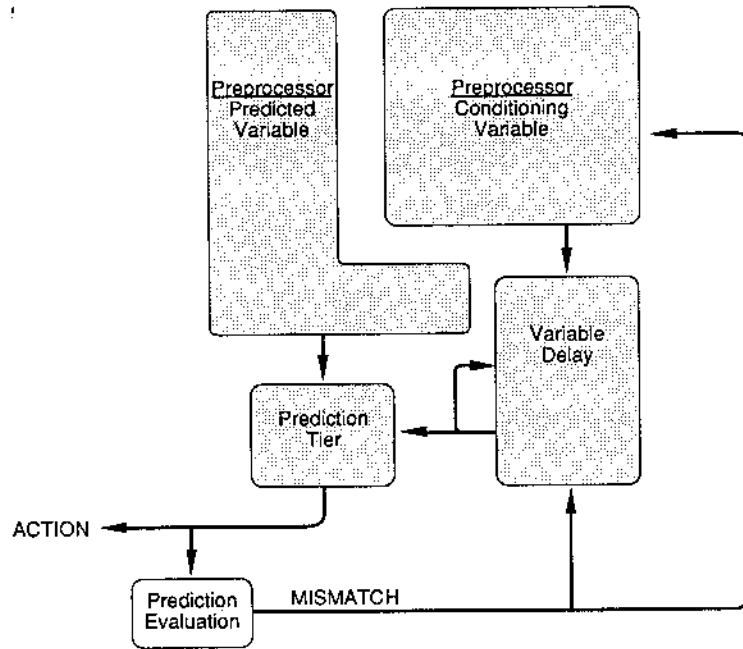


Fig. 7. Here the model network of Fig. 2B is shown with the addition of a prediction evaluation stage, which controls the rate of adaptive modifications Copyright © 1989 by William B Levy.

inputs to CA1; see Fig. 8.) Quite important to the mathematical details of the model here is the postulate that the event being predicted in CA1 is a binary-valued variable for each neuron regardless of which neuronal state is considered to be the predicted variable. Thus, in each CA1 cell, the predicted variable is a threshold-defined event, perhaps the action potential of the soma and initial segment or perhaps a dendritic spike.

There are two preprocessing systems: one for the conditioning variables and one for the predicted variables (see Fig. 8). These preprocessing systems do not correspond neatly to any single tier of principal neurons. For the predicted variable, preprocessing, which includes signal mixing and reducing statistical dependencies, occurs in the layer III cells of the entorhinal cortex, the distal dendrites of the CA1 principal neurons, and within the interneurons of CA1 stratum lacunosum. For the conditioning variable, preprocessing, which involves signal mixing, remixing, reducing statistical dependencies, and time-shifting, occurs in the layer II cells of the entorhinal cortex, the dentate gyrus including its infragranular layer (CA4 of Lorente de N6), and the CA3 region.

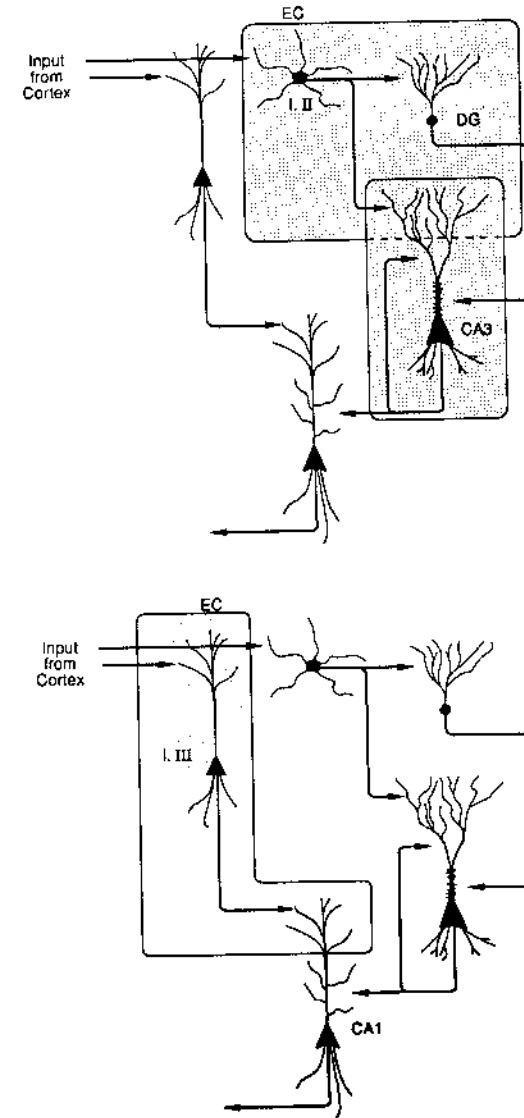


Fig. 8. Schematic illustration of the conditioning (upper panel) and predicted (lower panel) variable preprocessors. In the theory, the CA1 cells generate conditional probabilities and thus constitute the prediction representation space. Such conditional probabilities have two variables, a predicted variable and a conditioning variable, each of which can be preprocessed by the appropriate preprocessor. Copyright © 1989 by William B Levy.

It is also necessary for the conditioning variable preprocessor to perform a type of pattern recognition in order to be robust in its primary preprocessing task. More information about this pattern recognition process is found in Section VI. However, it is worth noting here that the pattern recognition process proposed below is not the information-losing, retrodictive process of categorization or classification that describes the function of many other neurallike networks. Instead it involves shifting the DG/CA3 activity state nearer previously used representations when a mismatch signal exists and when the current environment is similar to a previously experienced environment.

The prediction evaluation generator, as the mismatch-generating subsystem (shown in Fig. 7), keeps a running average of the amount of surprise that results with each successive current representation. When the accumulated, running average surprise exceeds a certain prespecified threshold level, the prediction evaluation generator activates a mismatch signal. To generate the correct amount of surprise, this evaluator needs to receive both current and predictive representations, that is, the outputs from the CA1 prediction tier. A more sophisticated version of the prediction evaluation generator, worthy of both experimental and theoretical investigation, posits a mismatch signal which controls the ease of modifiability and state change along a continuum which varies monotonically with the running average surprise. Because the mismatch detector's output is, in either case, a rather low-dimension signal, it differs from the high-dimension signal used in predictor-corrector systems or back-propagation schemes (Barto, Sutton, & Anderson, 1983; Rumelhart, Hinton, & Williams, 1986). However, the mismatch signal used here is similar to Sutton's adaptive critic (Sutton, 1984) because it quantifies the overall quality of the multidimensional output of the prediction tier. Mismatch generation is central to other theories as well, such as those of Gray (1982), Brooks (1986), and Carpenter and Grossberg (1987).

The mismatch signal is of low information content and should be thought of as a low-dimension output compared to all other signal lines illustrated in Fig. 7. The leading candidate for this signal is the septal input to the hippocampus; other candidates are the various monoaminergic inputs to the hippocampus.

When the statistics of the environment shift, the prediction evaluation generator at the output end of the network signals mismatch (see Fig. 7). If the pattern recognition process triggered in the DG/CA3 region by the mismatch signal fails to find anything recognizable, the CA3 activity state takes a random jump to another state and the mismatch signal allows synaptic modification in the DG/CA3 system.

An ancillary support function is also required of the network by this theory and, like the mismatch signal, may be one of the low-dimension nonspecific neuronal systems mentioned above. Any low-dimension signal, perhaps one mediated through the entorhinal cortex, seems like a reasonable possibility. The required support function, in effect, provides a distinctive label so that the prediction evaluation generator is able to tell the difference between predictive and current representations. This signal's function is to (1) force these two representation types to alternate as they arrive in CA1 and (2) inform the prediction evaluation generator which signal from CA1 is a predictive representation and which is a current representation. Alternation could be at regular or variable intervals without affecting the proposed scheme.

The shared representation space, which has CA1 alternating between predictive and current representations, is the process by which the network satisfies properties (2) and (3) of the definition of a predictive representation (see Section III, D, 3).

With more knowledge, including quantitative anatomy and the appropriate physiological observations, the present hippocampal model could be expanded to incorporate computations performed in the subicular regions (see Fig. 1), entorhinal cortex, and perhaps the modern Papez circuit. Although anatomical and physiological observations may suggest additional computational issues, many of the same types of computational strategies (abstract prediction and preprocessing to mix, time shift, and lower the statistical dependence of signals) would be reapplied. As currently envisioned, the reapplication of these strategies would produce a more interesting theory because of expanded network capabilities. These expanded capabilities extend the range of the predictions so that the predictions being generated concern events further in the future. The added circuitry clearly provides (see Fig. 1) more high-dimension, positive feedback loops which could mediate additional signal remixing and time-shifting.

A. THE CA1 COMPUTATION

Let the CA3 inputs, j , to CA1 be binary-valued variables, $\{0,1\}$, with the state of the j th input at time t designated as $Y_j(t)$, and suppose that Y is formed from entorhinal or cortical inputs X as $f: X \rightarrow Y$ with an f such that $f^{-1}(f(x)) = X$. Let $Z_k(t+1)$ be a binary-valued variable which denotes the state taken by a CA1 neuron k in response to a layer III input from entorhinal cortex. The strength of the synapse between j and k is denoted as $W_{jk}(t)$. Suppose that synaptic modification of the CA3-CA1

synapses occurs so that each synapse takes on values which are proportional to a conditional correlation. For example suppose, as is detailed below, that this class of synapses modifies according to the equations:

$$W_{jk}(t + 1) = W_{jk}(t) + \Delta W_{jk}(t, t + 1) \quad (8)$$

and

$$\Delta W_{jk}(t, t + 1) = \epsilon \cdot Z_k(t + 1) \cdot [Y_j(t) - c \cdot W_{jk}(t)]. \quad (9)$$

Then synaptic strength will asymptotically converge to $\bar{P}(Y_j = 1 | Z_k = 1)$, that is, approximately the value $E[Y_j | Z_k = 1]$. Therefore each $P(Y_j | Z_k = 1)$ is, at least implicitly, available at each synapse (jk).

At this point Bayes's theorem and MRE inference instruct us to hypothesize one particular computation when we work under the supposition that CA1 generates an optimal type II predictive representation which minimizes average surprise. This computation is defined by Eqs. (10)–(12). For each CA1 neuron k ,

$$P^*(Z_k = 1 | X) = P^*(Z_k = 1 | Y) = \frac{P^*(Z_k = 1, Y)}{P^*(Z_k = 0, Y) + P^*(Z_k = 1, Y)} \quad (10)$$

The first equality follows from the invertibility of f . The second equality is just Bayes's statement. Equations (11) and (12) follow from MRE inference, from the postulate just above that averages are stored at the jk synapses (see also Section V,B,2), and from the added presumption that the average activity level $E(Z_k = 1)$ is available at each CA1 neuron. (These points and their biological plausibility are discussed in greater detail in Levy *et al.*, in press.)

$$P^*(Z_k = 1, Y) \leq \bar{P}(Z_k = 1) \prod_j \bar{P}(Y_j | Z_k = 1)^{Y_j} [1 - \bar{P}(Y_j | Z_k = 1)]^{1 - Y_j} \quad (11)$$

$$P^*(Z_k = 0, Y) \leq \bar{P}(Z_k = 0) \prod_j \bar{P}(Y_j | Z_k = 0)^{Y_j} [1 - \bar{P}(Y_j | Z_k = 0)]^{1 - Y_j} \quad (12)$$

where \leq indicates the MRE inferred probability which is optimal in the context of Shore and Johnson (1980), and \bar{P} indicates the sample frequency or the Bayesian, sample-based, updated average.

We note that we have a form quite similar to those seen in Hopfield (1987), Hinton and Sejnowski (1983), and Golden (1988) because our assumptions and MRE yield the conditional independence assumed by these other approaches.

The reader has probably noticed that Eqs. (10)–(12) are written in terms of individual cells, k , but that the probability distribution desired, $P^*(Z | X)$, is over the aggregate of these individual neurons. If a postpro-

cessing network such as the indicated prediction evaluation generator (see Fig. 7) just adds up surprise across the CA1 neurons, it is essentially making an assumption of independence. We can equivalently think of such an additive function in the prediction evaluation generator as obeying the dictates of MRE when no further information is available. However, a more sophisticated generator might be able to bring further information to bear as it evaluates the output from CA1. If the prediction evaluation generator has stored statistics which indicate some of the dependencies between the CA1 neurons, k , then the outputs of these cells can be appropriately weighted to account for such interactions. Whatever the case, the binary version of the generator generates a relatively low-dimension, mismatch signal when there is too much surprise.

In hypothesizing Eqs. (10)–(12) or their equivalents, as computations performed in CA1, we hypothesize the existence of several neuronal functions which have not been identified nor even looked for by neurobiologists but which seem at least feasible. Equation 11 does not quite look like a computation of a neurallike processing element until we change the multiple products to multiple sums via logarithms and exponentiation. For instance, the multiplicative form is easily accommodated by spatial summation if synapses perform a logarithmic operation. Spatial summation might occur as

$$\log P^*(Y, Z_k = 1) \stackrel{\log \bar{P}(Z_k = 1) +}{=} \sum_j Y_j \log W_{jk}^1 - (1 - Y_j) \log(1 - W_{jk}^1) \quad (13)$$

for Eq. (11), and

$$\log P^*(Y, Z_k = 0) \stackrel{\log \bar{P}(Z_k = 0) +}{=} \sum_j Y_j \log W_{jk}^0 - (1 - Y_j) \log(1 - W_{jk}^0) \quad (14)$$

for Eq. (12) with the appropriate exponentiation following these steps and where

$$W_{jk}^1 = \bar{P}(Y_j | Z_k = 1) \text{ and } W_{jk}^0 = \bar{P}(Y_j | Z_k = 0). \quad (15)$$

The idea that the synapse performs a logarithm is plausible since the generation of voltage, as described by Nernst and Goldman-type equations, is a logarithmic function of conductance. More unusual is the consideration that an inactive synapse [note the $(1 - Y_j)$ terms] contributes to the computation. To produce this interaction, we hypothesize that the synapses are on dendritic spines and that the nonsynaptic conductance of the spine membrane can act as a current source of some significance.

The denominator in Eq. (10) is notably problematic for the theory since

it is not obvious where the terms $\bar{P}(Y_j | Z_k = 0)$ are created (however, see Levy *et al.*, in press, for further discussion). We can offer several suggestions including (1) a new modification rule at each synapse jk ; (2) implicit generation of $P^*(X)$ (see, e.g., Levy & Desmond, 1985a, rule 2) which eliminates the need for $P^*(Y | Z_k = 0)$; (3) inhibitory synaptic modification which apparently has the difficult, perhaps impossible, task of sorting and weighting the terms j for each k ; (4) a process which guarantees that the denominator remains constant; or (5) a mathematical short-cut which bypasses the computation in the denominator altogether.

It should be pointed out that if Eqs. (10)–(12) are implemented, each cell k can work with any subset $\{Y_j\}$ of the complete set Y and still, via MRE inference, produce a prediction conditioned on the full Y space. Note that the missing Y_j inputs will cancel each other out in Eq. (10) when they assume their MRE value, so full connectivity from CA3 to CA1 is neither postulated nor required.

$\bar{P}(Z_k = 1)$ is postulated to be adaptively encoded at each cell k . This is essentially the postulate that a cell knows about its own activity history and modifies its excitability threshold accordingly. In studies of long-term potentiation there is a phenomenon which suggests this postulate. The phenomenon is observed as a shift in the amount of synaptic excitation needed to fire cells and is called an i-o (for input conversion into output) curve shift (see, e.g., Wilson, Levy, & Steward, 1979, 1981). The essential observation of these studies is that, following high-frequency activation of the excitatory inputs to neurons in the dentate gyrus, less synaptic activation is needed to fire the cell than before the induction of long-term potentiation.

B. ASSOCIATIVE SYNAPTIC MODIFICATION

Associative synaptic modification is central to the present theory. Most of our detailed understanding of the characteristics of the rule(s) which govern associative modification comes from electrophysiological research in the dentate gyrus which, by analogy, applies to the CA3 synapses on the CA1 spiny pyramids. To some extent, understanding comes from newer studies in CA1 itself.

1. Self-Supervised Modification

Most of the synapses in the present model are presumed to modify in an essentially unsupervised (perhaps better expressed as a self-supervised) way. The three distinct self-supervised systems in the hippocampus are the entorhinal synapses formed on the cells of the dentate gyrus, on the cells of CA3, and on the cells of CA1. There is abundant evidence

for associative modification in the dentate gyrus (see Levy & Desmond, 1985b; Desmond & Levy, 1988, for reviews). Here postsynaptic excitation is permissive for change and presynaptic activity determines the amount and sign of the modification. In CA3 there have been no associative synaptic modification experiments of the entorhinal synapses, but the anatomical similarities between the entorhinal synapses of the dentate gyrus and those of CA3 lead us to hypothesize the identical nature of their associative modifiability. Finally, there are two reports of synaptic potentiation at the entorhinal–CA1 synapses (Doller & Weight, 1985; King & Levy, 1986).

Self-supervised synaptic modification is useful for reducing statistical dependency in representations because such modification tends to combine the effects of converging and correlated inputs on cell firing. This type of modification also assists the pattern recognition aspect of the network.

Even though associative modification in the dentate gyrus appears to be self-supervised, the research in the dentate gyrus can be used as an analog for the reinforced type of modification which is proposed to exist at the CA3–CA1 synapses. This analogy is possible because the experimental paradigms used in the dentate gyrus use a weak and a strong entorhinal input to create many of the characteristics of a reinforced system. Some of these characteristics include (1) the nonlinear, permissive nature of convergent excitation (Burger & Levy, 1987; Levy & Burger, 1987a, 1987b; Levy & Steward, 1979; McNaughton, Douglas, & Goddard, 1978; Wilson *et al.*, 1979, 1981); (2) the existence of a long-term depression which complements long-term potentiation so that an erasure mechanism exists which allows the synapse to function as a running averager (Levy, Brassel, & Moore, 1983; Levy & Steward, 1979, 1983); (3) a specific timing rule which defines the meaning of “associated” in the temporal domain (Levy & Steward, 1983); and (4) a limited amount of interaction along the proximo–distal dendritic axis which defines the meaning of “associated” in the spatial domain (White, Levy, & Steward, 1988).

2. Synaptic Weights as Averages

The experimental observations of Levy and Steward (1979) and Levy *et al.* (1983) using the entorhinal–dentate gyrus synapses directly support a specific class of synaptic modification rules which are related to a variety of earlier proposals (e.g., Amari, 1977; Grossberg, 1976; Kohonen, 1972, 1984). The formulation is best written as Eq. (8) above and as:

$$\Delta W_{jk}(t, t + 1) = \epsilon f(Z_k) [Y_j - cW_{jk}(t)]. \quad (16)$$

Equation (16) is the simplest member of a class of equations which fits the experimental observations and has the properties desired by many theoreticians. The argument t is time in some discrete units and is implicit in terms Z_k and Y_j . The variable W_{jk} is the strength of the synapse formed between afferent j and postsynaptic cell k . The variable $\Delta W_{jk}(t, t + 1)$ is the change in synaptic strength over one unit of time. The variable c is a positive constant of appropriate units so that Y_j and the product $[cW_{jk}(t)]$ have the same units; ϵ is a small positive number; Y_j , the presynaptic input, is nonnegatively valued. If $f(Z_k) > 0$, the difference $(Y_j - cW_{jk})$ determines whether potentiation or depression of synaptic strength W_{jk} occurs. This difference term also keeps the synaptic strength W_{jk} within limits, that is, between 0 and the maximum value of Y_j/c . The argument Z_k is some postsynaptic event in the k th neuron. The postsynaptic term $f(Z_k)$ is nonnegative and nondecreasing in the argument Z_k so that this term is permissive for change. For an unsupervised system, $f(Z_k)$ is just a function of the inputs and their synapses on k , as $f(Y, W_k)$. For a reinforced system, $f(Z_k)$ is a function of a set of inputs and synapses on cell k other than the set described by Y_j . Further details on this distinction between unsupervised and reinforced forms of synaptic modification are found below (Section V,B,3).

If the environment is stationary and strong mixing, it is sensible to average both sides of the equation over time (indicated by expectation operator $E[\]$) (see Geman, 1981; Levy & Geman, 1982, for examples). Equation (16) then becomes

$$E[\Delta W_{jk}] = E[\epsilon f(Z_k)(Y_j - cW_{jk})] \quad (17)$$

$$= \epsilon E[f(Z_k)Y_j] - \epsilon c E[f(Z_k)W_{jk}]. \quad (18)$$

If ϵ is small enough, W_{jk} changes very slowly. We can then rewrite Eq. (18) as

$$E[\Delta W_{jk}] = \epsilon E[f(Z_k)Y_j] - \epsilon W_{jk} E[f(Z_k)] \quad (19)$$

where c has been given the value of one.

In a stationary, strong mixing environment, $E[\Delta W_{jk}]$ converges to zero so that we have, after rearranging and dividing,

$$W_{jk} = \frac{E[f(Z_k), Y_j]}{E[f(Z_k)]}. \quad (20)$$

Now, if $f(Z_k)$ can only take on the values zero or one, we can write

$$\frac{E[f(Z_k) = 1, Y_j]}{E[f(Z_k) = 1]} = E[Y_j | f(Z_k) = 1]. \quad (21)$$

It is notable that, without the term $(-W_{jk})$ corresponding to long-term depression (e.g., Burger & Levy, 1985), this development [Eq. (16)–(20)] would not go through.

The running averager form of synaptic modification just described only covers to a delta neighborhood determined by the size of ϵ .

A mathematically different class of rules also fits the physiological observations and provides a similar final value of each W_{jk} as does the running averager. This other set of rules is Bayesian in their updating method (Levy & Desmond, 1988). These rules produce changes in synaptic strength which coverage to exact averages. A seeming disadvantage of the Bayesian adaptive form of associative synaptic modification is that, when $f(Z_k) > 0$ after many trials (actually the Bayesian form is most sensible when $f(Z_k)$ can take on values zero or one), new associative events have very little affect on the synaptic weights.

An alternative modification rule, which shares some of the features of the running averager and the Bayesian form, uses Eq. (16) and a dynamically controlled $\epsilon(t)$, $0 \leq \epsilon(t) \leq 1$. The variable $\epsilon(t)$ would be controlled by the prediction evaluation generator so as to be identified with a continuous mismatch signal and would be adjusted to larger values in novel environments and to smaller values as more and more time is spent in the same environment. Thus the delta region of approximate convergence which exists in the stochastic averaging method would go to zero. That is, there would be exact convergence when the input environment is stationary in its statistical correlations.

3. Reinforced Modification

A variety of researchers distinguish between unsupervised and reinforced-type supervised modification rules (e.g., Amari, 1977; Kohonen, 1972, 1984). Most important for the development of a network which can perform prediction of future events is an associative synaptic modification rule and appropriate neural circuitry in which the synaptic strength of afferents representing early events is reinforced by converging afferent activity representing later events. The relevant characteristics of a reinforced synaptic modification rule for our model are

1. There should be two distinct classes of excitatory inputs. The two classes are presumably the entorhinal cortical and CA3 inputs to CA1.
2. Sufficient activity of one class of inputs should permit synaptic modification of the other class of inputs. Presumably entorhinal cortical activity is permissive for modification of the CA3–CA1 synapses.
3. The opposite interaction, which reverses the permissive input and the modified input, should not occur. Here the entorhinal–CA1 syn-

apses can remain unmodified even while their activity permits modification of the CA3-CA1 synapses.

An additional characteristic we find useful in our network models is:

4. The permissive class of inputs is capable of self-supervised modification though not necessarily while simultaneously permitting modification in the other class of inputs. (Such unsupervised modifiability of entorhinal-CA1 synapses is suggested by other work from our laboratory.)

The possibility that CA1 is a reinforced type of supervised system follows from the experimental demonstration of these four properties by Moore and Levy (1988; Levy, 1988). These experiments studied associative synaptic modification of CA1 in the hippocampal slice. This study was able to find stimulation parameters such that the CA3 input required an associated distal dendritic excitation for CA3-CA1 potentiation to occur. The reverse interaction, in which potentiation of the distal synapses required CA3 activation, did not obtain, however. These and other observations (King & Levy, 1986) suggest that modification of the CA3-CA1 synapses, as controlled by the entorhinal input to CA1, bears a fundamental resemblance to the reinforced type of supervised synaptic modification.

C. TIME CONSTRAINTS ON ASSOCIATIVE SYNAPTIC INTERACTIONS

The temporal dependencies governing associative synaptic modification are crucial to any theory of the hippocampus as a generator of predictions. Experiments using the bilateral projections from the entorhinal cortex to the dentate gyrus were the first to observe a rather special temporal relationship which defines association at the cellular level (Levy & Steward, 1983; Lopez, Burger, & Levy, 1985; Lopez, Burger, Dickstein, Desmond, & Levy, submitted; see also Levy & Desmond, 1985b, for review). These experiments show that the pre- and postsynaptic activities, Y_j and $f(Z_k)$ of Eq. (16), need not be synchronous so long as their temporal association falls within the constraint of a specific temporal window and has a special temporal ordering. The temporal ordering constraint allows associative potentiation when the presynaptic activity Y_j precedes the postsynaptic excitation $f(Z_k)$. In fact, the two activities need not overlap at all.

Associative long-term potentiation can be induced with as much as 30 msec between the first pulse of a presynaptic train and the beginning of a powerful postsynaptic depolarization. The reverse ordering is not a potentiating condition at all. Rather, long-term depression occurs when the

permissive, reinforcing input precedes the associated presynaptic activity. Functionally, then, the rule looks at the preceding $Y_j = 0$ associative with the $f(Z_k) > 0$ rather than the activation of j which follows activation of k .

Similar experiments have now been performed in area CA1. Here the entorhinal input to CA1 is the permissive input for modification of the CA3-CA1 synapses. As in the dentate gyrus, the permissive excitation can follow, but not precede, synaptic activity to produce long-term potentiation at the active CA3-CA1 synapses (Moore & Levy, 1986; Levy, 1988). Figure 9 shows the temporal nature of the pulse trains used in this experiment to induce associative long-term potentiation. The reverse ordering does not produce potentiation.

The ordered timing requirements of an associative synaptic modification rule are useful for constructing a network that learns sequences. For example, if one particular event represented by the CA3 inputs regularly

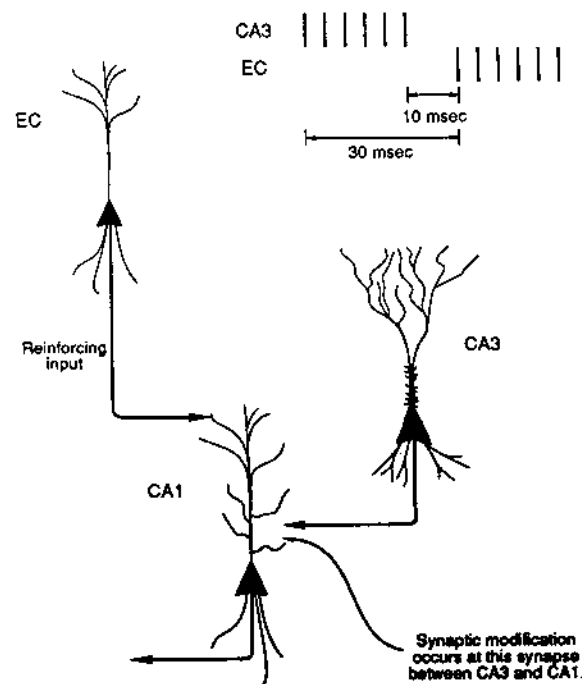


Fig. 9. Associative potentiation can span time. Associative long-term potentiation of CA3-CA1 synapses is induced by the input of entorhinal cortical (EC) neurons. Note the temporal relationship between CA3 and EC activity which induces potentiation; the reverse ordering does not produce potentiation. Copyright © 1989 by William B Levy.

precedes a particular event represented by the entorhinal layer III cells, then the CA3 event can be used to predict in advance, by about 20 msec, the entorhinal event or the CA1 cell firing induced by the entorhinal event. The observed timing requirement thus allows the CA3 input to predict CA1 activity induced monosynaptically by the entorhinal input.

Of course, the rather limited temporal window of 30 msec must be extended to be of behavioral consequence. That is, the temporal window of such associations is too brief to make timely predictions if we consider the whole animal trying to predict its environment in the real world. In this case, a useful predictive representation must be produced many tens, hundreds, or thousands of milliseconds before the event being predicted actually occurs so that the animal can act on its predictions in time. In fact, we interpret the results of these synaptic modification studies as truly limiting, and, therefore, we must consider other mechanisms which might be used to extend the temporal context of the predictions.

D. TIME-SHIFTING FOR ASSOCIATION AND TIMELY PREDICTIONS

1. Delay Lines Shift Representations Later in Time

Although an event which occurs in the real world cannot be shifted in time, it is possible to shift representations of this event in time. The more obvious shift is to move a representation later in time. Shifting a representation later in time is a workable scheme used to produce an association of two events at the cellular level even though these two events are widely separated in time in the real world (e.g., Zipser, 1986). In fact there is a history of research in which sequences longer than two are treated by adding delay lines and feedback (Fukushima, 1973; Grossberg & Kuperstein, 1986; Hopfield, 1987; Kleinfeld, 1986; Kohonen, 1984) to solve pattern recognition problems. However, pattern recognition problems are retrodictions while our interest here is in predictions.

In order to shift representations in time, action potential conduction down an axon could act as a delay line on the scale of 2–4 msec; more important, however, is the capacitance of each neuron. The so-called synaptic delay, usually defined as the time it takes a synaptic event on a dendrite to charge the neuron's cell body, is of the order of 10–12 msec. So, for example, by chaining together a sequence of 10 neurons, a shift in time of 100–120 msec might be achieved. This is still a rather short interval to be of use for many of the sequences encountered in the real environment, such as the sequence of representations generated as a rat moves through a maze.

The top portion of Fig. 10 illustrates a delay line circuit called a tapped delay line because of the multiple delay times available. Such tapped de-

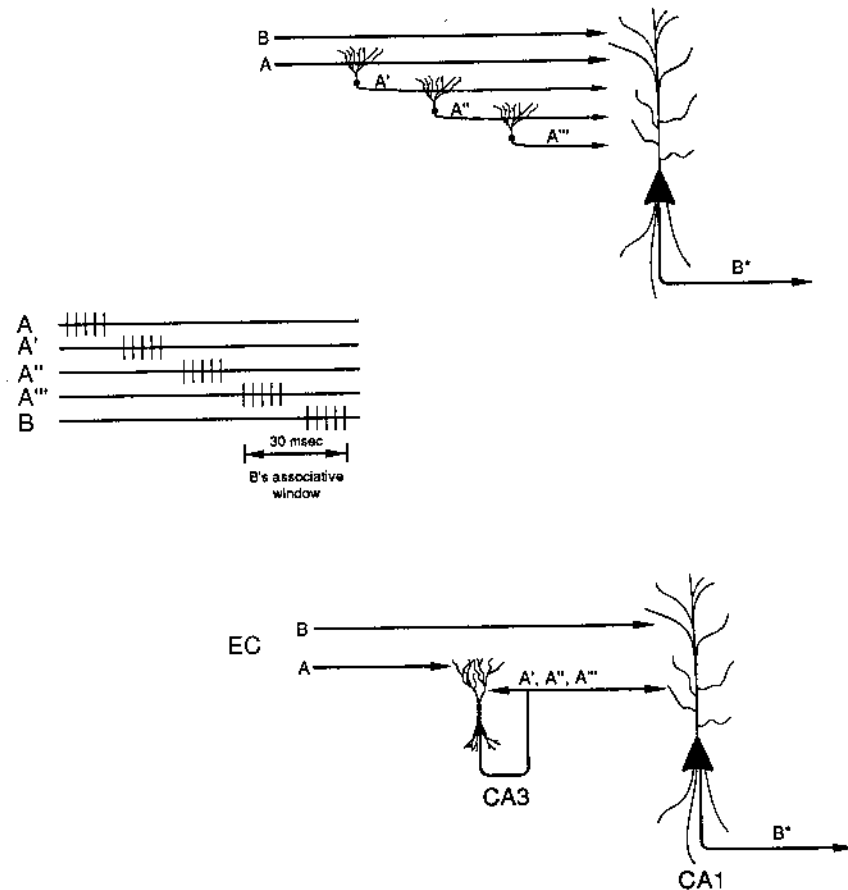


Fig. 10. Temporal compression of event A into the associative time window of event B. Associative synaptic modification spans a short time window. Tapped delay lines (upper panel) allow a multiplicity of delayed representations. In a multidimensional representation, however, such delay lines produce an undesirable delay of predictive representations. In contrast, a positive feedback loop (lower panel) can span longer time windows than the delay lines without delaying prediction generation (see also Fig. 11). Copyright © 1989 by William B Levy.

lay lines are, however, unsuitable for producing a predictive representation in a timely fashion because this same delay circuitry, which brings pre- and postsynaptic events together in time to allow associative synaptic modification, must also delay generation of the predictive representation itself. This undesirable delay is inevitable because the conditioning

input A which generates the prediction (B^*) is delayed by this same circuitry. As a result, by our definition of prediction (see Section III), an overly delayed prediction is not a prediction at all. Thus, for the prediction problem, as opposed to the retrodiction problem of pattern recognition, this imposed delay must be removed during prediction since the prediction must be delivered in a timely fashion to be a predictive representation.

To be more specific, consider some realistic estimates. A single delay of 10 msec, which just shifts an input into a 20-msec associative window, will create a prediction with 10 msec to spare using the modified synapses. However, if more than 20 msec of delay are interposed to shift a conditioning representation into the associative window, then the prediction later generated by the conditioning representation through this same synapse will not precede the event being predicted (see also Fig. 11 and below).

Thus, to implement the pure delay line strategy not only requires a multiplicity of delays but also requires some way of shifting the delayed response earlier in time or abandoning the originally modified synapse. This problem is solvable for unidimensional signals by using a cascaded sequence of associative modification. For multidimensional signals, however, we have been unable to find an implementation within the spirit of the delay line that does not require too many neurons or too many synapses or that does not violate the local computational principle. The simple picture at the top of Fig. 10 may conjure up the idea that a feedforward input could by-pass the delay circuitry after learning has occurred, but the figure omits the fact that the signals are multidimensional and that successive stages involve thousands of neurons working in parallel while intermixing signals, rather than just the single neurons illustrated.

2. Feedback Networks as Multiple Delay Lines

Networks which can obtain approximately stable states, however, such as the feedback networks of Hopfield (1984), Cohen and Grossberg (1983), and Shaw, Silverman, and Pearson (1985), are suitable for time-shifting delay and for what is effectively reverse time-shifting. That is, if a sequence of representations is highly similar, then associative synaptic modification can occur using the later members of such a sequence, but the modified synapses will be accessed by earlier members of the sequence when the sequence reoccurs as well. Thus feedback networks, including both the short CA3-CA3 loop and the longer limbic loops through subicular cortex (see Fig. 1) (see, e.g., Deadwyler, West, Cotman, & Lynch, 1975), may be able to function as a variable time-shifting

device which, in the sense of the similarity approximation achieved, dissolves the arrow of time.

3. Feedback Networks Can Time Shift without the Problems of Delay Lines

The proposed time-shifting mechanism is an approximation scheme which uses preprocessors to make successive signals similar enough that any one signal can substitute for the other to perform prediction but yet different enough that there is no information loss, $H(X | Y)$, from this transformation. This idea is illustrated at the bottom of Fig. 10 and in Fig. 11. Figure 10 shows how feedback would shift a representation into the associative window of synaptic modification. Early representations are

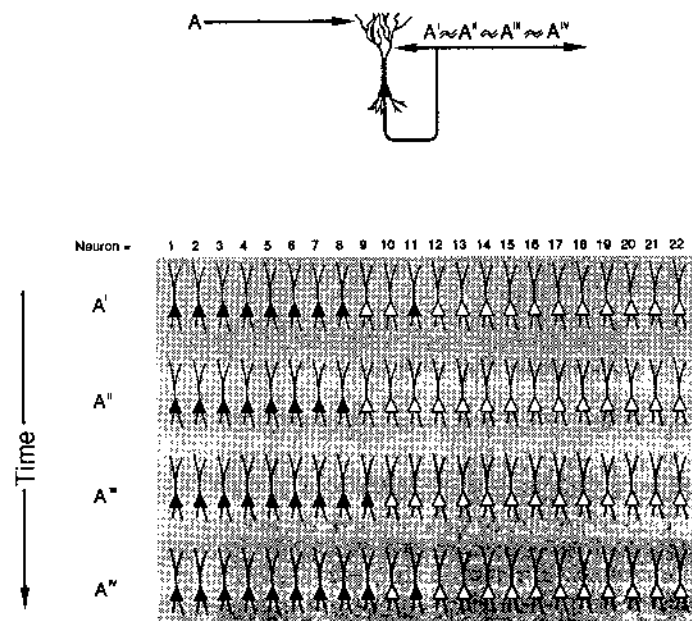


Fig. 11. Time shifting via approximate sequence stability of CA3 firing patterns. A feedback network such as CA3 can help span longer time windows while still maintaining a sequence of distinct representations. The lower portion of the figure indicates a sequence (A^I - A^{IV}) of output activity patterns of 22 CA3 neurons. Solid neurons are on; open neurons are off. Because the activity patterns A^I and A^{IV} in CA3 are nearly the same, A^I can access almost the same CA1 synapses as A^{IV} including those synapses which have associatively modified. In this way, feedback loops can span longer time windows without delaying the generation of a predictive representation. Copyright © 1989 by William B Levy.

almost identical to late representations (see bottom of Fig. 11), so even when synaptic modification in CA1 occurs to late (delayed) representations, the early representations will use the same synapses to evoke timely predictions.

At this point the combinatorial explosion begins to work in the network's favor because, in the high-dimension space (i.e., lots of neurons) used for neuronal representations, only a very small difference is needed to distinguish successive signals while the approximately identical nature of these signals is maintained. That is, perfect signal reproduction is not needed in a high-dimension space because small differences are a trivial fraction of the total representation signal. The feedback loop which produces approximately stable representations thus allows the network to span a variable time range, a long interval for association and little or no interval for prediction.

Some calculations illustrate how well the feedback network can work. Suppose we have 10^5 CA3 neurons code a representation with 10% active and 90% inactive neurons. Of these 10,000 active neurons, let 1% vary randomly while the network keeps the other 99% at the same high probability of firing. Each successive pattern will be almost like the previous pattern even though the network has some 2^{100} distinct states available if the 1% of the neurons undergoing random firing have a 50% chance of being active. Thus the network can uniquely represent any sequence of differing patterns, which allows it to avoid information loss [$H(X|Y)$], while the first representation and the last representation remain almost identical, within 1%. In this way, associative synaptic modification using the delayed form of the representation is good enough because both the early, and what we might call the reverse time-shifted, representations use almost the exact same set of synapses.

Figure 12 depicts a low-dimension example of eight different representations which accomplish temporal compression because they are nearly identical. Larger groups of neurons will do exponentially better in terms of the possible number of their distinct representations and their similarity.

4. Extending the Forecasted Period

In sum, four mechanisms of time shifting have been presented (see Fig. 14): (1) a delay line which uses the RC time constant of neurons to slow propagation of signals through a network; (2) a high-information content, positive feedback using a short loop; (3) a high-information content, positive feedback that uses long loops; and (4) the associative synaptic modification rule, which can associate two different inputs across a gap in time. In the next section we suggest that the feedback circuitry of the hippo-

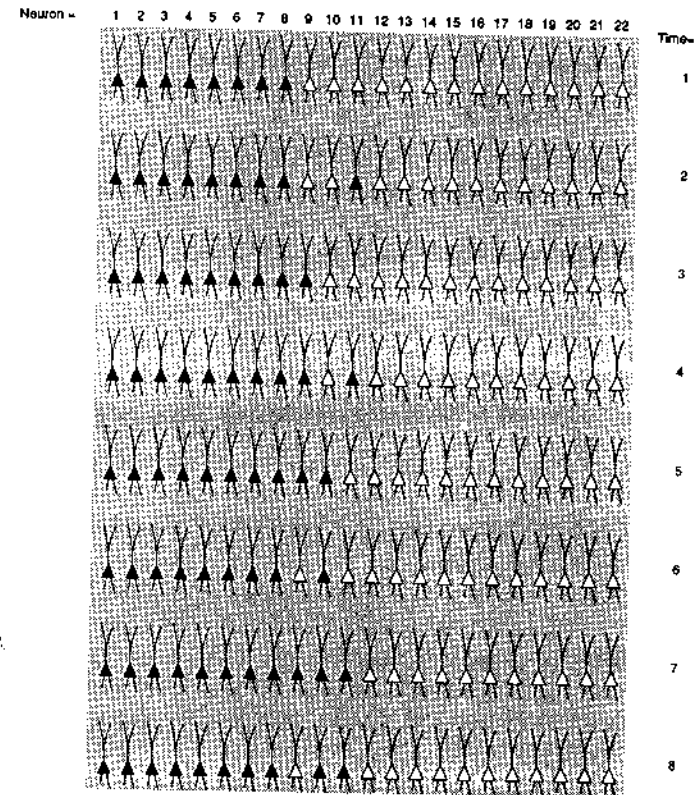


Fig. 12. One tier of neurons, changing over time, could achieve a highly similar but different series of representations. Solid neurons are on; open neurons are off. Copyright © 1989 by William B Levy.

campus (and associated limbic structures as well) can be used both to create predictions which look farther ahead into the future and to solve the problem of decreasing the statistical dependency of representations.

VI. An Algorithmic Process as a Hypothesis for the Function of DG/CA3

This section presents the second half of the argument that the function of the DG/CA3 system is to produce a sequence of similar representations. We have just seen that such a sequence is useful for solving the time-shifting problem. Here we sketch an argument which hypothesizes that such a class of sequences can also solve the problem of lowering the

intrinsic statistical dependency [Eq. (5)] of the conditioning variable $[Y$ of Eq. (10)].

A. THE TIGHT PACKING ALGORITHM

For the sequence $X(t)$, we are seeking a way to implement a transformation, $f: X(t) \rightarrow Y(t+1)$, that is provably optimal for reducing statistical dependence. In other words, we want to find an f which creates the smallest $\Sigma H(Y_j)$ while preserving information, $H(X|Y)$.

Below we outline an algorithm which achieves the desired transformation. Figure 13 illustrates how, after removing enough redundancies by some as yet unspecified process, it is sometimes possible to represent the same information with fewer active neurons. We call this transformation "tight packing" because it allows a reduction in the required size of the representation space.

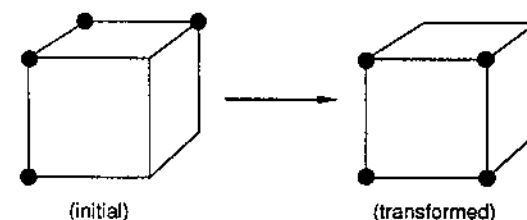
In fact, there exists a deterministic algorithm which implements tight packing. Just below we prove its optimality in a rather restricted setting. Then we describe a relaxed version of the algorithm which a neural network might be able to implement. We call this approximation method "rough counting." It is easy to imagine that the same DG/CA3 circuitry which performs time-shifting would simultaneously accomplish rough counting. Finally we point out that the algorithm is unsatisfactory in a nonstationary setting and suggest a possible solution to this failing.

The algorithm works for the space, $X \in \{0,1\}^n$, when the space is sparsely sampled, that is, when the sample size $N \ll 2^n$.

It can be claimed that the mapping which takes an arbitrary sequence $X(t)$ into the sequence $Y(t+1)$ as illustrated in Table I produces a Y of minimal statistical dependence. It should be clear that the illustrated sequence Y in the third partition of Table I could just be the result of counting a binary starting at zero. This mapping is a tight packing because it minimizes the Hamming distance among the points enumerated. In fact counting up from zero is not the only way to achieve minimal statistical dependence. Any mapping, which may start at any point in Y , minimizes statistical dependence if the mapping minimizes the Hamming distance among the points in Y space.

The proof that statistical dependence is minimized is trivial when the sample size $N = 2^m$, m is an integer, and when probability is taken to be the sampled relative frequency. Suppose $Y = (Y_1, \dots, Y_j, \dots, Y_n)$. All those dimensions j which remain unchanged, as in the first and second partitions of Table I, have associated marginal probabilities $\{P(Y_j = 1), P(Y_j = 0)\} = \{0,1\}$. All other dimensions (those in the third partition) have $P(Y_j = 1) = P(Y_j = 0) = .5$. The probability of any sampled Y is 2^{-m} , and this is the same probability as computed from the marginal values as

A Representations in a geometric space



B The same 4 representations in a neural version of the geometric space

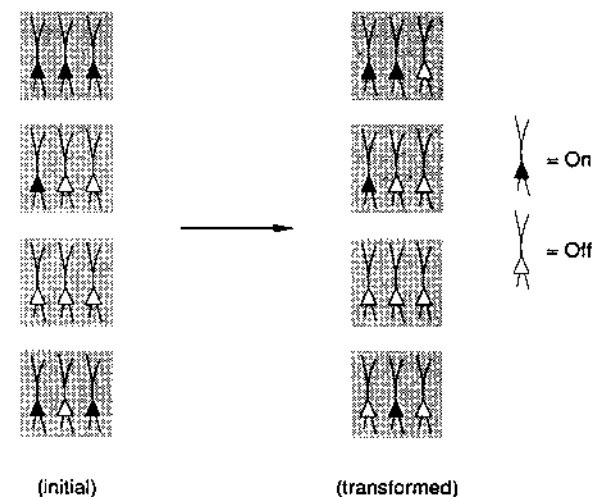


Fig. 13. Tight packing can lower the complexity of a representation. A, Four representations (\bullet) in a three-dimensional space (a cube) are transformed into four representations in a two-dimensional space (a face of the cube). B, If a neuron corresponds to a dimension of the cube, three neurons are needed for the four representations scattered around the eight corners of the cube initially. However, only two neurons are needed for the representation after the transformation. Copyright © 1989 by William B Levy.

the product $(1^{n-m})(.5^{-m})$. For the configurations Y of relative frequency zero, the marginal-based calculations give probability zero. Thus we have shown that the mapping produces complete independence under these circumstances because the full distribution $P(Y)$ is reproduced by multiplication of marginal probabilities, $[P(Y_j)]$.

It is too much to expect a neural network to "count" in perfect order but this condition can be relaxed.

TABLE I

PARTITIONING OF PROCESSORS AFTER 2^m
SAMPLES

N	Partition		
	First	Second	Third
1	{000...0	111...1	000...00,
2	000...0	111...1	000...01,
3	000...0	111...1	000...10,
4	000...0	111...1	000...11,
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
2^m	000...0	111...1	111...11}

B. ROUGH COUNTING

1. Sampling in Bunches

Consider the case in which there is a single generating distribution, $P\{X(t)\}$. Samples are sequences of X drawn in a group, say 2^6 at a time. If a single sampling period covers 2.5 msec, then 2^6 samples represent a sampling period of only 160 msec, so sampling in bunches is quite conceivable for the brain or regions within the brain.

2. The Algorithm Is a Dynamic Partitioning

Suppose there are n processors (about 10^5 – 10^6 neurons in a subregion of the limbic system that inspires this theory) in the X space, such as layer II cells of the entorhinal cortex, and the same number in the Y space, such as CA3 spiny pyramids. We will now see that there is a dynamic, that is, time-dependent, partitioning of the n processors in the Y space. This dynamic partitioning requires three distinct sets of processors $\{1,2,3\}$ with $n_1(t)$, $n_2(t)$, and $n_3(t)$ processors in each set in such a way that $2^{n_1} > 2^{n_2} \gg 2^{n_3}$ holds. For instance, n_2 might be 1–10% of n_1 ; n_3 will never be larger than 40 neurons, and 16–20 seems enough for most situations for two reasons. First, the third partition will by itself distinguish the sampled Y s so as to stand in one-to-one correspondence with the X samples. Second, 2^{20} processors allows a unique representation in Y of more than one million samples from X .

The nature of the partitioning of the neurons of CA3 can be described as follows:

1. The n_1 processors in the first partition of Y have never fired.
2. The n_2 processors in the second partition have always been firing.
3. The n_3 processors in the third partition have been on half the time in a very special way.

Specifically then—and here we again assume for simplicity that N is a power of two, say 2^m — $n_3(N) = m$ and no two vectors in Y are the same.

Table I shows such a partitioning of processors after N samples.

When the set of states in Table I is ordered as indicated, the third partition embodies a counting algorithm. As pointed out above, such a counting-like ordering is induced by the way we have ordered the processors Y_j and their partitions for illustrative purposes. Such an ordering does not actually occur in a network so that any numerical label on a processor is just an arbitrary mathematical convenience. [It may also help some readers to note that the transformations under which the measure of statistical dependence is invariant include (1) complementation of any Y_j because $P(Y_j)$ is a binary distribution and complementation does not alter $H(Y_j)$ and (2) permutation of the ordering of the Y_j values because $\sum_i H(Y_j)$ is unaffected by permutations of the variable ordering 1, 2, 3, . . . , n .]

Not only is optimization invariant over permutations of neurallike processors Y_j , but it is adequate to count in a rather slipshod manner. Since samples arrive in groups before a prediction is required, the network does not need to do a proper, straight-through job of counting. Any ordering within each group, including a randomly created one, is good enough so long as each group more or less methodically occupies the counting subspace: the third partition.

The dynamic aspect of the partitioning involves shifting a processor from either the first or second partition into the third partition. Tables IIA and IIB show the partitioning after two and four samples respectively. Note that we have suppressed group sampling to make the tables simpler. Tables IIA and IIB precede Table I in that the Table II partitionings later grow into the partitioning shown in Table I. Thus, in this dynamic partitioning, a processor of the first partition has been off (i.e., a zero) for the last N samples and then shifts into partition three by turning on (i.e., a one) for the next N samples. A similar scheme describes the shifting of a processor from partition two to partition three. The size of the third partition thus grows at a logarithmic rate with sample size so that there is more than enough space for any number of samples that may be realistically encountered.

TABLE II
THE GROWTH OF THE DYNAMIC
PARTITIONING

N	Partition		
	First	Second	Third
A. After two samples			
1	{000...0	111...1	0,
2	000...0	111...1	1}
B. After four samples			
1	{00...0	111...1	00,
2	00...0	111...1	01,
3	00...0	111...1	10,
4	00...0	111...1	11}

3. Neural Hypotheses

The exact mechanism which would accomplish this algorithm has not been tested. However, the mechanism seems to be compatible with the computations of neurallike processors, particularly feedback networks. The extreme constancy of successive patterns (note that they differ from each other by much less than 1% for the number of processors envisioned) could be produced by a Shaw *et al.* (1985) network or by a Hopfield (1984) or Cohen-Grossberg (1983) feedback network that has nearly converged. The slight movement away from an absolutely identical sequence of states $Y(t + 1)$ would be guaranteed by the nonconstancy of the input patterns $X(t)$.

C. NONSTATIONARY ENVIRONMENTS

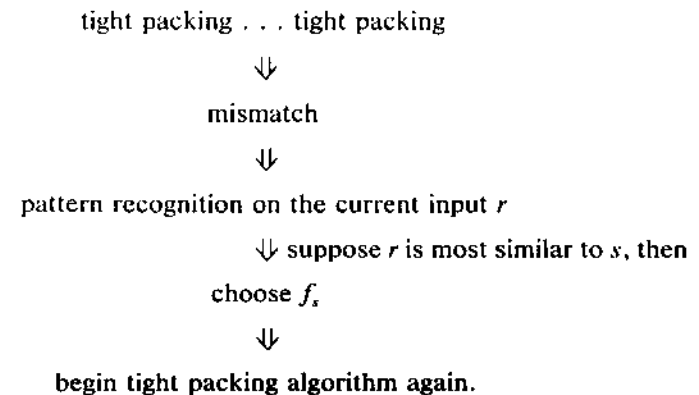
A complex environment might contain a sequence of generating functions. In this case even the rough counting algorithm leaves something to be desired. For example, when the environment shifts from one generating process to another and then shifts back again to a previously experienced environment, the preprocessor should be able to shift the dynamic partitioning appropriately. In particular, the network should be able to use the previously acquired conditional expectations in CA1 for predictions in the previously experienced environment. However, it will only be possible to use these acquired expectations and to update them adaptively if the network can return, at least approximately, to the preprocessing state, that is, dynamic partitioning, used for the recurring environ-

ment. Unfortunately, naive implementation of the rough counting algorithm can destroy the usefulness of previously acquired conditional expectations because the algorithm can be implemented so as to invalidate the previously acquired expectations.

To put this a little more precisely, we are now considering the problems which result from nonstationary environments. These environments can be thought of as a sequence of generating distributions indexed by r . In such circumstances, rather than seeking overall independence, the network should seek the independence of the representation Y conditional on r . Thus, rather than seek a good transformation f , we are sequentially seeking good transformations f_r .

The added difficulty of prediction in the nonstationary complex environment is how and when to change the transformation $f_r: X \rightarrow Y$ to another f_s . It is particularly critical to discover a method of changing f , which allows the network to return to an $f_r = f_s$ when s reoccurs. What is required then is an adaptive version of this dynamic partitioning so as not to destroy previously acquired relationships. The proposed solution to this problem is to interpolate a pattern recognition process between environments r .

One scheme we are considering is described by the following sequence of operations:



The *tight packing* portion of the algorithm is the rough counting described above. Rough counting continues until a mismatch signal occurs. *Mismatch* means the detection of a series of sufficiently poor predictions such that it is worthwhile to assume r has changed. The prediction evaluation box of Fig. 7 is the mismatch detector. The evaluation process is an adaptive procedure which produces a running average measure of the

quality of the predictions relative to the corresponding actual outcomes. Mismatch detection terminates the tight packing mode and switches the network into a pattern recognition mode. The network performs as a *pattern recognition* device on the current environment. This pattern recognition process allows the DG/CA3 system to take advantage of previous adaptive modifications and to shift its preprocessing transformation along with the shifting statistics of the entorhinal cortical inputs. This pattern recognition process also allows the network to return to old transformations when the environment shifts back to previously encountered statistics. If a previously encountered r is sensed, the network configuration moves, that is, shifts its activity state, to a configuration that is similar to the previous representations of r . If a new r is encountered, any configuration that is not associated with any r already observed will do. It is very easy to find a new place to begin tight packing in such a large space ($\sim 2^{100,000}$) because no more than a random selection upon the neuronal activities is necessary to implement the beginning of the three-way partitioning which is, with statistical certainty, a novel configuration. A novel environment would then require synaptic modification to set up relatively stable conditions which are the first and second partitions of the dynamic partitioning.

After the network activity pattern moves to the appropriate configuration of activity, that is, to the first and second partitions appropriate to the environment r , tight packing begins again from about where it left off for this environment, with increases in the number of processors in the third partition as necessary.

In sum, then, we propose that the same mechanisms which are used for time-shifting representations can also be used for solving the sensory fusion problem. The similarities between Fig. 14A and 14B make the same point by juxtaposition. In fact, it is only natural that a system which evolved for helping mammals move to and from the nest as efficiently as possible would have both the capability of sensory signal fusion and of time-shifting representations.

Finally, we can posit a relationship between the psychological theories of hippocampal function and the computational theory presented here. The time-shifting feedback loops which produce sequences of similar representations can be put to other uses. Each of these feedback systems can function as one of the short-term memory systems in the brain. In addition, these short-term memory systems are particularly suited for the signal mixing task of lowering statistical dependence and of connecting the past with the future—functions that could well describe the fundamental computations of a working memory system.

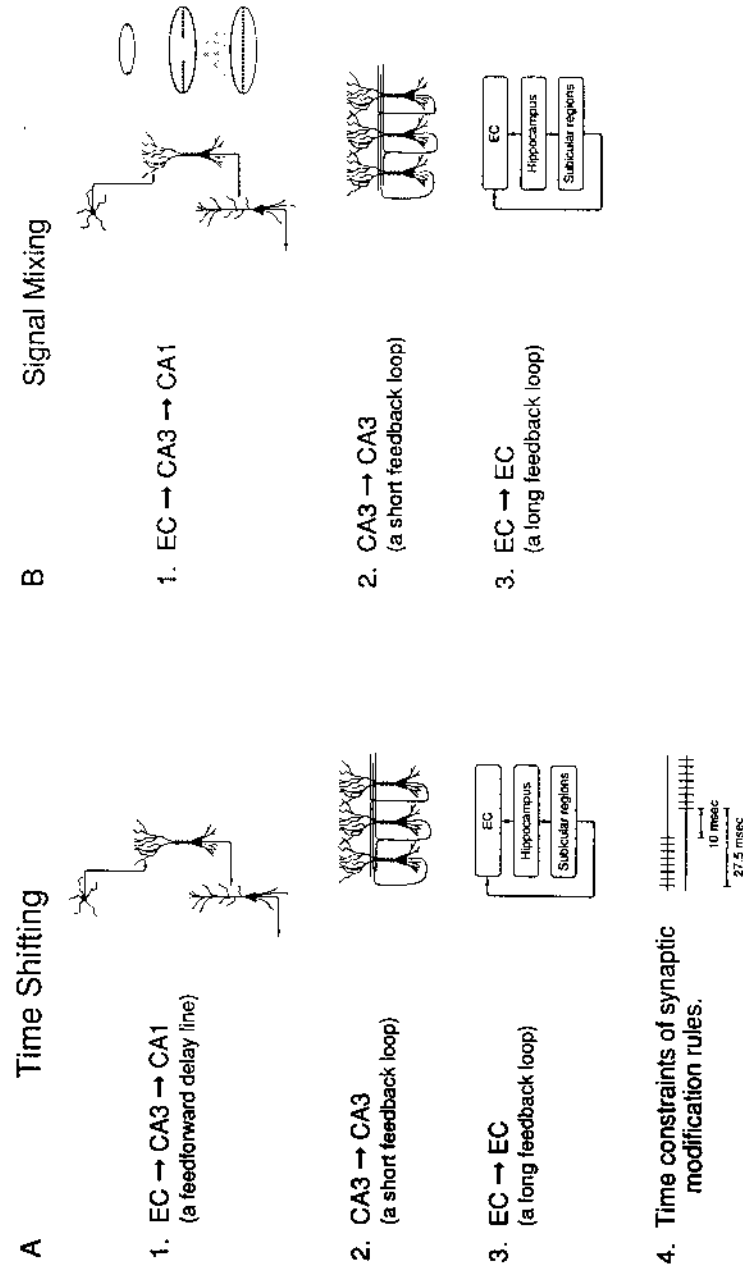


Fig. 14. Similarities between processes that time shift and signal mix. A, Four types of time shifting in the model. Synaptic modification (4) and delay lines (1) can provide a small amount of time shifting and temporal compression of representations. However, feedback networks (2 and 3) can span much longer periods, thereby allowing associative modifications over longer times while still producing a prediction before the event being predicted. B, The same circuitry used for time shifting can also be used for signal mixing. Note that divergence and convergence as in (1) can take place at many successive levels of the system. Copyright © 1989 by William B Levy.

ACKNOWLEDGMENTS

WBL is supported by NIMH RSDA MH00622 and by the Department of Neurological Surgery. John A. Jane, Chairman, provided the environment which allowed me to work on ideas somewhat outside of mainstream approaches to neuroscience. Interactions with F. H. C. Crick stimulated the development of some of the philosophical ideas about prediction. I gratefully acknowledge the suggestions and criticisms supplied by C. M. Colbert, S. Shoemaker, and particularly by N. L. Desmond, whose contributions to this paper are pervasive. I am also grateful for help with the more mathematical sections given by my collaborators, D. L. Costa and D. St. P. Richards. J. Sullivan helped develop the figures.

REFERENCES

- Amaral, D. G. (1987). Memory: Anatomical organization of candidate brain regions. In F. Plum (Ed.), *Handbook of physiology: Sect. 1. The nervous system* (Vol. V, pp. 211-294). New York: Oxford University Press.
- Amari, S.-I. (1977). Neural theory of association and concept-formation. *Biological Cybernetics*, *26*, 175-185.
- Ashby, W. R. (1956). Design for an intelligence-amplifier. In C. E. Shannon & J. McCarthy (Eds.), *Automata studies* (pp. 215-234). Princeton, NJ: Princeton University Press.
- Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *Mechanisation of thought processes* (Vol. II, pp. 537-559). London: Her Majesty's Stationery Office.
- Barlow, H. B. (1961a). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217-235). Cambridge, MA: MIT Press.
- Barlow, H. B. (1961b). The coding of sensory messages. In W. H. Thorpe & O. L. Zangwill (Eds.), *Current problems in animal behaviour* (pp. 331-360). London: Cambridge University Press.
- Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuron-like elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, Cybernetics*, *SMC-13*, 835-846.
- Bellman, R. (1961). *Adaptive control processes: A guided tour*. Princeton, NJ: Princeton University Press.
- Breese, C. R., Hampson, R. E., & Deadwyler, S. A. (1989). Hippocampal place cells: Stereotypy and plasticity. *Journal of Neuroscience*, *9*, 1097-1011.
- Brooks, V. B. (1986). How does the limbic system assist motor learning? A limbic comparator hypothesis. *Brain Behavior and Evolution*, *29*, 29-53.
- Burger, B., & Levy, W. B. (1985). Long-term associative potentiation/depression as an analogue of classical conditioning. *Society for Neuroscience Abstracts*, *11*, 493.
- Burger, B., & Levy, W. B. (1987). An intensity-dependent threshold-like effect controls both LTP and LTD. *Society for Neuroscience Abstracts*, *13*, 974.
- Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*, 54-115.
- Cohen, M. A., & Grossberg, S. (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*, 815-826.
- Colbert, C. M., & Levy, W. B. (1988). What is the code? *Proceedings of the International Neural Network Society*, *1*, 246.
- Cox, R. T. (1961). *The algebra of probable inference*. Baltimore, MD: Johns Hopkins Press.
- Cox, R. T. (1978). Of inference and inquiry, An essay in inductive logic. In R. D. Levine & M. Tribus (Eds.), *The maximum entropy formalism*. (pp. 119-167). Cambridge, MA: MIT Press.
- Csiszár, I., & Körner, J. (1981). *Information theory: Coding theorems for discrete memoryless systems*. New York: Academic Press.
- Dawkins, R. (1976). *The selfish gene*. New York: Oxford University Press.
- Deadwyler, S. A., West, J. A., Cotman, C. W., & Lynch, G. S. (1975). Physiological studies of the reciprocal connections between the hippocampus and the entorhinal cortex. *Experimental Neurology*, *49*, 35-57.
- Desmond, N. L., & Levy, W. B. (1988). Anatomy of associative long-term synaptic modification. In P. W. Landfield & S. A. Deadwyler (Eds.), *Long-term potentiation: From biophysics to behavior* (pp. 265-305). New York: Alan R. Liss.
- Doller, H. J., & Weight, F. F. (1985). Perforant pathway-evoked long-term potentiation of CA1 neurons in the hippocampal slice preparation. *Brain Research*, *333*, 305-310.
- Eichenbaum, H., & Cohen, N. J. (1988). Representation in the hippocampus: What do hippocampal neurons code? *Trends in Neuroscience*, *11*, 244-248.
- Foster, T. C., Christian, E. P., Hampson, R. E., Campbell, K. A., & Deadwyler, S. A. (1987). Sequential dependencies regulate sensory evoked responses of single units in the rat hippocampus. *Brain Research*, *408*, 86-96.
- Fukushima, K. (1973). A model of associative memory in the brain. *Kybernetik*, *12*, 58-73.
- Gamow, G. (1961). *One two three . . . infinity*. New York: Viking Press.
- Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability. A guide to the theory of NP-completeness*. New York: Freeman.
- Geman, S. (1981). The law of large numbers in neural modelling. *SIAM AMS Proceedings*, *13*, 91-105.
- Golden, R. M. (1988). Probabilistic characterization of neural model computations. In D. Z. Anderson (Ed.), *Neural Information Processing Systems* (pp. 310-316). New York: American Institute of Physics.
- Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In F. Plum (Ed.), *Handbook of physiology: Sect. 1. The nervous system* (Vol. V, pp. 373-417). New York: Oxford University Press.
- Good, I. J. (1963). Maximum entropy for hypotheses formulation especially for multidimensional contingency tables. *Annals of Mathematical Statistics*, *34*, 911-934.
- Gray, J. A. (1982). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system*. New York: Oxford University Press.
- Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121-134.
- Grossberg, S., & Kuperstein, M. (1986). *Neural dynamics of adaptive sensory-motor control: Ballistic eye movements*. Amsterdam: Elsevier/North-Holland.
- Hamming, R. W. (1980). *Coding and information theory*. New York: Prentice-Hall.
- Hinton, G. E., & Sejnowski, T. J. (1983). Optimal perceptual inference. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pp. 448-453.
- Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences U.S.A.*, *81*, 3088-3092.
- Hopfield, J. J. (1987). Learning algorithms and probability distributions in feed-forward and feed-back networks. *Proceedings of the National Academy of Sciences U.S.A.*, *84*, 8429-8433.

- Hopfield, J. J., & Tank, D. W. (1985). "Neural" computation of decisions in optimization problems. *Biological Cybernetics*, *52*, 141-152.
- Insausti, R., Amaral, D. G., & Cowan, W. M. (1987). The monkey entorhinal cortex. II. Cortical afferents. *Journal of Comparative Neurology*, *264*, 356-395.
- Jaynes, E. T. (1978). Where do we stand on maximum entropy? In R. D. Levine & M. Tribus (Eds.), *The maximum entropy formalism*. (pp. 15-118). Cambridge, MA: MIT Press.
- Jerison, H. J. (1973). *Evolution of the brain and intelligence*. New York: Academic Press.
- Johnson, R. W., & Shore, J. E., (1983). Comments on and correction to "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy." *IEEE Transactions on Information Theory*, *IT-29*, 942-943.
- Jones, E. G., & Powell, T. P. S. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain*, *93*, 793-820.
- Karp, R. M. (1975). On the complexity of combinatorial problems. *Networks*, *5*, 45-68.
- King, M. A., & Levy, W. B. (1986). Heterosynaptic depression of hippocampal CA3 afferents to CA1 accompanies long-term potentiation of convergent entorhinal afferents. *Society for Neuroscience Abstracts*, *12*, 505.
- Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*, 671-680.
- Kleinfeld, D. (1986). Sequential state generation by model neural networks. *Proceedings of the National Academy of Sciences U.S.A.*, *83*, 9469-9473.
- Kohonen, T. (1972). Correlation matrix memories. *IEEE Transactions on Computers*, *C-21*, 353-359.
- Kohonen, T. (1984). *Self-organization and associative memory*. Berlin: Springer-Verlag.
- Kosel, K. C., Van Hoesen, G. W., & Rosene, D. L. (1982). Non-hippocampal cortical projections from the entorhinal cortex in the rat and rhesus monkey. *Brain Research*, *244*, 201-213.
- Levy, W. B. (1985). An information/computation theory of hippocampal function. *Society for Neuroscience Abstracts*, *11*, 493.
- Levy, W. B. (1988). A theory of the hippocampus based on reinforced synaptic modification in CA1. *Society for Neuroscience Abstracts*, *14*, 168.
- Levy, W. B., Brassel, S. E., & Moore, S. D. (1983). Partial quantification of the associative synaptic learning rule of the dentate gyrus. *Neuroscience*, *8*, 799-808.
- Levy, W. B., & Burger, B. (1987a). An intensity-dependent threshold-like effect controls both LTP and LTD. *Society for Neuroscience Abstracts*, *13*, 974.
- Levy, W. B., & Burger, B. (1987b). Electrophysiological observations which help describe an associative synaptic modification rule. *Proceedings IEEE First Annual International Conference on Neural Networks*, *IV*, 11-15.
- Levy, W. B., Colbert, C. M., & Desmond, N. L. (in press). Elemental adaptive processes of neurons and synapses: A statistical/computational perspective. In M. A. Gluck & D. E. Rumelhart (Eds.), *Neuroscience and connectionist models*. Hillsdale, NJ: Erlbaum.
- Levy, W. B., & Desmond, N. L. (1985a). The rules of elemental synaptic plasticity. In W. B. Levy, J. Anderson, & S. Lehmkühle (Eds.), *Synaptic modification, neuron selectivity and nervous system organization* (pp. 105-121). Hillsdale, NJ: Erlbaum.
- Levy, W. B., & Desmond, N. L. (1985b). Associative potentiation/depression in the hippocampal dentate gyrus. In G. Buzsáki & C. H. Vanderwolf (Eds.), *Electrical activity of the archicortex* (pp. 359-373). Budapest: Akadémiai Kiadó.
- Levy, W. B., & Desmond, N. L. (1988). Characteristics of associative potentiation/depression. In H. L. Haas & G. Buzsáki (Eds.), *Synaptic plasticity in the hippocampus* (pp. 93-95). Berlin: Springer-Verlag.

- Levy, W. B., & Geman, S. (1982). *Limit behavior of experimentally derived synaptic modification rules* (Reports in Pattern Analysis No. 121). Providence, RI: Brown University, Division of Applied Mathematics.
- Levy, W. B., & Steward, O. (1979). Synapses as associative memory elements in the hippocampal formation. *Brain Research*, *175*, 65-78.
- Levy, W. B., & Steward, O. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, *8*, 791-797.
- Lopez, H., Burger, B., Dickstein, R., Desmond, N. L., & Levy, W. B. (1989). *Long-term potentiation and long-term depression in the hippocampal dentate gyrus: quantification of dissociable modifications*. Manuscript submitted for publication.
- Lopez, H., Burger, B., & Levy, W. B. (1985). The asymptotic limits of long-term potentiation/depression are independently controlled. *Society for Neuroscience Abstracts*, *11*, 930.
- Mathai, A. M., & Rathie, P. N. (Eds.). (1975). *Basic concepts in information theory and statistics*. New York: Wiley.
- McNaughton, B. L., Douglas, R. M., & Goddard, G. V. (1978). Synaptic enhancement in fascia dentata: Cooperativity among coactive afferents. *Brain Research*, *157*, 277-293.
- Minsky, M. L., & Papert, S. A. (Eds.). (1969). *Perceptrons*. Cambridge, MA: MIT Press.
- Moore, S. D., & Levy, W. B. (1986). Association of heterogeneous afferents produces long-term potentiation. *Society for Neuroscience Abstracts*, *12*, 504.
- Moore, S. D., & Levy, W. B. (1989). *Heterogeneous synaptic activation can permit long-term potentiation in the hippocampus*. Manuscript submitted for publication.
- Murray, E. A., & Mishkin, M. (1987). Experimental studies of memory in monkeys. Implications for understanding human memory disorders. *National Forum*, *67*, 33-37.
- O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. London: Oxford University Press.
- Olton, D. S. (1978). Characteristics of spatial memory. In S. H. Hulse, H. Fowler, & W. K. Honig (Eds.), *Cognitive processes in animal behavior* (pp. 341-373). Hillsdale, NJ: Erlbaum.
- Olton, D. S. (1985). The temporal context of spatial memory. *Philosophical Transactions of the Royal Society of London, Series B*, *308*, 79-86.
- Pandya, D. N., & Kuypers, H. G. J. M. (1969). Cortico-cortical connections in the rhesus monkey. *Brain Research*, *13*, 13-36.
- Raffaële, K. C., & Olton, D. S. (1988). Hippocampal and amygdaloid involvement in working memory for nonspatial stimuli. *Behavioral Neuroscience*, *102*, 349-355.
- Ranck, J. B., Jr. (1985). Head direction cells in the deep cell layer of dorsal presubiculum in freely moving rats. In G. Buzsáki & C. H. Vanderwolf (Eds.), *Electrical activity of the archicortex* (pp. 217-220). Budapest: Akadémiai Kiadó.
- Rosene, D. L., & Van Hoesen, G. W. (1987). The hippocampal formation of the primate brain. A review of some comparative aspects of cytoarchitecture and connections. In E. G. Jones & A. Peters (Eds.), *Cerebral cortex* (Vol. 6, pp. 345-456). New York: Plenum.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. W. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing, explorations in the microstructure of cognition* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.
- Saper, C. B. (1982). Convergence of autonomic and limbic connections in the insular cortex of the rat. *Journal of Comparative Neurology*, *210*, 163-173.
- Shannon, C. E. (1950). A chess-playing machine. *Scientific American*, *182*, 48-51.

- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.
- Shaw, G. L., Silverman, D. J., & Pearson, J. C. (1985). Model of cortical organization embodying a basis for a theory of information processing and memory recall. *Proceedings of the National Academy of Sciences U.S.A.*, **82**, 2364-2368.
- Shore, J. E., & Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, **IT-26**, 26-37.
- Sorensen, K. E. (1985). Projections of the entorhinal area to the striatum, nucleus accumbens, and cerebral cortex in the guinea pig. *Journal of Comparative Neurology*, **238**, 308-322.
- Sorensen, K. E., & Shipley, M. T. (1979). Projections from the subiculum to the deep layers of the ipsilateral presubicular and entorhinal cortices in the guinea pig. *Journal of Comparative Neurology*, **188**, 313-334.
- Squire, L. R. (1987). *Memory and brain*. New York: Oxford University Press.
- Staddon, J. E. R., & Hinson, J. M. (1983). Optimization: A result or a mechanism? *Science*, **221**, 976-977.
- Sutton, R. S. (1984). *Temporal credit assignment in reinforcement learning*. Unpublished doctoral dissertation, Amherst, MA: University of Massachusetts, Department of Computer and Information Science.
- Swanson, L. W., & Cowan, W. M. (1977). An autoradiographic study of the organization of the efferent connections of the hippocampal formation in the rat. *Journal of Comparative Neurology*, **172**, 49-84.
- Swanson, L. W., Köhler, C., Björklund, A. (1987). The limbic region. 1: The septohippocampal system. In A. Björklund, T. Hökfelt, & L. W. Swanson (Eds.), *Handbook of chemical neuroanatomy* (Vol. 5, pp. 125-277). Amsterdam: Elsevier.
- Thompson, R. F., & Spencer, W. A. (1966). Habituation: A model phenomenon for the study of neuronal substrates of behavior. *Psychological Review*, **73**, 16-43.
- Van Hoesen, G. W., & Pandya, D. N. (1975). Some connections of the entorhinal (area 28) and perirhinal (area 35) cortices of the rhesus monkey. III. Efferent connections. *Brain Research*, **95**, 39-59.
- Van Hoesen, G. W., Pandya, D. N., & Butters, N. (1972). Cortical afferents to the entorhinal cortex of the rhesus monkey. *Science*, **175**, 1471-1473.
- Van Hoesen, G. W., Pandya, D. N., & Butters, N. (1975). Some connections of the entorhinal (area 28) and perirhinal (area 35) cortices of the rhesus monkey. II. Frontal lobe afferents. *Brain Research*, **95**, 25-38.
- Watanabe, S. (1969). *Knowing and guessing. A quantitative study of inference and information*. New York: Wiley.
- White, G., Levy, W. B., & Steward, O. (1988). Evidence that associative interactions between afferents during the induction of long-term potentiation occur within local dendritic domains. *Proceedings of the National Academy of Sciences U.S.A.*, **85**, 2368-2372.
- Wilson, R. C., Levy, W. B., & Steward, O. (1979). Functional effects of lesion-induced plasticity: Long term potentiation in normal and lesion-induced temporodentate connections. *Brain Research*, **176**, 65-78.
- Wilson, R. C., Levy, W. B., & Steward, O. (1981). Changes in the translation of synaptic excitation to dentate granule cell discharge accompanying long term potentiation. II. An evaluation of mechanisms utilizing the dentate gyrus dually innervated by surviving ipsilateral and sprouted crossed temporodentate inputs. *Journal of Neurophysiology*, **46**, 339-355.

- Winston, P. H. (1977). *Artificial intelligence*. Reading, MA: Addison-Wesley.
- Young, J. Z. (Ed.). (1970). *The life of mammals*. Oxford: Clarendon Press.
- Zipser, D. (1986). A model of hippocampal learning during classical conditioning. *Behavioral Neuroscience*, **100**, 764-776.
- Zucker, S. W. (1981). *Computer vision and human perception*. Paper presented at the Seventh Joint International Conference on Artificial Intelligence, Vancouver.