

Neural Networks 18 (2005) 1242-1264.

The original publication is available at  
<http://www.sciencedirect.com/science/journal/08936080>

## **Interpreting Hippocampal Function as Recoding and Forecasting**

William B Levy<sup>1</sup>, Ashlie B. Hocking<sup>1,2</sup>, Xiangbao Wu<sup>1</sup>

<sup>1</sup> Department of Neurosurgery, University of Virginia, Charlottesville, VA

<sup>2</sup> Department of Computer Science, University of Virginia, Charlottesville, VA

Correspondence should be sent to:

Dr. William B Levy

University of Virginia Health System

P.O. Box 800420 Neurosurgery

Charlottesville, VA 22908-0420, USA

Phone: (434) 924-5014, Fax: (434) 982-3829, email: [wbl@virginia.edu](mailto:wbl@virginia.edu)

## **Abstract**

A model of hippocampal function, centered on region CA3, reproduces many of the cognitive and behavioral functions ascribed to the hippocampus. Where there is precise stimulus control and detailed quantitative data, this model reproduces the quantitative behavioral results. Underlying the model is a recoding conjecture of hippocampal computational function. The expanded conjecture includes a special role for randomization and, as recoding progresses with experience, the occurrence of sequence learning and sequence compression. These functions support the putative higher-order hippocampal function, i.e., production of representations readable by a linear decoder and suitable for both neocortical storage and forecasting. Simulations confirm the critical importance of randomly driven recoding and the neurocognitive relevance of sequence learning and compression. Two forms of sequence compression exist, on-line and off-line compression: both are conjectured to support neocortical encoding of context and declarative memory as described by Cohen and Eichenbaum (1993).

*Keywords:* Prediction; Context; Recoding; CA3; Synaptic modification; Randomization; Chaotic; Cognition; Chunking; Transverse patterning; Transitive inference; Trace conditioning; Cognitive mapping

## **The Theory**

The fundamental observations of Milner (1972) concerning hippocampal function and its subsequent refinement by Cohen and Squire (1980) and by Cohen and Eichenbaum (1993) direct our ideas about hippocampal function concerning learning and declarative memory. Based on their ideas and observations as well as the anatomical relationship of the hippocampus with neocortex and the intrinsic anatomy of the hippocampus itself, we have concentrated on a recoding theory of hippocampal function (Levy 1985, 1989, 1990a, 1994). At the same time, the animal literature, with its emphasis on spatial, contextual, and configural learning, led us (Levy 1989) to include a sequence prediction aspect to this theory as we incorporated the insights of O'Keefe and Nadel (1978), Hirsh (1974), Kesner and Hardy (1983), and eventually Rudy and Sutherland (1995).

Thus, our theory (Levy 1989) arises from the confluence of several ideas: the basic function of the hippocampus as a cognitive map; the particular anatomy and detailed connectivity of the hippocampus (sparse recurrence in CA3 with divergence of entorhinal cortex (EC) inputs and recurrent signals versus convergence coming out of CA3 via the sequential projections to CA1, subiculum, and EC); the general need for a device to find associations that the neocortex would have trouble creating due to a lack of connectivity (Levy 1994); and last but not least, the need to encode correlations across behaviorally relevant time-spans for the purpose of forecasting. Thus, the combination of these perspectives leads to a hippocampal theory that conceptualizes a sequence learning device, as well as conjecturing a random recoder.

Others who model hippocampal function as sequence learning include Abbott and Blum (1996), Hasselmo's laboratory (e.g., Hasselmo et al. 2002; Molyneaux & Hasselmo 2002), Tsodyks et al. (1996), Mehta et al. (1997), Schmajuk (2002), and Treves (2004). On the other hand, McClelland et al. (1995), and Rolls et al. (1997) advocate more conventional pattern recognition models. In terms of pattern recall, speed of

convergence is quite rapid for the integrate-and-fire model (Panzeri et al. 2001, Rolls and Treves 1998, Treves 1993, Treves et al. 1997). Moreover, recently the Rolls' laboratory has begun to investigate sequence learning in their models (Stringer et al. 2004). From our viewpoint, almost all neurons are pattern recognition devices. Therefore, there is no argument about the existence of this ability in the hippocampus. Rather, sequence-learning models have pattern recognition as one of several capabilities (see Levy 1996 for a summary).

To establish and refine the viability of our theoretical perspective, we have produced a series of computer simulations. The first successful simulations of sequence learning were reported in Minai and Levy (1993b) and Minai et al. (1994). More recently we have been able to move beyond such qualitative observations. Our hippocampal model is now able to reproduce the quantitative data of hippocampal-dependent phenomena where the relevant stimuli are under precise control (see Appendix). Here we emphasize the stochastic dependence of the recoding dynamics.

The fundamental recoding by the hippocampal formation occurs in the CA3 subregion of the hippocampus (Levy 1989). Therefore, we study hippocampal function with models that emphasize this subregion.

[Insert figure 1 and table 1 near here]

### **A Family of Models**

Instead of a single model, we use a family of CA3 models. All the members of this family share certain basic biological properties that are summarized in Table 1. The primary neurons are spike passing, i.e., communication is binary  $\{0,1\}$ . Where we use McCulloch-Pitts neurons, the updating of internal excitation and the axonal communication lag are the same. This is not true for models using integrate-and-fire neurons. The input layer (Fig. 1a) is a merging of the entorhinal cortex (EC) and dentate gyrus (DG) inputs to CA3. Fundamentally, the CA3 model is a sparsely interconnected

feedback network (Fig. 1b), typically with thousands of neurons in a simulation. All direct, recurrent connections between primary cells are excitatory. There is an interneuron mediating feedforward inhibition, and one mediating feedback inhibition. Inhibition is of the divisive form, but activity in the free-running models is only imperfectly controlled because of a delay in the feedback that activates these inhibitory neurons.

To date, region CA3 is modeled as a randomly connected network. Each excitatory neuron randomly connects to approximately  $n \cdot c$  other neurons, where  $n$  is the number of neurons and  $c$  is the connectivity ratio (usually set to 0.1 but lower connectivities also work, e.g., Sullivan and Levy (2004) and Levy et al. (2005)). Given the output of neuron  $i$  at time  $t$ , here  $z_i(t)$ , the net internal excitation of neuron  $j$ ,  $y_j(t)$ , is

$$y_j(t) = \frac{\sum_{i=1}^n w_{ij} c_{ij} \phi(z_i(t-1))}{\sum_{i=1}^n w_{ij} c_{ij} \phi(z_i(t-1)) + K_{FB} \sum_{i=1}^n w_{il} z_i(t-1) + K_0 + K_{FF} \sum_{i=1}^n x_i(t)} \quad (1)$$

where  $w_{ij}$  represents the weight value between neurons  $i$  and  $j$  at time  $t-1$ , and  $c_{ij}$  is a binary variable  $\{0,1\}$ , indicating whether or not there is a connection from neuron  $i$  to  $j$ . The term  $\sum w_{ij} c_{ij} \phi(z_i(t-1))$  represents the excitatory synaptic conductance for the  $j^{\text{th}}$  neuron. Parameters  $K_{FB}$  and  $K_{FF}$  are constants that scale the feedback and feedforward inhibitions, respectively. The constant  $K_0$  controls the magnitude and stability of activity oscillations and is analogous to a shunting rest conductance (Smith et al. 2000). Weights  $w_{il}$  are the positively valued synaptic strengths between each pyramidal cell  $i$  and the feedback inhibitory neuron at time  $t-1$ . The binary external input to neuron  $j$  at time  $t$  is indicated by  $x_j(t)$ . If either  $x_j(t) = 1$  or  $y_j(t) \geq \theta$ , neuron  $j$  fires (i.e.  $z_j(t) = 1$ ), where  $\theta$  is a threshold fixed at 0.5.

Synaptic failures can be included via a synaptic failure channel represented by the function  $\phi(z_j(t))$  for the connection from neuron  $i$  to neuron  $j$  (Sullivan & Levy 2003a,

2004). Here  $\phi(z_j = 0) = 0$ . A synaptic failure,  $\phi(z_j = 1) = 0$ , occurs with probability  $f$ , and successful synaptic activation,  $\phi(z_j = 1) = 1$ , with probability  $(1 - f)$ ; i.e., the failure process is a Bernoulli random variable that acts independently on each synapse at each time-step. The addition of failures allows successful simulations to run at lower activity levels (Sullivan & Levy 2004).

The model uses a biologically-inspired postsynaptic associative modification rule with potentiation and depression (Levy & Steward 1979) and a time staggering between pre- and postsynaptic activity (Levy & Steward 1983). A minimal version of this synaptic modification takes the form

$$w_{ij}(t+1) = w_{ij}(t) + \mu z_j(t) (z_i(t-1) - w_{ij}(t)) \quad (2)$$

where  $i$  is input,  $j$  is output, and  $\mu$  is the synaptic modification rate.

For more biological simulations, synaptic modification spans multiple time-steps, (see, e.g., August & Levy 1999; Rodriguez & Levy 2001); specifically in the case of Rodriguez and Levy

$$w_{ij}(t+1) = w_{ij}(t) + \mu z_j(t) (\bar{z}_i(t-1) - w_{ij}(t)), \quad (3)$$

where

$$\bar{z}_i(t-1) = \begin{cases} \bar{z}_i(t-2)\alpha & \text{if } \phi(z_i(t-1)) = 0 \\ 1 & \text{if } \phi(z_i(t-1)) = 1 \end{cases} \quad (4)$$

and  $\alpha$  represents the decay time constant of the NMDA-receptor (NMDA-R). The decay of the glutamate-like priming of the NMDA-R is assumed to be exponential (see Levy & Sederberg 1997, Mitman et al. 2003, August & Levy 1999).

For better control of activity, a rule for modification of pyramidal-to-interneuron synaptic strengths has been incorporated in recent work (Sullivan & Levy 2003b).

Specifically,

$$w_{iI}(t+1) = w_{iI}(t) + \lambda z_i(t-1) \left[ \frac{\sum z_i(t)}{n} - a \right], \quad (5)$$

where  $w_{iI}$  is the weight of excitatory connection from neuron  $i$  to the feedback interneuron,  $\lambda$  is the pyramidal-interneuron synaptic modification rate constant, and  $a$  is the desired percentage of active neurons. Although this additional biology is unnecessary if the parameters  $K_0$ ,  $K_{FB}$ , and  $K_{FF}$  of Eq. (1) are set with extreme care, it greatly simplifies investigations that are devoted to parameter sweeps.

### **Recoding With a Purpose**

Barlow (1959), Watanabe (1961), and Dretske (1981) interpreted sensory and even cognitive processing as lossless recoding to enhance simplicity. Our early thinking (Levy 1985) was inspired by these proposals, and application of their perspective to sequences led to our tight packing interpretation of CA3 encodings (Fig. 2). However, the goal of recoding is not some idyllic Dretskean representation suitable for philosophical or scientific contemplation.

[Insert figure 2 near here]

The goal of recoding and neural representation, in general, is the organism's survival and propagation. Presumably, the learning paradigms used to define hippocampal function (e.g., trace conditioning, goal finding, and configural problem solving) help ensure that the recoding we study is behaviorally and cognitively relevant, and thus relevant to the organism's propagation. That is, we explicitly define the goal of recoding as the production of representations that are suitable for forecasting, readable by a linear decoder, and appropriate for neocortical encoding. Often such recoding entails an information-theoretic simplification, but such simplification is by no means equivalent to forecasting which is suitably accurate and linearly decodable. The successes of the model across cognitive/behavioral tasks requiring normal hippocampal function (see Appendix and summary in Table 2) support our claims of relevance.

[Insert table 2 near here]

Here we reiterate and extend the claim that recoding, in the service of neocortex and for forecasting, is fundamental for understanding hippocampal function. Generically, neocortical representations need help finding associations for two reasons: lack of neocortical connectivity and the incompatible timescale of associations in the world versus associative capabilities at a synapse (Levy 1989, 1994, Levy & Steward 1983, Holmes & Levy 1990). Mechanistically, the same set of processes combine to solve both of these generic problems (see Fig. 14 in Levy 1989). Furthermore, we claim that the recoding perspective unifies the various expository theories of hippocampal function referenced earlier. To explain the unification, we must explain the codes in terms of their utility. Then we describe our understanding of the mechanisms underlying recoding.

*What Recoding Does.* An appropriate theme for interpreting computation and information processing in the brain is signal transformation. Note that the hippocampus lacks both direct sensory input and direct motor control. This insight, that it does not perceive nor decide to move a muscle, is consistent with observations such as the case of H.M. On the other hand, a reciprocal relationship exists between the entorhinal cortex and other association cortices (Swanson & Köhler 1986, Swanson et al. 1987). There are no other direct associative cortical recipients with the exception of orbital and ventromedial prefrontal cortex. Thus, the hippocampal formation must be a recoder for the reciprocally connected cortices, and it is the last one in the hierarchy of sensory information processing. As the recoder of last resort, the episodic theory (Cohen & Squire 1980) implies that the hippocampus serves the neocortex to produce context-based memories and more. However, the hippocampus can, in some cases, solve problems before declarative knowledge is available (Chun & Phelps 1999, Greene et al. 2001). The Greene et al. (2001) result implies an additional function of region CA3, the ability to produce recodings suitable for sequence prediction that solve configural

learning problems (e.g., see Levy 1996, Rudy & Sutherland 1995, and Cohen & Eichenbaum 1993.) In any event, it is central to our theory that the hippocampus is the recoder used by association cortices when they fail to encode the appropriate associations.

The argument is supported by specific examples which are hard to interrelate at the cognitive level. That is, the set of prototypical learning problems that define the hippocampal contribution to cognitive function are disparate, and the underlying similarities seem far from obvious. The functions that CA3 recoding supports include: (1) contextual associations (Hirsh 1974), (2) configural learning that includes the quite distinct demands of the transitive inference and the transverse patterning tasks (Alvarado & Rudy 1992, 1995; Dusek & Eichenbaum 1997), (3) cognitive mapping (O'Keefe & Nadel 1978), and (4) trace conditioning (Solomon et al. 1986). Moving from the cognitive perspective to a machine learning perspective of hippocampal-dependent tasks still leaves us with a disparate set of functional requirements; at minimum, the functional requirements include forecasting, conditional pattern recognition, generalization, and discrimination. Historically, the thread running through most, if not all hippocampal theories, is an attempt to unify such disparate cognition (e.g., Rolls & Treves 1998, McClelland et al. 1995) or functional processing (e.g., Cohen & Squire 1980, O'Keefe & Nadel 1978, Cohen & Eichenbaum 1993, and many others). It is our thesis (Levy 1989, 1996) that recoding considerations produce the most encompassing unification by explaining the hippocampal contribution to disparate functional or cognitive processes.

[Insert figures 3 and 4 near here]

*Characterizing the hippocampal recoding.* Creating suitable contextual memories requires creating context codes. Such codes consist of modestly generalized, subsequence detection devices called local context neurons (see Fig. 3). Using an

orthogonal input sequence, Fig. 4 illustrates an extreme version of the recoding problem and its solution via the repetitively firing, local context neurons. Note that there is no shared similarity of input-activated neurons representing the successive inputs in Fig. 4 panel 1, although each input has a dwell time of three time-steps and there is random firing due to recurrent activation. As training proceeds, the recoding evolves, and each neuron begins to detect a specific subsequence. The comparison between panels 3 and 4 of Fig. 4 illustrates this evolution; the approximately randomly-firing recurrent neurons of panel 3 autonomously emerge as the local context neurons appearing in panel 4. As a result of these recurrent local context neurons, the sequence of CA3 state space vectors in panel 4 are far from the orthogonal sequence that characterizes the inputs. This result is fundamental. Due to the randomly interspersed local context neurons, the recoding moves rather smoothly through a sequence of state space representations. For the same reason, adjacent CA3 vectors resemble each other more than the corresponding input vectors resemble each other. This enhanced resemblance of successive CA3 vectors is the key idea for recoding a sequence that appears in the world with some statistical regularity even if successive states in the world are quite different, as is the case in Fig. 4. In fact, the enhancement of similarity between successive vectors by recoding is the encoding of whatever statistical regularity exists across presentations of the sequence (see Levy and Wu 1996). That is, longer-lived local context neurons are equivalent to greater generalization of a subsequence, in the sense that the longer a neuron fires, the longer it is reporting the equivalence of successive patterns in a sequence of patterns.

This enhancement of similarity among successive representation vectors has been discussed in terms of compression in a geometric coding space in Levy (1989). There it is seen as being equivalent to both representational and temporal compressions (see

Fig. 2), where representational compression is automatically a context code. What was called signal mixing and tight packing in our earlier work we now call recoding.

The signal mixing interpretation is important because it, too, is fundamental in explaining how the hippocampus can solve a problem the neocortex cannot. Although we call the recurrent connectivity or even the EC to CA3 connections sparse, it is much denser than the relative connectivity between different association cortices. Signal mixing and the discovery of associations at the neuronal level is then accomplished by a spatial divergence of the EC input to DG and CA3, and by the recurrent connectivity of CA3, which spreads along the septo-temporal axis of CA3 as well as within a lamella. All this divergence leads to a random convergence of the information originally carried by the different input axons. Such signal mixing leads to discovery of pairwise associations that are encoded by time-spanning associative synaptic modification (Levy 1989). In this regard, our computational studies have emphasized the importance of recurrent connections where the input codes to be associated are orthogonal. Such studies include the simple sequence in Fig. 4 as well as configural learning problems and trace conditioning (see Appendix for details).

Regardless of what it is called, tight packing (Fig. 2) or local context codes (Fig. 3), the form of the recodings is exactly what the model produces under the appropriate regularity of training experiences. Such context codes are generalizations in the sense that they represent the approximate equivalence of successive members of a subsequence (Fig. 3). However, CA3 is not limited to generalization; it can also discriminate.

*Discrimination.* In a sense, discrimination is the opposite of generalization. Nevertheless, neurons of CA3 – particularly those with stronger external excitation – can act like neuronal decoders (e.g., Levy et al. 1990). In contrast to a CA1 decoder,

however, random neurons in CA3 are using the sparse random recurrent connectivity to discover representations that allow discrimination.

Consider the transverse patterning problem, which requires discrimination between the overlapping stimulus pairs AB, BC, and CA (see Appendix for details). For each stimulus pair, there exists a subspace of recurrent neurons that fire exclusively to that pair. Given the existence of such subspaces, the subset of neurons excited by the appropriate decision code will modify the proper synapses and contribute to the correct forecast. By working in one of these subspaces of CA3 state space, some neurons can pull out predictions that differ from the generalizations that are occurring at the level of the full dimension of the state space vector. Then, for these subspace neurons, time-spanning synaptic modification amplifies the correlated inputs and attenuates certain anti-correlated inputs. Thus, the CA3 output interpreted by a linear-decoding decision system solves transitive inference by amplifying the appropriate similarities (see Wu & Levy 1998) and solves transverse patterning by amplifying the contextual differences. (See Appendix which illustrates the same parameterized network solving both transitive inference and transverse patterning.)

This ability – the production of neurons that discriminate between different overlapping sequences based on information in that sequence – implies an additional utility of the CA3 recodings. Specifically, these recodings will be appropriate for neocortical function in the sense described by Cohen and Eichenbaum (1993). That is, the CA3 codes are appropriate for tasks requiring reconfiguration, or equivalently, the flexible reuse of chunked encodings. Thus, the Cohen-Eichenbaum generalization of declarative memory is solved by linear decoders working in the appropriate representational subspaces produced by CA3 recoding.

Additional discussion about local context firing and temporal compression follows, as we further characterize recoding and point out its general relevance to the more cognitive theories of hippocampal function.

*Context codes, Prediction, and Cognition.* Contextual representation of an experience is relevant to the episodic aspects of declarative memory. That is, in the sense that an episode is an interrelated sequence of events, it also defines a context. The interrelationship is encoded by strengthening the connections between the firing neurons that represent the salient features of the episode. Local context encoding enhances similarities in CA3 state space representations, and in this way, it defines or represents context across time. When sequences are fluctuating but fundamentally repetitive, region CA3 encodes the spatio-temporal information that constitutes an episode particularly well. The prototypical example is a rat exploring an open field for an extended period of time; such wandering exploration produces a variety of partially overlapping sensory sequences (e.g., Nadel & Willner 1980). Our simulations of simple spatial-like tasks, including a circular sequence and various maze-type learning problems (see Appendix) produce place cells that interrelate neighboring locales across space due to temporal contiguity of the experiences during training. For example, in the disambiguation problem the recent past (context past) becomes usable for decision making. This utility depends on training-induced synaptic modification which allows recent, within-trial information to propagate in time by virtue of compressed representation of a sequential encoding.

Local context neuron firing and sequence compression are at the heart of context-based prediction and forecasting. An individual local context neuron, as a generalized subsequence recognizer, is a “place cell” in a spatial task (O’Keefe & Nadel 1978). However, such a neuron cannot properly be called a place cell in transitive inference (Wu & Levy 1998), transverse patterning (Levy et al. 1996, Shon et al. 2002, Wu et al.

1998), or trace conditioning (Levy & Sederberg 1997, Rodriguez & Levy 2001) because there is no physical space relevant to the most critical aspect of the encoding. Although space is always present and may be part of such encodings, it is not relevant to the solution of these tasks.

[Insert figure 5 near here.]

*Temporal compression via overlapping representations.* Enhanced overlap of the CA3 state space vectors automatically implies forecasting; such anticipatory prediction is an inevitable outcome of temporal compression and backward, i.e., earlier in time, cascade of local context neurons (see Fig. 5). This temporal compression is a CA3 recoding that anticipates future states and, therefore, is a faster than real-time representation. Extending the shared neurons of successive representations implies the future merely by the representation of the present (Levy 1989, 1994, 1996). Such compressed and overlapped recodings lead to forecasting by simple sequence completion (see Fig. 11 in Levy 1989 and Fig. 4 here). In addition, such compressed recodings are presumably incorporated into neocortical circuitry where they can be used as chunks (Miller 1956) and reconfigurable elements (Cohen & Eichenbaum 1993).

In sum, temporal compression with a backward cascade produces a recoding that is suitable for generating predictions (forecasts) based on the hippocampal recodings themselves – so long as a decoder exists. Thus we conjecture that temporally compressed sequences, followed by their neocortical encoding, allow the neocortex to forecast without the hippocampus. This conjecture rests on the supposition that the sequence completion problem in the hippocampus becomes a pattern completion problem in neocortex. In particular, the highly compressed and overlapped encodings are suitable for rapid pattern completion by the more symmetrically connected recurrent networks of neocortex.

It is worth distinguishing two, non-exclusive mechanisms of temporal compression by CA3, off-line and on-line compression. These two forms of compression are available as hippocampal output at distinctly different times. Off-line compression is a high-speed replay phenomenon, whereas on-line compression occurs in real-time.

*Off-line compression.* This compression occurs when the animal is in neocortically defined slow-wave sleep (SWS) or occasionally while the animal is awake and performing relatively mindless tasks such as grooming or eating (based on Buzsáki (1996) sharp wave observations). Compared to on-line compression, which occurs when CA3 cell firing is sluggish, off-line compression is correlated with high levels of hippocampal cell firing. Off-line compression arises spontaneously from CA3 itself (Buzsáki 1996), and this spontaneous form of compression is easily reproduced in the model (August & Levy 1996, 1997, 1999). Thus the model isolates the critical biology for the fast-replay, off-line compression.

[Insert figure 6 near here]

According to our simulations, off-line compression (see Fig. 6 and Appendix for details) occurs when the hippocampus is allowed to free-run with little external excitation but with greater total activity than when the neocortex is driving the hippocampus. Allowing increased activity by reducing inhibition was inspired by the old observations of Green and Arduini (1954) and the fact, well-known to hippocampal neurophysiologists, that hippocampal single unit activity increases substantially in SWS (e.g., Thompson & Best 1989). The compression in our high temporal resolution integrate-and-fire simulations is 15-30 fold (August & Levy 1999), which compares well with published neurophysiological results (Wilson & McNaughton 1994).

*On-line compression.* This second type of compression occurs in the attentive animal whenever the CA3 recodings develop by extension of local context lengths (see Fig. 4 for an example of the formation of local context neurons with training). These new

representations are driven directly by episodic experience. In turn, this extended firing enhances the similarity between successive representations of a sequence compared to its driving input codes.

Two dynamic processes that are hard to separate when explaining the development of on-line compression across training are the backward cascade and the development of local context neurons. Specifically, the training-induced increase in context length of a subset of individual neurons co-occurs with earlier firing of most of these neurons. Contrast panel 1 versus panel 2 in Fig. 4 and upper and lower graphs of Fig. 7 to see that many neurons turned on externally tend to turn on before the external activation. This tendency to fire earlier, i.e., backward cascading as training proceeds, depends on the relative freedom a neuron has in terms of its position within a sequence (Mitman et al. 2003). For example, in regard to this freedom to recode, externally activated neurons are, in large part, anchored to a position within a sequence while recurrently activated neurons are relatively free to be repositioned (compare Fig. 7, which shows external firing after training, to Fig. 5 which shows recurrences). The phenomena of backward cascade and extension of context length occur virtually without exception in the paradigms we have studied. Nevertheless, they are best understood in trace conditioning and in simple sequence learning.

[Insert figures 7 and 8 near here]

The autonomous recoding solution to the trace conditioning problem is arguably the prototypical example of on-line compression (see Appendix). To solve the escape version of trace conditioning, the unconditioned stimulus (UCS) must be predicted before its actual occurrence (Solomon et al. 1986). A prediction (forecast) takes the form of activating enough UCS neurons via recurrent connections before the external UCS activation occurs. The backward cascade of UCS encoding neurons occurs by a backward extension of local context length because UCS neurons are strongly anchored

by their external activation. Figure 8 shows an example of a backward cascade by extension of context length. That is, the backward cascade of UCS neurons occurs by virtue of the onset firing times of some of these neurons coming earlier within a trial. Other neurons also change both their onset and offset time, and this is predominantly in the direction of earlier firing. For more details of the complexity of the dynamic of such a cascade see Fig. 4 of Levy et al. (2005).

### **Why the Network Anticipates Its Own States and a Quantitative Definition of Context**

Very little in the way of useful recoding would occur without the asymmetric associative modification rule. It produces the backward cascade, creates local context neurons that are generalized subsequence recognizers, and produces the discriminative firings needed for transverse patterning. The explanation may seem strange to those unfamiliar with Bayesian inversion and the effect of stationarity, but the explanation is rather simple once the appropriate time-spanning synaptic modification rule is in effect.

By taking the expected value of Eq. (2) or (3) and presuming that  $E[\Delta w_{ij}] \rightarrow 0$  over training, we see that synapses tend to take values proportional to  $E[Z_i(\textit{past})|Z_j(\textit{present})=1]$  where  $i$  is an input to cell  $j$  and the pre- and postsynaptic random variables are binary  $\{0,1\}$  and where we are calling  $t$  in Eq. (2) and (3) the present and  $(t-1)$ , or earlier, the past. However, wherever this expectation is limited to a subsequence that is approximately stationary, we can translate in time:

$$E[Z_i(\textit{past})|Z_j(\textit{present})=1] \equiv E[Z_i(\textit{present})|Z_j(\textit{future})=1].$$

Thus the conditional expectation on the right makes explicit a statistical relationship between the future and present that is stored at  $j$ 's afferent synapses.

Along the same lines, if average neuronal firing is approximately guaranteed (or is quantified by the postsynaptic neuron  $j$ ), then

$E[Z_j] \equiv E[Z_j(\textit{present})] \equiv E[Z_j(\textit{future})]$  and likewise for  $E[Z_i]$ . Furthermore, via

a Bayesian inversion and via a method of statistical inference such as maximum entropy (Jaynes 1979), a probability distribution is implicit (Levy 1990b),

$E[Z_i(\textit{present})|Z_j(\textit{future})=1] \Rightarrow P(Z_j(\textit{future})|Z_i(\textit{present}))$ .

As pointed out by Hocking and Levy (2005), even the crude minimal model with its simplistic form of neuronal integration, Eq. (1), is approximately a Bayesian inversion, and the possibility that hippocampal neurons produce a computation that is even closer to Bayesian inversion has not escaped comment (Levy et al. 1990, Hocking & Levy 2005). More to the point, prediction of future states leads to a backward cascade due to the extension of cell firing in an earlier direction.

Via this time-spanning synaptic modification, earlier-firing neurons influence the activation of later firing neurons. **As a result, these later firing neurons begin to fire earlier – the equivalent of a forecast of their own future activations.**

Finally, it is notable that if each neuron is producing a conditional probability, then each conditioning variable is itself a microscopic context. Thus, context is quantitatively defined as the state of the conditioning variable of a conditional probability. Therefore each neuron is a microscopic context for its postsynaptic targets.

### **What Drives Recoding?**

We have studied a variety of factors that influence recoding and the development of local context neuron firing. These studies produced insights as reflected by certain characteristic relationships of network properties, e.g., a relationship between the values of synaptic weights and firing patterns (Amarasingham & Levy 1998). Other examples include the cell firing patterns that develop as a function of training; that is, longer local

context neuronal firings develop when there is more repetition of sequential input information (Wu et al. 1998) and when there is high reliability/low variability among the external input sequences across the training experience (Wu et al. 1996). Also influencing local context neuron firing is the parameterization of the synaptic modification rule (Eq. (3)) (e.g., Mitman et al. 2003) and the rate at which the input changes within a sequence (Levy & Wu 1996, Wu et al. 1998). However, at the heart of the theory (Levy 1989) is random recoding. Therefore, this last section concentrates on random processes because of their fundamental role in recoding.

Metaphorically, the idea inspiring a role for randomness in recoding is Jaynes' (1979) maximum entropy idea, applied to the hierarchy of neocortical sensory recoding systems. The differing anatomies (i.e., connectivities) and physiologies of various neocortical regions reflect different prior distributions on the received information, with the hippocampal system abandoning the neocortical priors. In regard to the hippocampal system as the associator of last resort, the maximum entropy philosophy says that when all else has failed, one should abandon the preconceived biases of the neocortical regions projecting into the upper layers of the EC and use the flattest prior appropriate to the information actually available. In terms of physiology, the a priori probability of firing should be the same for any primary neuron in the network. Likewise, the anatomy should reflect a flat prior. The contrast with neocortex makes the point. The topological nature of neocortical connectivity – the Mexican hat function of adjacent neocortical tissue and the sparsity of distant connections – is a biasing of the encodable associations. The anatomy of the EC-CA3 and CA3-CA3 systems is a connectivity characterized by an input spreading divergence (see Levy 1989) and conjectured random connectivity. Such spreading and randomization tends to overcome much of the preconceived (and appropriate) biases that are inherent within the input codes themselves.

Amplifying this divergence and most fundamental to recoding is the sparse, excitatory, random, recurrent connectivity of CA3 (Levy 1989). Such sparse connectivity has been part of our model from the first simulations, where an emphasis was placed on the role of such sparsity in producing a nonreciprocal (asymmetric) connectivity (Minai et al. 1994, Minai & Levy 1993a, b, c, 1994). This, in turn, helps produce the nonconvergent wandering through CA3 state space. This chaotic dynamic in CA3 state space is, in fact, the medium used for encoding sequential patterns, but this dynamic alone is not enough. Although the sparse connectivity of the model is the foundation of the recoding process, producing reliably useful recodings requires more. To take advantage of this sparse connectivity and chaotic dynamic, the model requires a trial-to-trial randomizing process, and it benefits from the appropriate time-spanning associative modification (Mitman et al. 2003). Here we concentrate on results that provide qualitative and quantitative insights into the role played by randomization. There are four processes we have studied in this regard:

- 1) chaotic, i.e., unpredictable deterministic, activity fluctuations
- 2) random initialization of state space,  $Z(0)$ , at the beginning of each trial
- 3) quantal synaptic failures at recurrent excitatory synapses, and
- 4) the relative strength of the external inputs.

Table 3 summarizes the variety of paradigms and methods used to study the effects of randomization, and it outlines interactions of randomization with such fundamental parameters as activity and connectivity. The overall interpretation of these results is that randomization drives an undirected code word “search”. Randomization does this by counteracting, from one training trial to next, a tendency for too much similarity between sequences of state space representations.

[Insert Table 3 near here]

*Chaotic activity fluctuations.* Sparse random connectivity of the model and delayed inhibition lead to chaotic fluctuations, which can be further randomized by other processes such as initial state randomization and quantal synaptic failures. Regarding these activity fluctuations, there are four relevant observations: (i) the activity fluctuations in the deterministic version of the free-running model are of the chaotic (aperiodic) type (Minai & Levy 1993a, b, c); (ii) for simulations of the transitive inference problem using a moderate number of neurons (1024), learning is almost nonexistent when these oscillations are suppressed by using a competitive mechanism for activity control (Levy & Wu 2000); (iii) the same can be said in the case of TP when simulated with 512 or 1024 neurons (Sullivan & Levy 2004); and (iv) transitive inference simulations parameterized to produce mild, high-frequency activity oscillations substantially outperform simulations with low-frequency oscillations, although total mean square fluctuations are of equal power (see Fig. 7 of Smith et al. 2000). On the other hand, overly large, chaotic activity fluctuations destroy performance (ibid.) because such large fluctuations destroy the flow of information (i.e., the probability of individual cell firings tends toward one or zero) across sequential states.

*Initial state randomization and sensitivity to initial conditions.* Presumably, in the time between training trials, there are small differences in sensory experiences that cause randomization of the initial state of CA3 at the beginning of the each training trial. When such initialization is the only randomizing process beyond the chaotic tendencies of the model, its role is critical for producing performance that matches behavioral observations. This can be shown in two ways. The length of  $Z(0)$  can be varied (Levy & Wu 2000) or for fixed length  $Z(0)$ , the amount of trial-to-trial randomization can be varied (Shon et al. 2002). Figure 9 shows the result of varying the randomization of the fixed length  $Z(0)$ . Overall the best learning of transverse patterning occurs when initial state randomization,  $Z(0)$ , is maximized. Without randomization, performance fails to reach the

standard criterion for any of the individual simulations run. As initial state randomization increases beyond one-half of its maximal value, criterion performance occurs with increasing frequency, but maximal randomization is what benefits performance the most (Wu & Levy 1999, Levy & Wu 2000, Shon et al. 2002).

[Insert figure 9 near here]

The explanation of this positive effect for enhanced trial-to-trial fluctuations is called “code word search.” In metaphorical language, initial state randomization leads to something like a random search through state space during critical training trials. However, we can go beyond metaphor and quantitatively define this code word search in simulations and a theorem. Figure 7 of Shon et al. (2002) shows how simulations with large  $Z(0)$  wander through state space when a novel sequence is introduced after training on another sequence. Because the different contingencies of TP training are so similar, substantial changes in state space sequences, even with the introduction of novel sequences, do not occur when  $Z(0)$  is small. However, with large  $Z(0)$  randomizations, novel input sequences drive the system to try out rather different state space sequences. With the occurrence of sufficiently novel state space sequences, time-spanning associative synaptic modification, based on the regularity of decision and outcome, captures an appropriate state space sequence without destroying the previous learning of other input sequences.

After documenting various aspects of the effect of  $Z(0)$  randomization, Shon et al. (2002) refine the idea of random search for code words by quantifying the randomization-driven state space dynamics and then producing a theorem. This theorem makes precise how the variables in the model affect the trial-to-trial variation of the neurons that fire early in training. This variation is what is meant by the metaphorical term “search”. The theorem quantifies the average normalized hamming distance between  $Z(1)$  states as a function of number of neurons, activity, connectivity, external

activity, and firing thresholds. That is, the theorem quantifies how initial state randomization affects the variation of neuron firing patterns when new information is introduced during training. In the case of the initial training trials, the predictions of the theorem are confirmed by simulations (see Shon et al. 2002 for details). In sum,  $Z(0)$  randomization is required if these simulations are to learn TP, and this randomization exerts its beneficial effects by causing a simulation to try out different paths through state space when novel subtasks are introduced.

*Quantal synaptic failures.* Quantal synaptic failures can pass for a true randomizing process in the sense that the failure or success of transmitter release at one synapse of an active axon is statistically independent of what happens at another synapse of that axon. Moreover, the failure rates in the hippocampus are particularly high, perhaps in the range of 55-85% (Miles & Wong 1986, Stevens & Wang 1994, Thomson 2000).

Studying a model with particularly poor performance in the transverse patterning task (the competitive model using the simple, one time-step synaptic modification rule, Eq. (2)) leads to a surprisingly powerful demonstration of the cognitive enhancing power of this randomizing and information losing process (Levy & Baxter 2002). In particular, performance is substantially enhanced over a broad range of simulation parameters by adding quantal failures to the model. As Fig. 10 demonstrates, this more robust performance allows simulations to run at much lower activity levels, which are more biologically appropriate (Sullivan & Levy 2004). Such low levels of activity also produce greater sequence length memory capacity (Levy & Wu 1996).

This result, that synaptic failures benefit simulations run at lower activity, generalizes to the TI task in a free-running network (see Fig. 11, compare 0% failures to 50% failures).

[Insert figures 10 and 11 near here]

Counteracting these three randomization processes (deterministic chaos,  $Z(0)$ , and synaptic failures) is deterministic external activation. This activation dictates a subset of neurons that must fire.

*External activation.* Information is carried from the neocortex via external activation of CA3 neurons by layer II entorhinal cortex cells, both directly to CA3 and indirectly through the dentate gyrus. Without this information, there is nothing worth recoding. On the other hand, the stronger external excitation is relative to recurrent excitation, the greater the restrictions on CA3 recoding. That is, if relative external excitation is too powerful, then there is little freedom to recode because the firing patterns of CA3 will be selected too precisely by the inputs. In terms of following the dictates of the input sequence, some imprecision is not only desirable but necessary, because the input codes, as a representation of the information in the task, are a failed neocortical representation for the cognitive task to be learned. Therefore, these sequential input representations must be modified. The dual needs – to modify and to preserve input information – require a compromise in the relative amount of externally produced CA3 activity. This ratio is quantified as  $\sum_i X_i(t) (= m_e(t))$  divided by the total number of active neurons per timestep.

The best compromise ratio is in the vicinity of 30-35% for externally driven CA3 cell firing versus total firing (Polyn et al. 2000; see also Appendix here; Levy & Wu, submitted; see also Fig. 11). Note that this ratio of 30-35% for external activation is still valid when synaptic failures are introduced (see Fig. 11). This means that the best performance requires 65-70% of the firing to be recurrently driven. Arguably, these numbers are somewhat artificial because a spectrum of such ratios exists across the CA3c-to-CA2 axis. Nevertheless, in the spirit of minimal modeling, it is our simplification

of external activation,  $X_i(t) \in \{0,1\}$ , that makes clear the functional importance of this ratio, and it is this simplification that allows such a straightforward quantification.

Finally, in addition to the importance of the three randomizing processes (chaotic, initial state, and quantal failures) and the compromise ratio of external activation necessary to take appropriate advantage of the randomization processes, our simulations predict that randomization processes will also have a macroscopic manifestation. Specifically, the variation of individual learning across subjects, which is reproduced across simulations (Levy et al. 2003, and Fig. 12).

[Insert figure 12 near here]

In sum, our research demonstrates the utility of the recoding perspective, including quantifying the interactions between relevant biological variables. Nevertheless, we do not propose that the recoding perspective is the only correct one. Indeed, regarding higher, intermediate, and lower level hippocampal function, the multiplicity of interpretations guiding the research of our colleagues is valid and important. However, as a unifying theme and as a compass to direct research questions, the recoding perspective provides insights and emphasizes questions not yet addressed by other hypotheses.

## **Appendix**

### **Description of Various Paradigms and Results**

Of equal importance to the biological features of the model, are the restricted set of learning paradigms that the model claims as its purview. The model is restricted to those tasks that require a functioning hippocampus to learn, but after a suitable amount of training, lose their original hippocampal dependency. This restricted definition is not arbitrary but defines the essence of the special role the hippocampus plays in memory

formation (Milner 1972; Squire 1987). In this regard the family of minimal models reproduces a surprisingly long list of hippocampal-dependent phenomena. These include transitive inference, transverse patterning, the early stages of spatio-temporal context encoding and cognitive mapping, and the air-puff escape version of the trace conditioning paradigm (Solomon et al. 1986). The data that simulations of the model should reproduce in these paradigms include what is or is not learnable, the rate at which performance improves, and the learned cell firing characteristics, particularly during the time that hippocampal function is critical in the paradigm.

Here we summarize some of the paradigms mentioned earlier as well as other successful simulations, which reproduce hippocampal functions. Table 2 summarizes most of the model's successes.

### **Simple Sequence Completion**

As a sequence prediction device, the first task the model must be able to solve is simple sequence completion. The sequence is a temporally ordered set of patterns, each represented by a set of externally driven neurons. For example, in Fig. 4 each pattern corresponds to forty externally driven neurons. To test a network, we activate the first pattern in the sequence. Then, if the network activates the remaining patterns in order, the network has learned the sequence. Such learning has been shown in our model under a variety of circumstances (e.g. Minai & Levy 1993a; Wu et al. 1996).

### **Spontaneous Rebroadcast with Temporal Compression**

One neurophysiological observation, important to the role of the hippocampus as a recoder servicing the neocortex and a role that requires the appropriate scaling of time, is the temporal compression of a hippocampally encoded sequence, or episode, during spontaneous replay.

Spontaneous replay after training was one of the first properties we demonstrated (Minai & Levy 1993b). However, quantitatively accurate compression requires a more

complex model, i.e., at least an integrate-and-fire model (compare Levy et al. 1998 to August & Levy 1999). Temporal compression of a learned sequence or episode occurs in the hippocampus when such a sequence of neuronal firing patterns, which occurred earlier in the day, is spontaneously replayed. This replay occurs during sleep at an increased rate of speed without an external cue. Such temporal compression occurs during slow-wave sleep (SWS) and has also been observed during moments of relative inactivity while the animal is awake (Buzsáki et al. 1992, Pavlides & Winson 1989, Skaggs & McNaughton 1996, Skaggs et al. 1996, Wilson & McNaughton 1994). Skaggs et al. (1996) report compression ratios as high as 10:1 during SWS. It is hypothesized that this compressed replay is for purposes of consolidating memories in the cerebral cortex that were stored in the hippocampus (Skaggs & McNaughton 1996, Wilson & McNaughton 1994). Just as in the experiments, after learning a circular sequence (Fig. 6A), spontaneous replay occurs in simulations of the model (Fig. 6B). That is, after training, starting a network in a random initial activity state eventually leads to a fast replay of a previously learned sequence (August & Levy 1999; Levy et al. 1998). The measured compression corresponds to what neurophysiologists observe.

### **Finding a Shortcut**

Distinct from jump-ahead recall (Prepscius & Levy 1994) is the ability of a network to find a shortcut. Spatially, a shortcut is a continuous path leading from a source to a destination that is shorter than paths previously experienced. Similarly, we say a network has found a shortcut when it creates a sequence that is logically continuous and is shorter than other sequences it has encountered that bridge the first and last patterns. For a sequence to be logically continuous, the pair of patterns produced at time  $t$  and  $(t + 1)$  must be pairs that have been previously presented to the network in adjacent time-steps.

Abbott and Blum (1996) also demonstrate shortcuts in sequences. In their model, the hippocampus gradually shortens its path through an interaction with the environment and successive shortenings. This result differs from our demonstrations, in that simulations using our model produce shortcuts without this gradual approach. Figure 13A contains a looping sequence of 40 patterns with some of the patterns occurring twice. After this sequence is presented to a network, the network is able to create a shorter sequence that omits the looping subsequence (Levy et al. 1995).

[Insert figure 13 near here]

### **Subsequence Disambiguation**

A general problem in sequence learning is subsequence disambiguation. This scenario requires that the network correctly recall two sequences that share a common subsequence. For example, in Fig. 13B sequences 1 and 2 share the common subsequence  $\alpha\beta\gamma$ . After training on these two sequences, the network can disambiguate if it recalls the sequence ABC $\alpha\beta\gamma$ GHIJKL when presented with pattern A and recalls the sequence OPQ $\alpha\beta\gamma$ UVWXYZ when presented with pattern O. The challenge for the network lies in the common subsequence. After recalling pattern  $\gamma$ , the next correct pattern depends on whether it was initially presented with pattern A or O. Such a task requires that the network store some sort of context corresponding to past experiences.

Fukushima (1973) demonstrates that a neurally inspired network could solve this problem using a non-local learning rule with a fully connected network. However, neither the rule nor the connectivity corresponds to the biological reality of the hippocampus.

Networks using our more biologically based model can also solve this problem using a one time-step spanning associative modification rule (Eq. (2)) (Minai et al. 1994; Levy et al. 1995; Wu et al. 1996). Because the required context in this example is more than one time-step in the past, the network must rely on the recurrently driven neurons to create the necessary context.

### **Goal Finding Without Search**

Another useful cognitive ability is to imagine the correct route to a specific goal before actually taking that path. Just as subsequence disambiguation requires using context based on past experiences, goal finding requires using context derived from a future goal. For example, in the looping path problem (Fig. 13A), the network might be given the goal of reaching pattern 21. Since the network would normally avoid the loop, it must alter its behavior based on this goal. But what is a reasonable representation of a goal in a neural system? Imagine the end patterns are places containing desirable objects. For example, pattern 21 might contain a sub-pattern representing food. Therefore, we might stimulate the network to find this goal by activating a portion of the food neurons – in the same way that a hungry rat might have its food neurons activated by the hunger-mediating portion of the brain. Using the looping path sequence, Levy et al. (1995) demonstrated that by firing four of the 512 neurons used to represent pattern 21, the network will enter the loop and reach that pattern.

### **Piecing Together Subsequences**

A problem related to goal finding and subsequence disambiguation is piecing together subsequences to create new sequences. For example, in Fig. 13B, one could create the sequence ABC $\alpha$  $\beta$  $\gamma$ UVWXYZ or the sequence OPQ $\alpha$  $\beta$  $\gamma$ GHIJKL. This problem is distinguished from subsequence disambiguation by positing that the network is presented with both an initial position (or pattern) and a specific goal to reach from that position. So, for example, the network is started at pattern A and given the goal of reaching pattern Z. Just as in goal finding, simulations using our model indicate the goal by providing external input corresponding to a subset of the neurons that code for pattern Z. When a strong enough goal input is given, for example four out of 512 neurons, the network is able to create the appropriate sequence (Levy et al. 1995; Levy 1996; Levy & Wu 1997a; Wu & Levy 1996; Polyn et al. 2000).

## T-maze

Another example of goal finding is the T-maze . One version of this problem requires the animal to learn to choose either the right arm or the left arm of a T-shaped maze upon reaching the intersection. There are two different, distinguishable goals. We simulate an animal learning this T-maze problem by training the network on two partially overlapping sequences: a “left” sequence where the simulation chooses the left arm of the maze, and a “right” sequence where the simulation chooses the right arm of the maze. The T-maze is divided into 14 segments, 6 segments for the stem of the “T”, and 4 segments for each arm. For each four-segment arm, the first two segments are orthogonal to the fourth goal arm. A neural firing pattern composed of 15%-20% externally driven neurons out of the total activity represents the sensory input of each segment. In simulations, each segment lasts three time-steps (approximately 100 ms). A training trial always begins by traversing the stem of the “T”, and then, depending on the simulation, a left or right turn is made. This turn continues to the goal of that arm. After a series of such left and right training trials, we test the simulation’s ability to select a left-turn or right-turn in order to reach either goal. The desired goal is defined by an induced attractor (i.e., activation of a subset of external goal neurons). We presume that some other brain region decides which goal is desired and accordingly turns on a subset of the neurons that are in just one of the goals; these “goal neurons” are externally driven on every time-step of the test trial. During a test trial, the stem sequence is presented but the arm sequence is not. A simulation has successfully learned the T-maze if, after the choice point (the intersection of the stem and the two arms), the network activates more of the neurons coding the nearby portion of the arm with the desired outcome than the neurons coding the nearby portion of the other arm. We have shown that the model successfully learns the T-maze problem, but learning can easily deteriorate with overtraining (Monaco and Levy 2003).

In sum, the model has the fundamental properties needed for cognitive mapping, or at least for two-thirds of Tolman's (1948) definition of cognitive mapping. That is, we demonstrated that the network could learn a sequence (Levy et al. 1995; Wu & Levy 1996), find shortcuts, learn overlapping sequences, solve the disambiguation problem for these two sequences, and create novel sequences by piecing together different but overlapping parts of previously learned sequences in the goal finding problem. (See Fig. 13 and Levy (1996) for more details.)

### **Non-spatial Paradigms**

At this point, we turn our attention to particular cognitive paradigms that are known to be hippocampal but are explicitly non-spatial in nature. These paradigms are very well quantified and seem to bear little relationship to each other or to cognitive mapping. However, because the model can solve all of them, we can see the relationship in a computational sense among these disparate problems. There are two prominent non-spatial cognitive learning paradigms used in animal and learning experiments, transverse patterning (TP) and transitive inference (TI).

### **Configural Learning**

The configural learning paradigms which have been studied include transverse patterning (TP), transverse non-patterning (TNP), and transitive inference (TI). Both TP and TI are dependent on normal hippocampal function. We conjecture that this is also true for TNP when properly trained (Wu & Levy 2002). The TP and TNP problems have been studied in Rudy's laboratory (Alvarado & Rudy 1992). The TP task requires a choice from among three stimuli, A, B, and C, which are presented to the network as simultaneous pairs ( $\{A,B\}$ ,  $\{B,C\}$ , or  $\{C,A\}$ ). In TP experiments, subjects (human, animal, or a simulated neural network) are trained to choose the correct stimulus in each pair. We refer to trials for each of these three stimulus pairs as a subtask. Figure 14 shows in

terms of neural firing patterns how our simulations of our model learn the subtasks. For more details, see figures 3, 4, and 5 in Shon et al. (2002).

[Insert figure 14 near here]

In TP there is a perfect balance in the ambiguity of the individual stimulus with respect to the correct decision: stimulus A is correct for the pair {A,B}, stimulus B is correct for the pair {B,C}, and stimulus C is correct for the pair {C,A}. Thus, TP is noted as A+B-, B+C-, and C+A-. Because of the symmetry of right and wrong, the correct decision can only be made by using the contextual knowledge of the stimulus pair itself.

Transverse non-patterning (TNP) is similar to TP, except that the {B,C} pair is replaced with {D,C}. Thus, TNP can be represented as A+B-, D+C-, and C+A-. Because there is less balance of ambiguity (B is always wrong and D is always right), one might imagine that TNP would be easier to learn than TP. However, as shown by our simulations (Wu & Levy 2002) and experiments using rats (Alvarado & Rudy 1992), TNP is, on average, unlearnable when training uses a progressive learning paradigm.

The transitive inference (TI) problem is regarded as a logical problem in cognitive psychology. Transitive inference is based on a transitive relationship such as "greater than", and as such, is said to develop if subjects can infer from A is greater than B and B is greater than C, that A is greater than C. In psychological experiments, at least four pairs of five atomic items are used. After learning which atomic stimulus (A, B, C, or D) is the correct choice in each of these four pairs (A>B, B>C, C>D, and D>E), the subject gets the novel BD combination in which B should be the correct choice. Here the BD test is critical, but one can also test on the AE pair. By casting the TI problem as a sequence-learning problem, we have shown that a simplified hippocampal model solves the TI problem (Levy & Wu 1997b, 2000; Wu & Levy 1998), which was studied in some detail relative to network parameters at the end of Smith et al. (2000).

## Cognitive Difficulties

The TNP problem is interesting because Alvarado and Rudy report it as unlearnable. Using a progressive training paradigm (Alvarado & Rudy 1992), the experimental data indicate that the TP problem is learnable but the TNP problem is not. Simulations agree and predict that the TNP problem is unlearnable when training is as presented by Alvarado and Rudy. However, simulations predict that TNP is learnable using a different training paradigm (unpublished observations).

[Insert figure 15 near here]

Figure 15 compares the behavioral and simulated learning curves for the TP problem using the published training schedule (Alvarado & Rudy 1992, 1995). Here the model was parameterized to fit the learning rates and asymptotic performance. That is, multiple simulations were run with the same parameters while differing only in initial connectivity. The data are reported in blocks of 30 trials to correspond to the reported behavioral data. There were three phases of training (I, II, and III) corresponding to blocks 1-4, 5-7, and 8-10, respectively. A new problem set was added at the beginning of each phase. Thus, problem A+B- was always part of training; problem B+C- was present in Phases II and III; and problem C+A- was present only in Phase III. Note that performance changes across training sessions, and that there is reasonable agreement between our simulation results and the behavioral data. More importantly, the  $\chi^2$  tests of the histograms of SEM's show that our simulations are no more different from each rodent experiment than those two experiments are from each other (the average  $\chi^2$  of the simulation versus the two replicates is 4.21 while the  $\chi^2$  comparing the two behavioral experiments is 5.04;  $df = 4$  in both cases). In both the simulations and the behavioral data, performance on problem A+B- rose to approximately the same level in Phase I. In Phase II, the simulations slightly outperform the experimental results, but the two replicates of the

behavioral experiments are far enough apart that there is no reason to demand an overly exact fit by the simulations at this point. The introduction of the C+A- problem in Phase III brings the simulations and behavioral data very close to one another at the end of training.

Finally using the same parameterization of the network that produces the TP learning curves, we reinitialize the simulations and train on the TI task. Without parameter adjustment, we again reproduced the published learned performance in the appropriate number of trials (Fig. 15; lower right corner).

### **Trace Conditioning**

Trace conditioning was devised by Pavlov. A subject is given a brief stimulus called the conditioned stimulus (CS), followed by an interval of no stimulus (the trace interval). Finally, at the end of the trace interval comes the unconditioned stimulus (UCS or US). The unconditioned stimulus elicits an unconditioned response (UCR, Fig. 16).

Eventually, if the trace interval is not too long, the subject learns to anticipate the UCS by generating a conditioned response (CR) at an appropriate time. For more details, see Solomon et al. (1986). What is of interest here is the compressed encoding of sequence CS → trace interval → UCS.

[Insert figure 16 near here.]

If the UCS is unpleasant and the UCR is a means of mitigating the unpleasantness, then the trace conditioning task is an escape paradigm. The prototypical example of this is trace eyeblink conditioning, wherein a neutral CS (such as a tone) is paired with an air puff delivered to the eye, eliciting a blink. Solomon et al. (1986) demonstrate the critical importance of the hippocampus for learning the trace interval when an escape paradigm is used. In such tasks, the hippocampus forecasts the UCS (unconditioned stimulus). Presumably, by virtue of reciprocal connections with UCS-activated neurons in cerebral cortex, the hippocampus activates these neurons prior to the activated occurrence of the

UCS. Such activation can lead to a blink by virtue of these neurons' presumed relationship with the blink controlling neurons in motor cortex (although more direct connections with trochlear or cerebellar systems remains an open question).

The CS and UCS are represented in the model as patterns presented at specific times to a network simulation during each training trial (Levy & Sederberg 1997, Rodriguez & Levy 2001). If during testing, when presented with only the CS, the network successfully anticipates the UCS at the appropriate time, then we say that the network successfully acquired trace conditioning. For very short or very long trace intervals, a network simulation fails to anticipate the UCS. For intermediate trace intervals, almost all simulations successfully anticipate the UCS, and for certain boundary intervals, a smaller percentage of the networks successfully anticipate the UCS. The curve describing the successful learning as a function of trace duration, Fig. 17, closely parallels the work reported by Gormezano et al. (1983) involving rabbits. A quantified time scale is produced by assuming the off-rate time constant of the NMDA-receptor is 100 ms ( $\alpha$  of Eq. (4)). Also matching animal experiments is the sudden onset of the acquired behavior (see Fig. 8 of Rodriguez & Levy 2001).

[Insert figure 17 near here]

The neuron types after acquisition of trace conditioning also parallel the reported type in McEchron and Disterhoft (1997; compare to Table 2 of Rodriguez & Levy, 2001). Contrary to Pavlov's prediction, networks that acquire trace conditioning do not actually use a CS memory trace to anticipate the UCS at the appropriate time. Rather, a random bridge through state space allows a timely activation of a UCS representation (Fig. 8).

## **Acknowledgments**

This work was supported by NIH MH63855 to WBL, the Meade Munster Foundation, and by the Department of Neurosurgery. The authors thank Janine M. Vallee for her assistance.

## References

Abbott, L.F. & Blum, K.I. (1996). Functional significance of long-term potentiation for sequence learning and prediction. *Cerebral Cortex*, 6, 406-416.

Alvarado, M.C. & Rudy, J.W. (1992). Some properties of configural learning: An investigation of the transverse-patterning problem. *Journal of Experimental Psychology*, 18, 145-153.

Alvarado, M.C. & Rudy, J.W. (1995). Rats with damage to the hippocampal-formation are impaired on the transverse-patterning problem but not on elemental discriminations. *Behavioral Neuroscience*, 109, 204-211.

Amarasingham, A. & Levy, W.B (1998). Predicting the distribution of synaptic strengths and cell firing correlations in a self-organizing, sequence prediction model. *Neural Computation*, 10, 25-57.

August, D.A. & Levy, W.B (1996). Temporal sequence compression by a hippocampal network model. *INNS World Congress on Neural Networks*, 1299-1304.

August, D.A. & Levy, W.B (1997). Spontaneous replay of temporally compressed sequences by a hippocampal network model. In: J.M. Bower (Ed.), *Computational Neuroscience: Trends in Research*, 1997 (pp. 231-236). New York: Plenum Press.

August, D.A. & Levy, W.B (1999). Temporal sequence compression by an integrate-and-fire model of hippocampal area CA3. *Journal of Computational Neuroscience*, 6, 71-90.

Barlow, H.B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In: *Mechanisation of Thought Processes*. London: Her Majesty's Stationery Office.

Buzsaki, G. (1996). The hippocampo-neocortical dialogue. *Cerebral Cortex*, 6, 81-92.

Buzsaki, G., Horvath, Z., Urioste, R., Hetke, J. & Wise, K. (1992). High-frequency network oscillation in the hippocampus. *Science*, 256, 1025-1027.

Chun, M.M. & Phelps, E.A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature Neuroscience*, 2, 844-847.

Cohen, N.J. & Eichenbaum, H. (1993). *Memory, Amnesia, and the Hippocampal System*. Cambridge, MA: MIT Press.

Cohen, N.J. & Squire, L.R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science*, 210, 207-210.

Dretske, F.I. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dusek, J.A. & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences, USA* 94:7109-7114.

Fukushima K. (1973). A model of associative memory in the brain. *Kybernetik*, 12, 58-63.

Gomezano, I., Kehoe, E.J., & Marshall, B.S. (1983). Twenty years of classical conditioning research with the rabbit. *Progress Psychobiology & Physiological Psychology*, 10, 197-267.

Green, J.D. & Arduini, A.A. (1954). Hippocampal electrical activity in arousal. *Journal of Neurophysiology*, 17, 531-557.

Greene, A.J., Prepscius, C., & Levy, W.B (2000). Primacy versus recency in a quantitative model: Activity is the critical distinction. *Learning & Memory*, 7, 48-57.

Greene, A.J., Spellman, B.A., Dusek, J.A., Eichenbaum, H.B., & Levy, W.B (2001). Relational learning with and without awareness: Transitive inference using non-verbal stimuli in humans. *Memory & Cognition*, 29, 893-902.

Hasselmo, M., Bodelon, C., & Wyble, B. (2002). A proposed function of hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computing*, 14, 793-817.

Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory. *Behavioral Biology*, 12, 424-444.

Hocking, A.B & Levy, W.B (2005). Computing conditional probabilities in a minimal CA3 pyramidal neuron. *Neurocomputing*, 65-66, 297-303.

Holmes, W.R. & Levy, W.B (1990). Insights into associative long-term potentiation from computational models of NMDA receptor-mediated calcium influx and intracellular calcium concentration changes. *Journal of Neurophysiology*, 63, 1148-1168.

Jaynes, E.T. (1979). 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.

Kesner, R.P. & Hardy, J.D. (1983). Long-term memory for contextual attributes: dissociation of amygdala and hippocampus. *Behavioral Brain Research*, 8, 139-149.

Levy, W.B (1985). An information/computation theory of hippocampal function. *Society for Neuroscience Abstract*, 11, 493.

Levy, W.B (1989). A computational approach to hippocampal function. In: R.D. Hawkins & G.H. Bower (Eds.), *Computational Models of Learning in Simple Neural Systems* (pp. 243-305). New York: Academic Press.

Levy, W.B (1990a). Hippocampal theories and the information/computation perspective. IN: L. Erinoff (Ed.), *NIDA Monographs; Neurobiology of Drug Abuse: Learning and Memory* (pp. 116-125). Rockville, MD: U. S. Dept. of Health and Human Services, National Institute of Drug Abuse.

Levy, W.B (1990b). Maximum entropy prediction in neural networks. *International Joint Conference on Neural Networks*, 1-7—1-10.

Levy, W.B (1994). Unification of hippocampal function via computational considerations. INNS World Congress on Neural Networks, IV-661-666.

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

Levy, W.B & Baxter, R.A. (2002). Energy-efficient neuronal computation via quantal synaptic failures. *Journal of Neuroscience*, 22, 4746-4755.

Levy, W.B, Colbert, C.M., & Desmond, N.L (1990). Elemental adaptive processes of neurons and synapses: A statistical/computational perspective. In: M.A. Gluck & D.E. Rumelhart (Eds.), *Neuroscience and Connectionist Models* (pp. 187-235). Hillsdale, NJ: Lawrence Erlbaum Assoc., Inc.

Levy, W.B, Sanyal, A., Rodriguez, P., Sullivan, D.W., & Wu, X.B. (2005). The formation of neural codes in the hippocampus: trace conditioning as a prototypical paradigm for studying the random recoding hypothesis. *Biological Cybernetics*, 92, 409-426.

Levy, W.B & Sederberg, P.B. (1997). A neural network model of hippocampally mediated trace conditioning. *IEEE International Conference on Neural Networks*, I-372-376.

Levy, W.B, Sederberg, P.B., & August, D.A. (1998). Sequence compression by a hippocampal model: A functional dissection. In: J.M. Bower (Ed.), *Computational Neuroscience: Trends in Research, 1998* (pp. 435-439). New York: Plenum Press.

Levy, W.B & Steward, O. (1979). Synapses as associative memory elements in the hippocampal formation. *Brain Research*, 175, 233-245.

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

Levy, W.B & Wu, X.B. (1996). The relationship of local context codes to sequence length memory capacity. *Network*, 7, 371-384.

Levy, W.B & Wu, X.B. (1997a). Predicting novel paths to goals by a simple, biologically inspired neural network. In: J.M. Bower (Ed.), *Computational Neuroscience: Trends in Research*, 1997 (pp. 705-709). New York:Plenum Press.

Levy, W.B & Wu, X.B. (1997b). A simple, biologically motivated neural network solves the transitive inference problem. *IEEE International Conference on Neural Networks*, I-368-371.

Levy, W.B. & Wu, X.B (2000). Some randomness benefits a model of hippocampal function. In: . H. Liljenstrom, P. Arhem, & C. Blomberg (Eds.), *Disorder versus Order in Brain Function* (pp. 221-237). Singapore: World Scientific Publishing.

Levy, W.B & Wu, X. (2005). External Activity and the Freedom to Recode. *Neurocomputing*, 2005, submitted.

Levy, W.B, Wu, X., & Baxter R.A. (1995). Unification of hippocampal function via computational/encoding considerations. In: D.J. Amit, P. del Giudice, B. Denby, E.T. Rolls & A. Treves (Eds.), *Proceedings of the Third Workshop on Neural Networks: from Biology to High Energy Physics*. Singapore: World Scientific Publishing Intl. *J. Neural Sys.*, 6, (Supp.), pp. 71-80.

Levy, W.B., Wu, X.B., Greene, A.J., & Spellman, B.A. (2003). A source of individual variation. *Neurocomputing*, 52-54, 165-168.

Levy, W.B, Wu, X.B. & Tyrcha, J.M. (1996). Solving the transverse patterning problem by learning context present: A special role for input codes. *INNS World Congress on Neural Networks*, 1305-1309.

McClelland, J.L., McNaughton, B.L., & O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102, 419-437.

McEchron, M.D. & Disterhoft, J.F. (1997). Sequence of single neuron changes in CA1 hippocampus of rabbits during acquisition of trace eyeblink conditioned responses. *Journal of Neurophysiology*, 78, 1030-1044.

Mehta, M.R., Barnes, C.A., & McNaughton, B.L. (1997). Experience-dependent, asymmetric expansion of hippocampal place fields. *Proceedings of the National Academy of Science USA*, 94, 8918-8921.

Miles, R. & Wong, R.K. (1986). Excitatory synaptic interactions between CA3 neurones in the guinea-pig hippocampus. *Journal of Physiology*, 373, 397-418.

Miller, G. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.

Milner, B. (1972). Disorders of learning and memory after temporal lobe lesions in man. *Clinical Neurosurgery*, 19, 421-446.

Minai, A.A., Barrows, G.L., & Levy, W.B (1994). Disambiguation of pattern sequences with recurrent networks. *INNS World Congress on Neural Networks*, IV-176-181.

Minai, A.A. & Levy, W.B (1993a). The dynamics of sparse random networks. *Biological Cybernetics*, 70, 177-187.

Minai, A.A. & Levy, W.B (1993b). Sequence learning in a single trial. *INNS World Congress on Neural Networks*, II-505-508.

Minai, A.A. & Levy, W.B (1993c). Predicting complex behavior in sparse asymmetric networks. *Neural Information Processing Systems 5: Natural and Synthetic*, 556-563.

Minai, A.A. & Levy, W.B (1994). Setting the activity level in sparse random networks. *Neural Computation*, 6, 85-99.

Mitman, K.E, Laurent, P.A., & Levy W.B (2003). Defining time in a minimal hippocampal CA3 model by matching time-span of associative synaptic modification and

input pattern duration. International Joint Conference on Neural Networks (IJCNN) 2003 Proceedings, 1631-1636.

Molyneaux, B. & Hasselmo, M. (2002). GABA<sub>B</sub> Presynaptic Inhibition Has an In Vivo Time Constant Sufficiently Rapid to Allow Modulation at Theta Frequency. *Journal of Neurophysiology*, 87, 1196-1205.

Monaco, J.D. & Levy, W.B (2003). T-maze training of a recurrent CA3 model reveals the necessity of novelty-based modulation of LTP in hippocampal region CA3. International Joint Conference on Neural Networks (IJCNN) 2003 Proceedings, 1655-1660.

Nadel, L. & Willner, J. (1980). Context and conditioning: A place for space. *Physiological Psychology*, 8, 218-228.

O'Keefe, J. & Nadel L. (1978). *The Hippocampus as a Cognitive Map*. Oxford:Oxford University Press.

Panzeri, S., Rolls, E. T., Battaglia, F. and Lavis, R. (2001). Speed of feedforward and recurrent processing in multilayer networks of integrate-and-fire neurons. *Network: Computation in Neural Systems* 12: 423-440.

Pavlidis, C. & Winson J. (1989). Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *Journal of Neuroscience*, 9, 2901-2918.

Polyn, S., Wu, X.B., & Levy, W.B (2000). Entorhinal/dentate excitation of CA3: A critical variable in hippocampal models. *Neurocomputing*, 32-33, 493-499.

Polyn, S. & Levy, W.B (2001). Dynamic control of inhibition improves performance of a hippocampal model. *Neurocomputing*, 38-40, 823-829.

Prepscius, C. & Levy, W.B (1994). Sequence prediction and cognitive mapping by a biologically plausible neural network. *INNS World Congress on Neural Networks*, IV-164-169.

Rodriguez, P. & Levy, W.B (2001). A model of hippocampal activity in trace conditioning: Where's the trace? *Behavioral Neuroscience*, 115, 1224-1238.

Rolls, E.T. & Treves, A. (1998). *Neural networks and brain function*. Oxford: Oxford University Press.

Rolls, E.T., Treves, A., Foster, D. & Perez-Vicente, C. (1997). Simulation studies of the CA3 hippocampal subfield modelled as an attractor neural network. *Neural Networks*, 1559-1569.

Rudy, J.W. & Sutherland, R.J. (1995). Configural association theory and the hippocampal-formation: An appraisal and reconfiguration. *Hippocampus*, 5, 375-389.

Schmajuk, N. (2002). *Latent inhibition and its neural substrates*. Boston, MA: Kluwer Academic Publishers.

Shon, A.P., Wu, X.B., & Levy, W.B (2000). Using computational simulations to discover optimal training paradigms. *Neurocomputing*, 32-33, 995-1002.

Shon, A.P., Wu, X.B., Sullivan, D.W., & Levy, W.B (2002). Initial state randomness improves sequence learning in a model hippocampal network. *Physical Review E*, 65, 031914/1-15.

Skaggs, W.E. & McNaughton, B.L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science*, 271, 1870-1873.

Skaggs, W.E., McNaughton, B.L., Wilson, M.A. & Barnes, C.A. (1996). Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus*, 6, 149-172.

Smith, A.C., Wu, X.B., & Levy, W.B (2000). Controlling activity fluctuations in large, sparsely connected random networks. *Network*, 11, 63-81.

Solomon, P.R., Vander Schaaf, E R., Thompson, R.F., & Weisz, D.J. (1986). Hippocampus and trace conditioning of the rabbit's classically conditioned nictitating membrane response. *Behavioral Neuroscience*, 100, 729-744.

Squire, L.R. (1987). *Memory and Brain*. New York, NY: Oxford University Press.

Stevens, C.F. & Wang, Y. (1994). Changes in reliability of synaptic function as a mechanism for plasticity. *Nature*, 371, 704-707.

Stringer, S.M., Rolls, E.T., & Trappenberg, T.P. (2004). Self-organising continuous attractor networks with multiple activity packets, and the representation of space. *Neural Networks*, 17, 5-27.

Sullivan, D.W. & Levy, W.B (2003a). Quantal synaptic failures improve performance in a sequence learning model of hippocampal CA3. *Neurocomputing*, 52-54, 397-401.

Sullivan, D.W. & Levy, W.B (2003b). Synaptic modification of interneuron afferents in a hippocampal CA3 model prevents activity oscillations. *International Joint Conference on Neural Networks (IJCNN) 2003 Proceedings*, 1625-1630.

Sullivan, D.W. & Levy, W.B (2004). Quantal synaptic failures enhance performance in a minimal hippocampal model. *Network*, 15, 45-67.

Swanson, L.W. & Köhler, C. (1986). Anatomical evidence for direct projections from the entorhinal area to the entire cortical mantle in the rat. *Journal of Neuroscience*, 6, 3010-3023.

Swanson, L.W., Köhler, C. & Björklund, A. (1987) The limbic region. I: The septohippocampal system. In: A. Björklund, T. Hökfelt, & L.W. Swanson, (Eds.), *Handbook of Chemical Neuroanatomy*, Vol. 5. (pp. 125-277). Amsterdam: Elsevier Science Publishers.

Thompson, L.T. & Best, P.J. (1989). Place cells and silent cells in the hippocampus of freely-behaving rats. *Journal of Neuroscience*, 9, 2382-2390.

Thomson, A.M. (2000). Facilitation, augmentation and potentiation at central synapses. *Trends in Neurosciences*, 23, 305-312.

Tolman, E.C. (1948). Cognitive maps in rats and men. *Psychological Review*, 42, 189-208.

Treves, A. (1993). Mean-field analysis of neuronal spike dynamics. *Network* 4: 259-284.

Treves, A. (2004). Computational constraints between retrieving the past and predicting the future, and the CA3-CA1 differentiation. *Hippocampus*, 14, 539-556.

Treves, A., Rolls, E. T. and Simmen, M. (1997). Time for retrieval in recurrent associative memories. *Physica D* 107: 392-400.

Tsodyks, M.V., Skaggs, W.E., Sejnowski, T.J. & McNaughton, B.L. (1996). Population dynamics and theta rhythm phase precession of hippocampal place cell firing: a spiking neuron model. *Hippocampus*, 6, 271-280.

Watanabe, S. (1961). A note on the formation of concept and of association by information theoretical correlation analysis. *Information and Control*, 4(2-3), 291-296.

Wilson, M.A. & McNaughton, B.L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265, 676-679.

Wu, X.B., Baxter, R.A. & Levy, W.B (1996). Context codes and the effect of noisy learning on a simplified hippocampal CA3 model. *Biological Cybernetics*, 74, 159-165.

Wu, X.B. & Levy, W.B (1996). Goal finding in a simple, biologically inspired neural network. INNS World Congress on Neural Networks, 1279-1282.

Wu, X.B. & Levy, W.B (1998). A hippocampal-like neural network model solves the transitive inference problem. In: J.M. Bower (Ed.), *Computational Neuroscience: Trends in Research*, 1998 (pp. 567-572). New York, NY: Plenum Press.

Wu, X.B. & Levy, W.B (1999). Enhancing the performance of a hippocampal model by increasing variability early in learning. *Neurocomputing*, 26-27, 601-607.

Wu, X.B. & Levy, W.B (2002). Simulating the transverse non-patterning problem. *Neurocomputing*, 44-46, 1029-1034.

Wu, X.B. & Levy, W.B (2005). Increasing CS and US longevity increases the learnable trace interval. *Neurocomputing*, 65-66, 283-289.

Wu, X.B., Tyrcha, J., & Levy, W.B (1998). A neural network solution to the transverse-patterning problem depends on repetition of the input code. *Biological Cybernetics*, 79, 203-213.

**Table 1****A Minimal Hippocampal CA3 Model**

1. Neurons are threshold elements with inputs that are weighted and summed; the output is binary, a spike when threshold is exceeded and no spike otherwise.
2. Most connections are excitatory.
3. Synapses modify associatively based on a local Hebbian rule that is time-spanning between pre- and postsynaptic activations and includes LTP and LTD-like processes.
4. Recurrent excitation is sparse and randomly connected.
5. Recurrent excitation is stronger than external excitation.
6. One or more randomization processes exist.
7. Inhibitory neurons control activity, approximately.
8. Activity is low but not too low.

<b>Table 2</b> Cognitive, Behavioral, and Cellular Predictions and Explanations by the Model		
<u>Spatial Tasks</u>	<u>Configural Tasks</u>	<u>Trace Conditioning</u>
	<b>Paradigms</b>	
Simple sequence completion (various)	Transverse Patterning (8,11,12)	Trace Conditioning (15,16)
One trial learning (1,18)	Transverse Non-Patterning (NP1) (13)	
Jump ahead recall (2,3)	Transitive Inference (14,23)	
Circular sequence completion (4,5)		
Sequence Disambiguation (6,7,8)		
Shortcut finding (6)		
Goal finding (6)		
Combining appropriate subsequences (9,10)		
<b>Demonstrations and Observations</b>		
Simple sequence completion (1,17) 1. Memory capacity (9) 2. One trial learning (1)	Transverse Patterning 1. Learning paradigms (Concurrent, staged, progressive – 19) 2. Learning rates (12)	Maximum learnable trace interval (16, 22)
	NP1 learning rates (13)	Number of trials required to learn (16)
Circular sequence completion 1. Off-line Compression (2,4,5) 2. On-line Compression vs. noise, context length (8) 2. Off-line Spontaneous replay (3,4,5)	Transitive Inference population variability (20)	Different classes of neurons that bridge the trace interval (16)
Cell firing ahead of place (15,16, 22,24,25 See also on-line compression)		Jump in performance across training (16)
<b>Predictions</b>		
on-line T-maze choice point decision (21)	Transverse Non-Patterning (NP2) (unpublished observations)	Lack of stimulus encoding neurons during trace interval (16)
	Relative neuronal codes (19)	Increasing CS/US longevity increases learnable trace interval (22)

1. Minai and Levy, 1993b; 2. August and Levy, 1996; 3. Prepisci and Levy, 1994; 4. Levy, Sederberg and August, 1998; 5. August and Levy, 1999; 6. Levy, Wu, And Baxter, 1995; 7. Minai, Barrows, and Levy, 1994; 8. Wu, Baxter, and Levy 1996; 9. Levy and Wu, 1996; 10. Wu and Levy, 1996; 11. Levy, Wu, and Tyrcha, 1996; 12. Wu, Tyrcha and Levy, 1998; 13. Wu and Levy, 2002; 14. Wu and Levy, 1998; 15. Levy and Sederberg, 1997; 16. Rodriguez and Levy, 2001; 17. Amarasingham and Levy 1998; 18. Greene et al. 2000; 19. Shon, Wu and Levy 2000; 20. Levy et al. 2003; 21. Monaco and Levy 2003; 22. Wu and Levy 2005; 23. Smith, Wu and Levy 2000; 24. Mitman et al. 2003; 25. Levy et al. 2005.

**Table 3** Randomization Supports Recoding

<i>Cognitive Paradigms</i>	<i>Manipulation of Randomization</i>	<i>Results Relative to Learning Cognitive Task</i>	<i>Citations</i>
TI	frequency of chaotic oscillations (variance constant)	high frequency outperforms low frequency	Smith et al. 2000
TI	free-running with Z(0) versus competitive	competitive model has no activity oscillations and fails	Levy & Wu 2000
TP	size (length) of fully randomized Z(0)	no Z(0) fails and longer Z(0) is better up to $\ Z(0)\  = a \cdot n$	Wu & Levy 1999
TP	size of fully randomized Z(0)	performance correlates ( $r = 0.85$ ) with before training sensitivity to initial conditions	Wu & Levy 1999
TP	fraction of Z(0) randomized	no Z(0) randomization fails and greater randomization is better	Shon et al. 2002
TP	synaptic failures	inverse monotonic relationship between failure rate and activity	Sullivan & Levy 2003a, 2004
Disambiguation & Goal Finding	size of $m_e$	~35% is best	Polyn & Levy 2001
TI	size of $m_e$	~35% is best	Levy & Wu 2005

Figure 1.

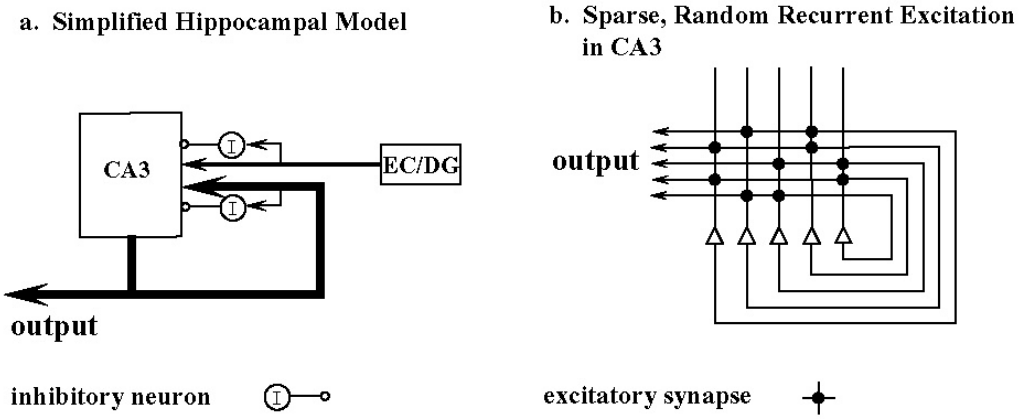
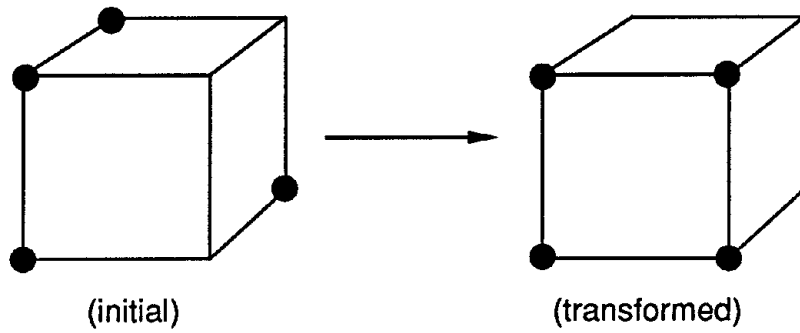


Figure 2.

A. Representations in a geometric space.



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

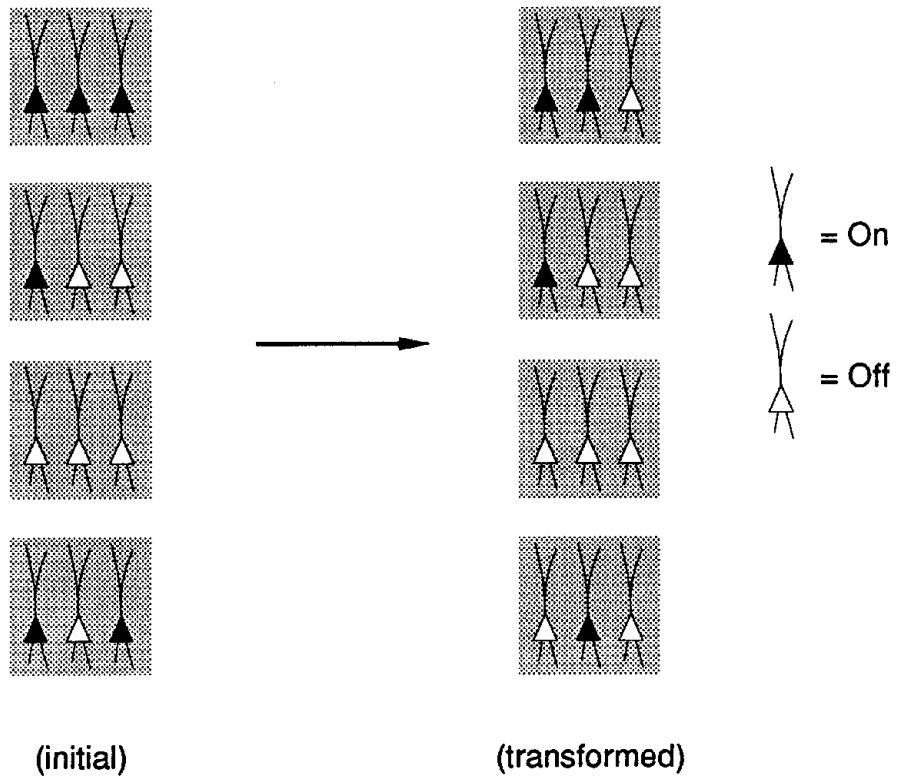


Figure 3.

Approximate Stability of a Sequence of CA3 Firing Patterns:  
How one tier of neurons (22 are illustrated) could change over time and  
achieve a highly similar but different series of representations.

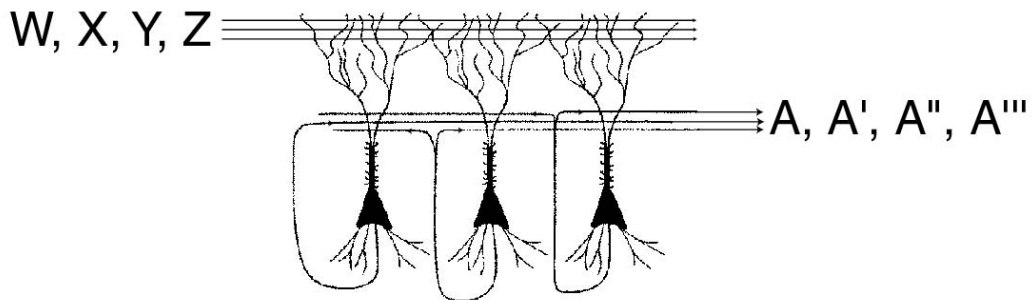
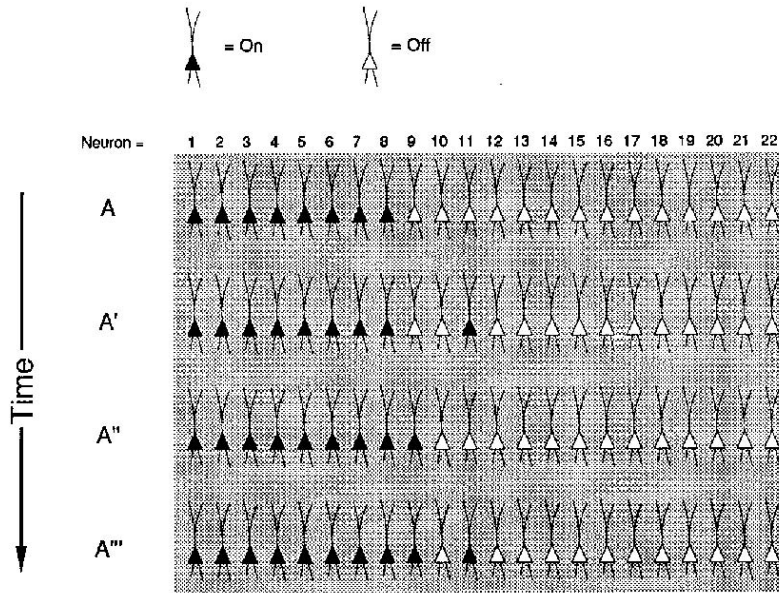


Figure 4.

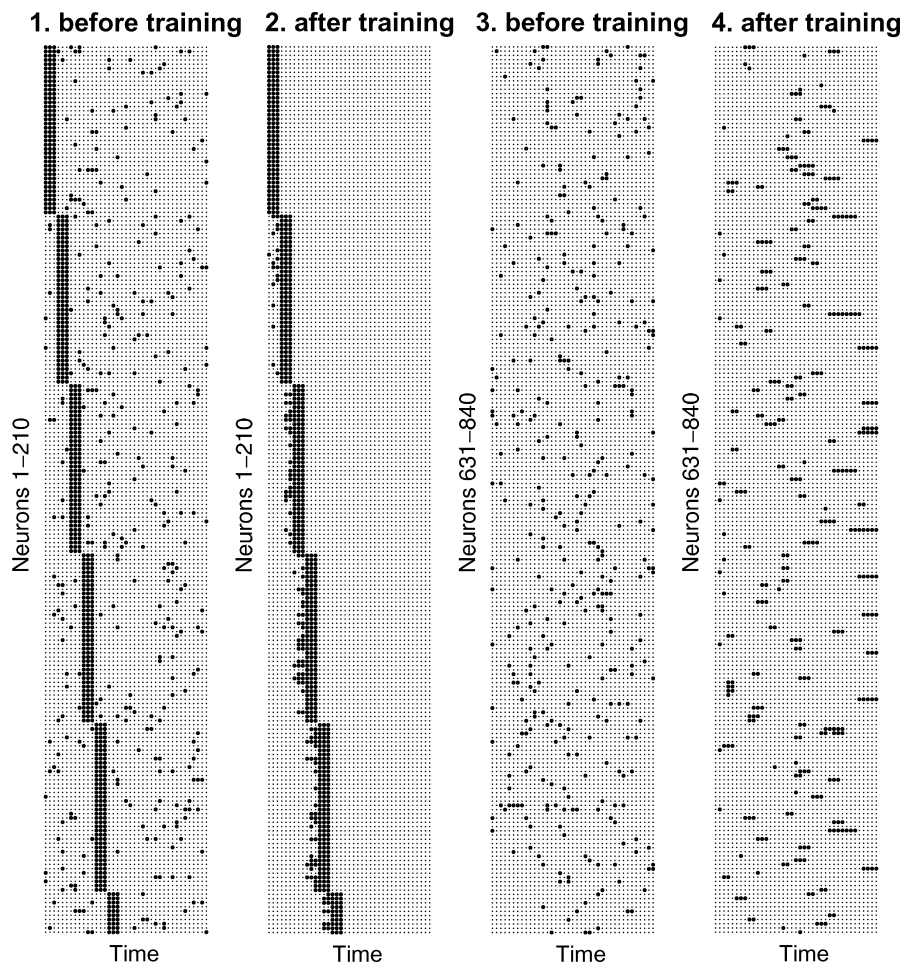


Figure 5.

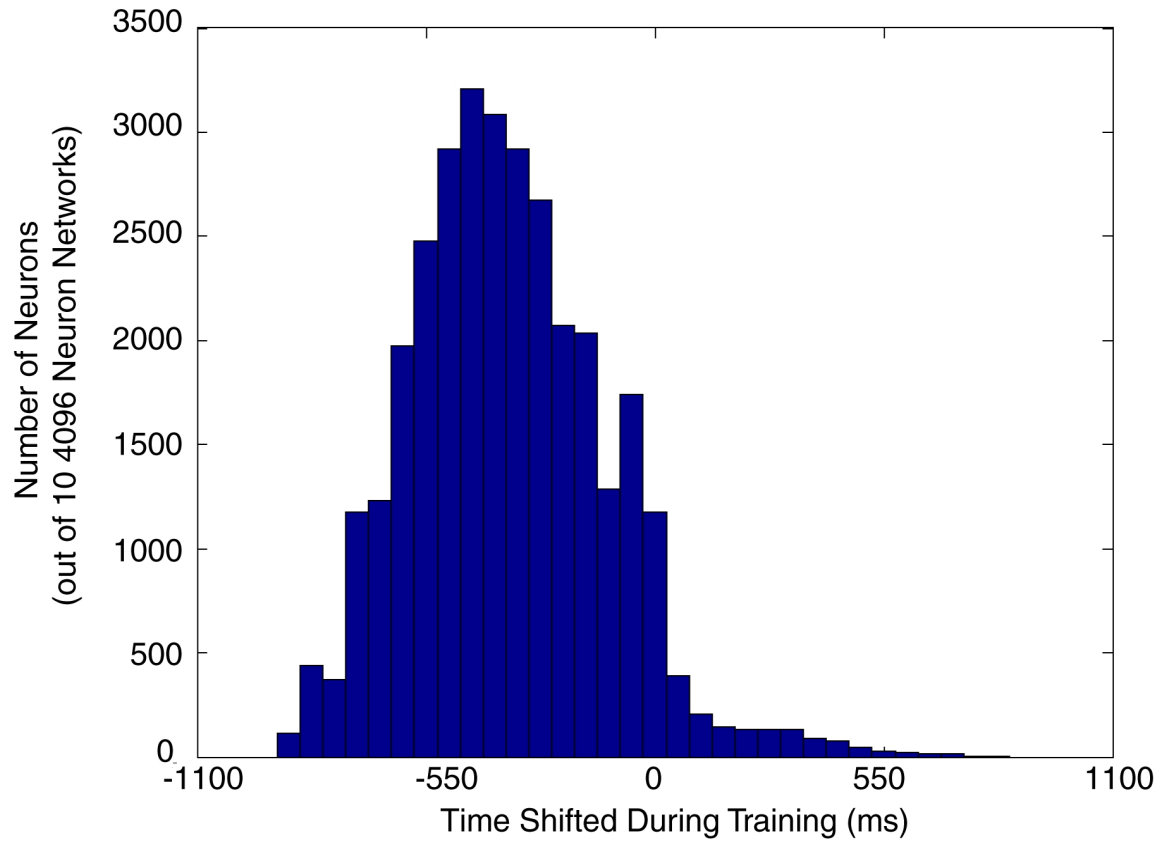


Figure 6.

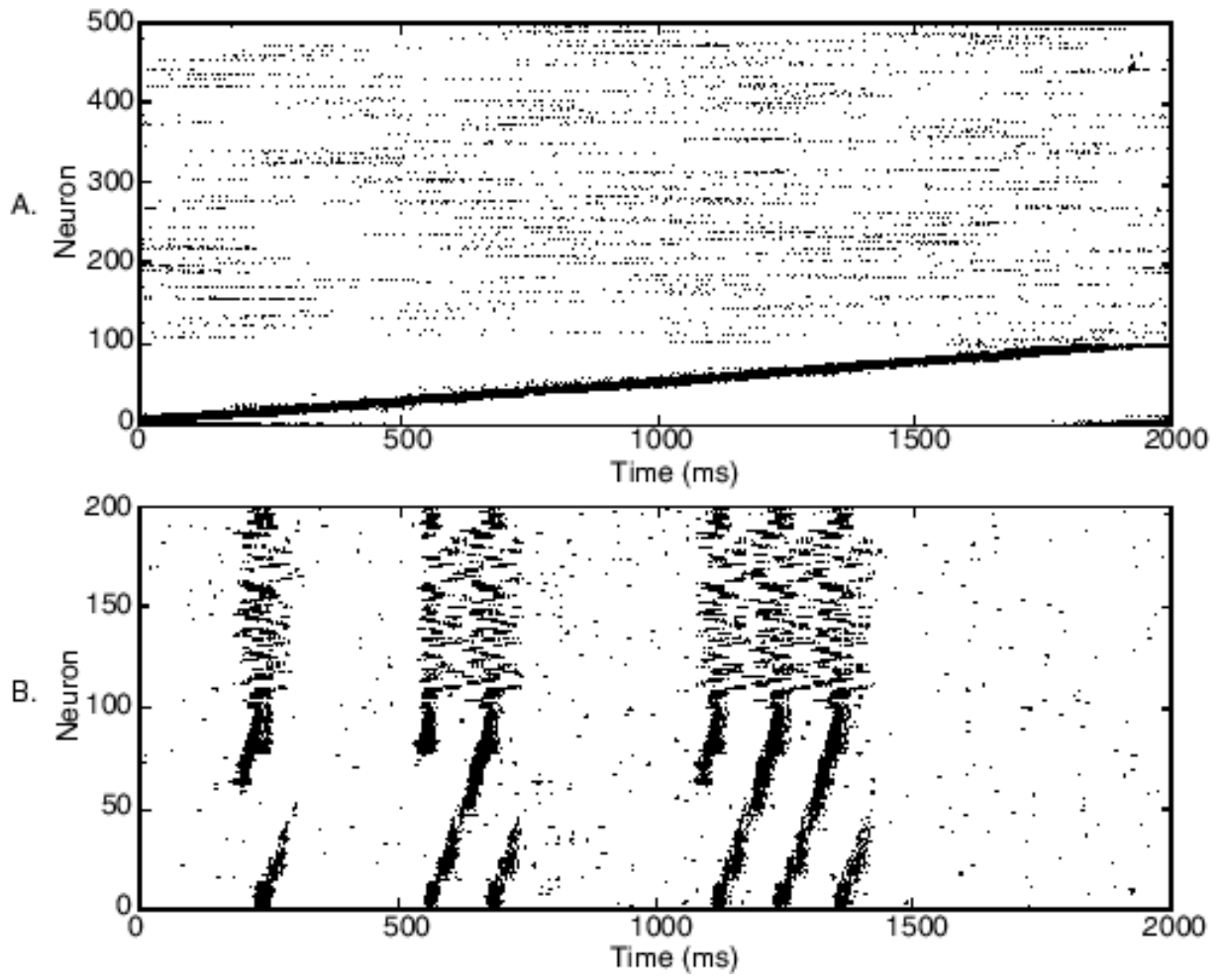


Figure 7.

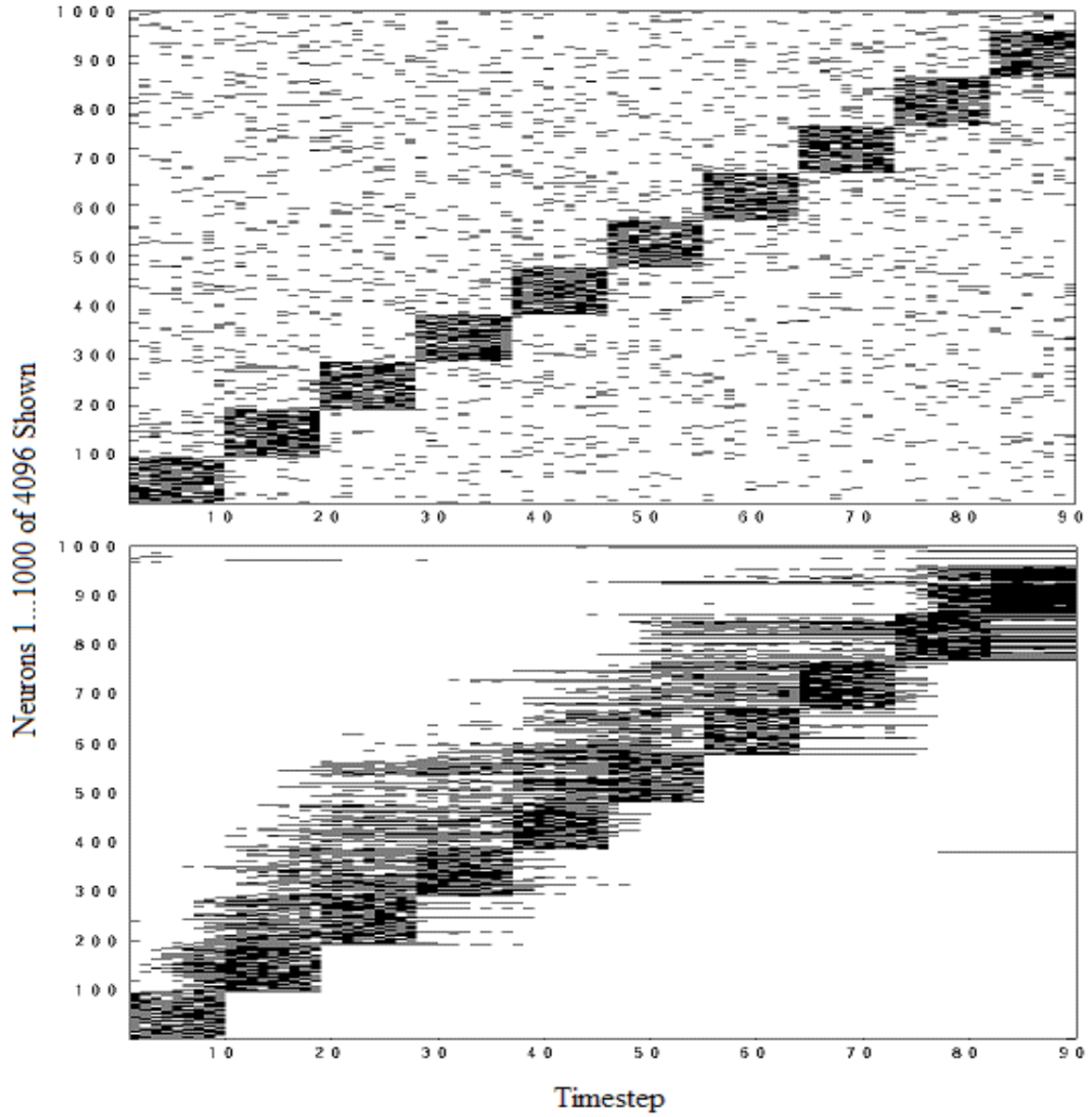


Figure 8.

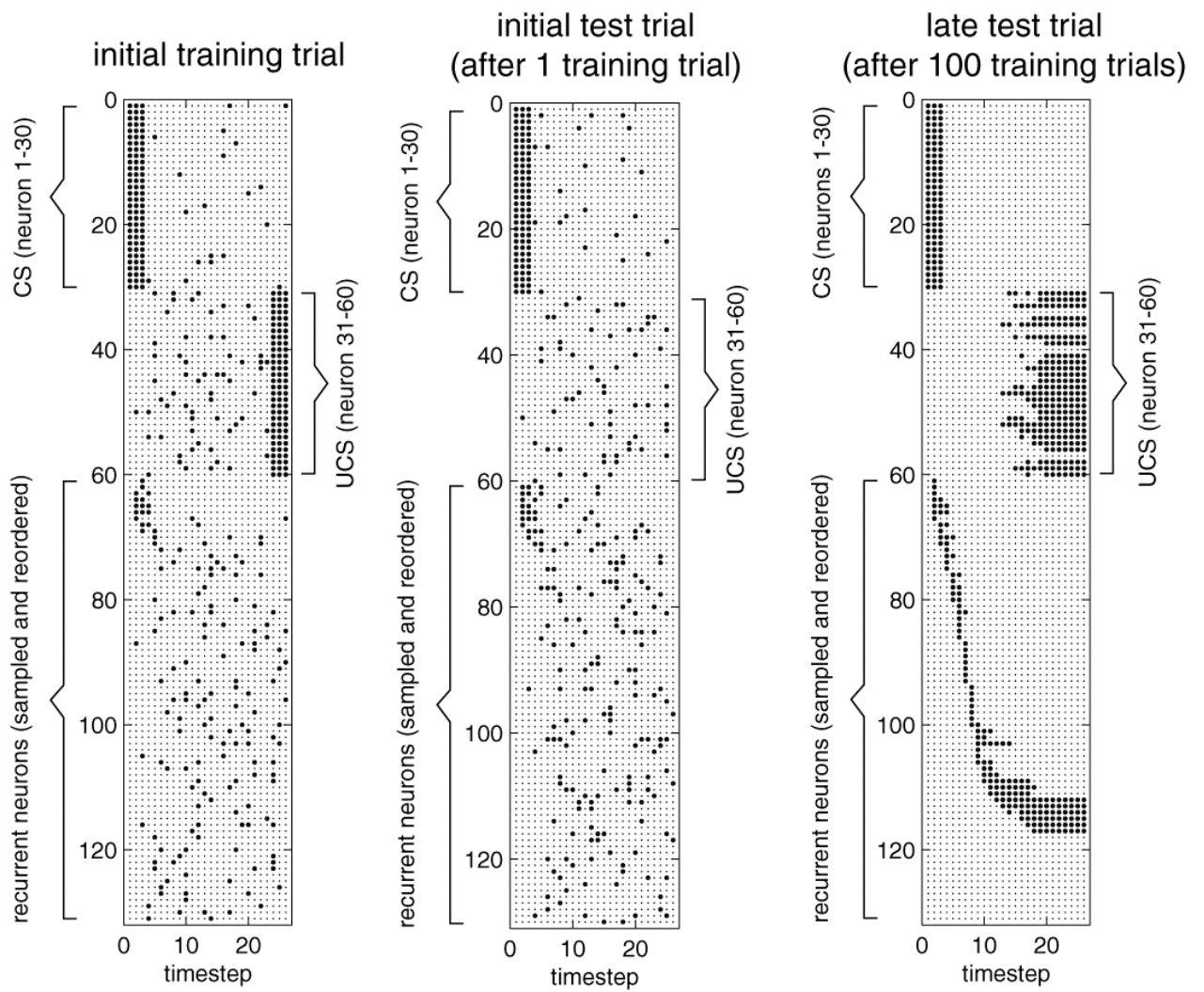


Figure 9.

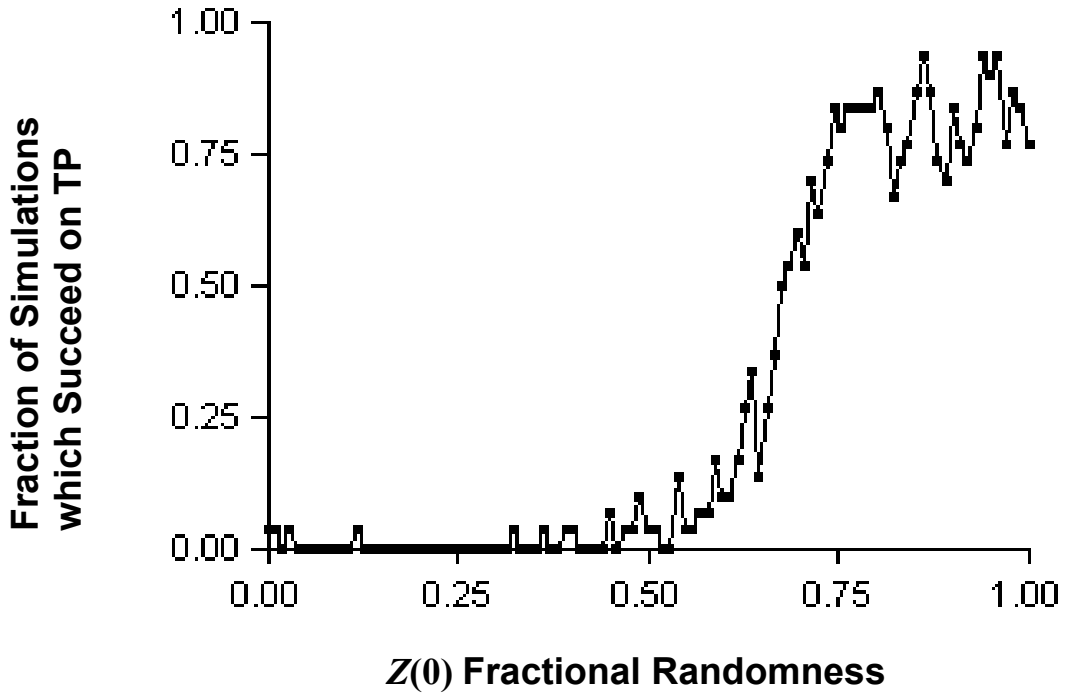


Figure 10.

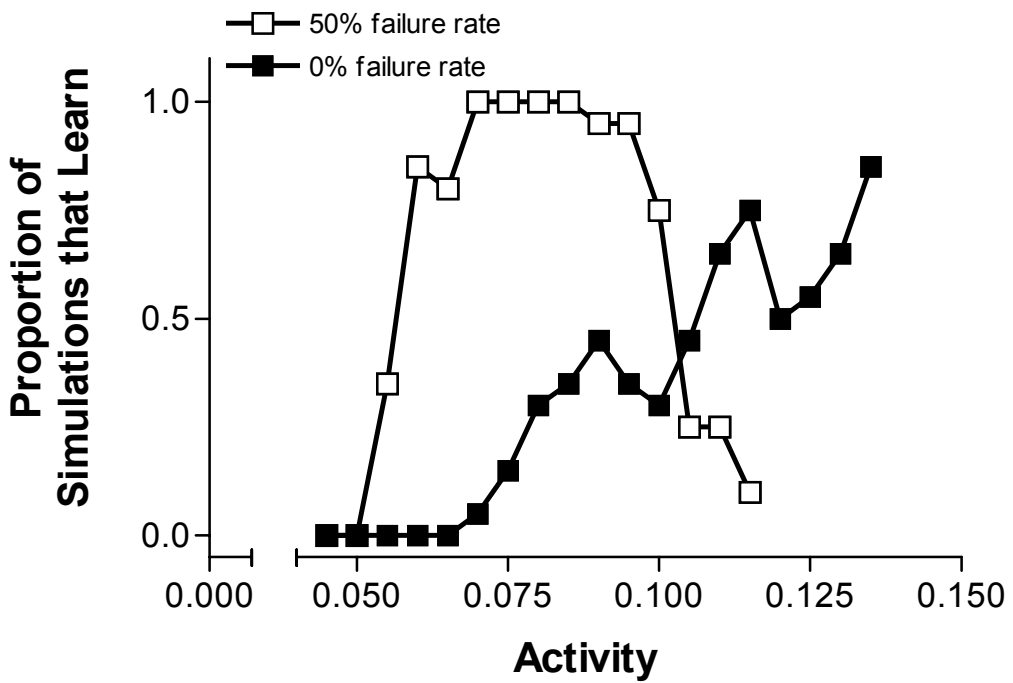


Figure 11.

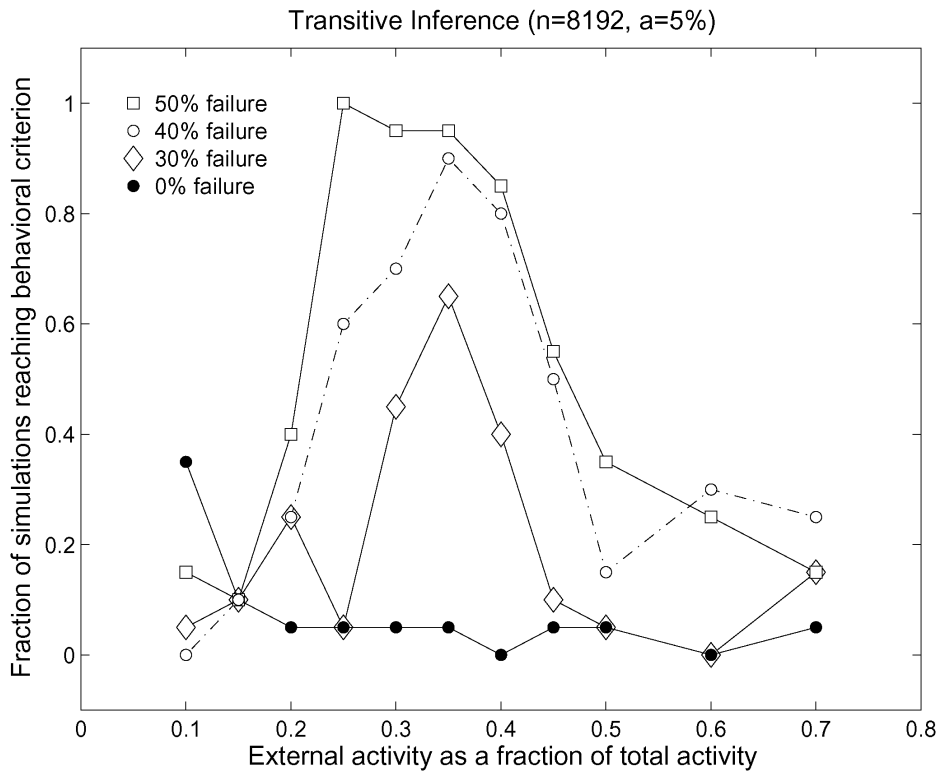


Figure 12.

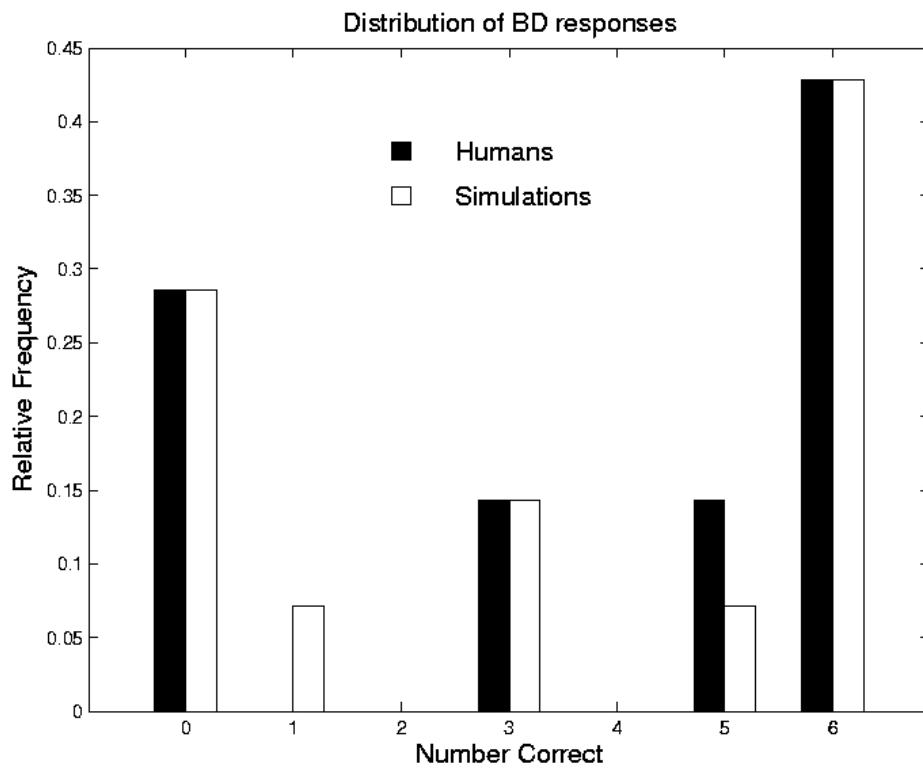


Figure 13A.

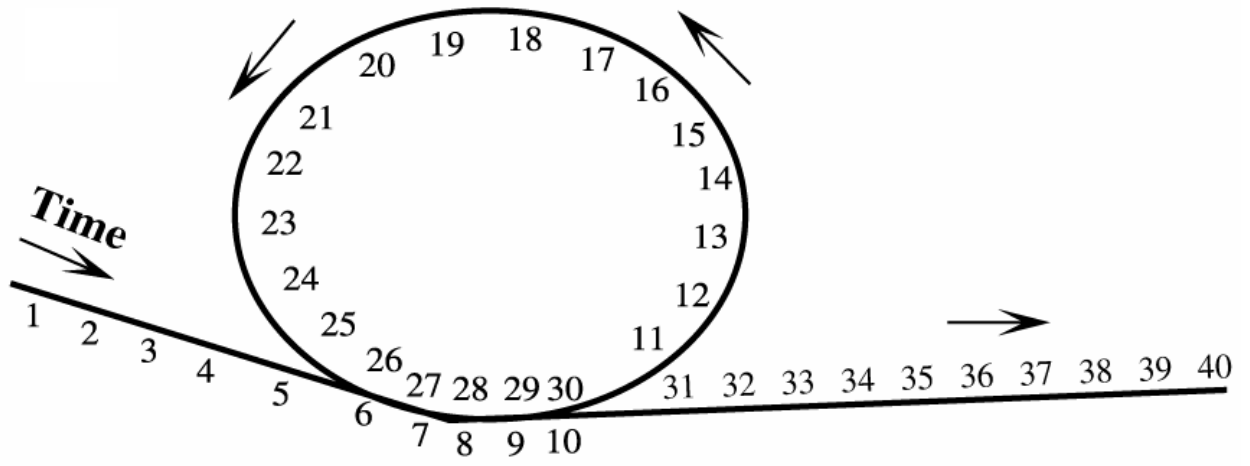


Figure 13B.

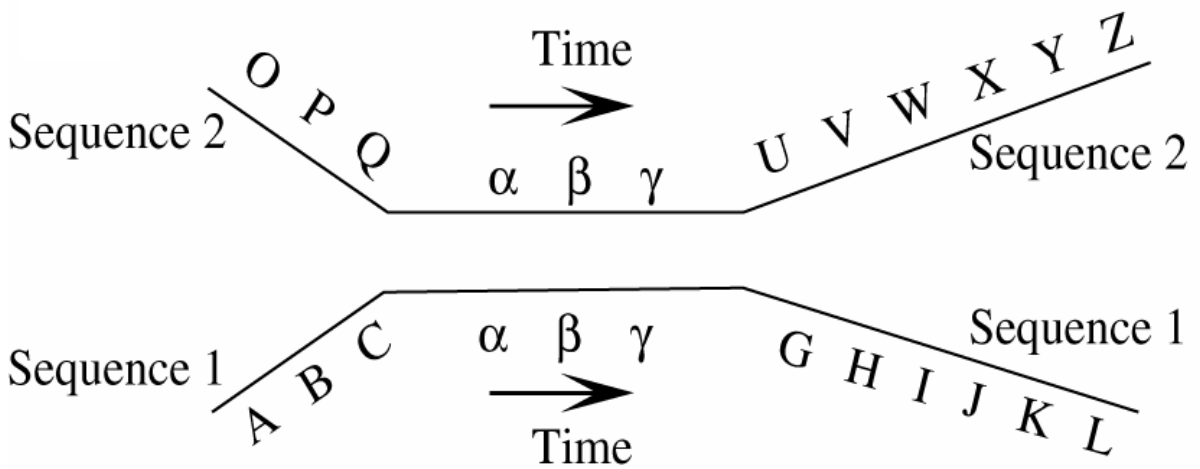
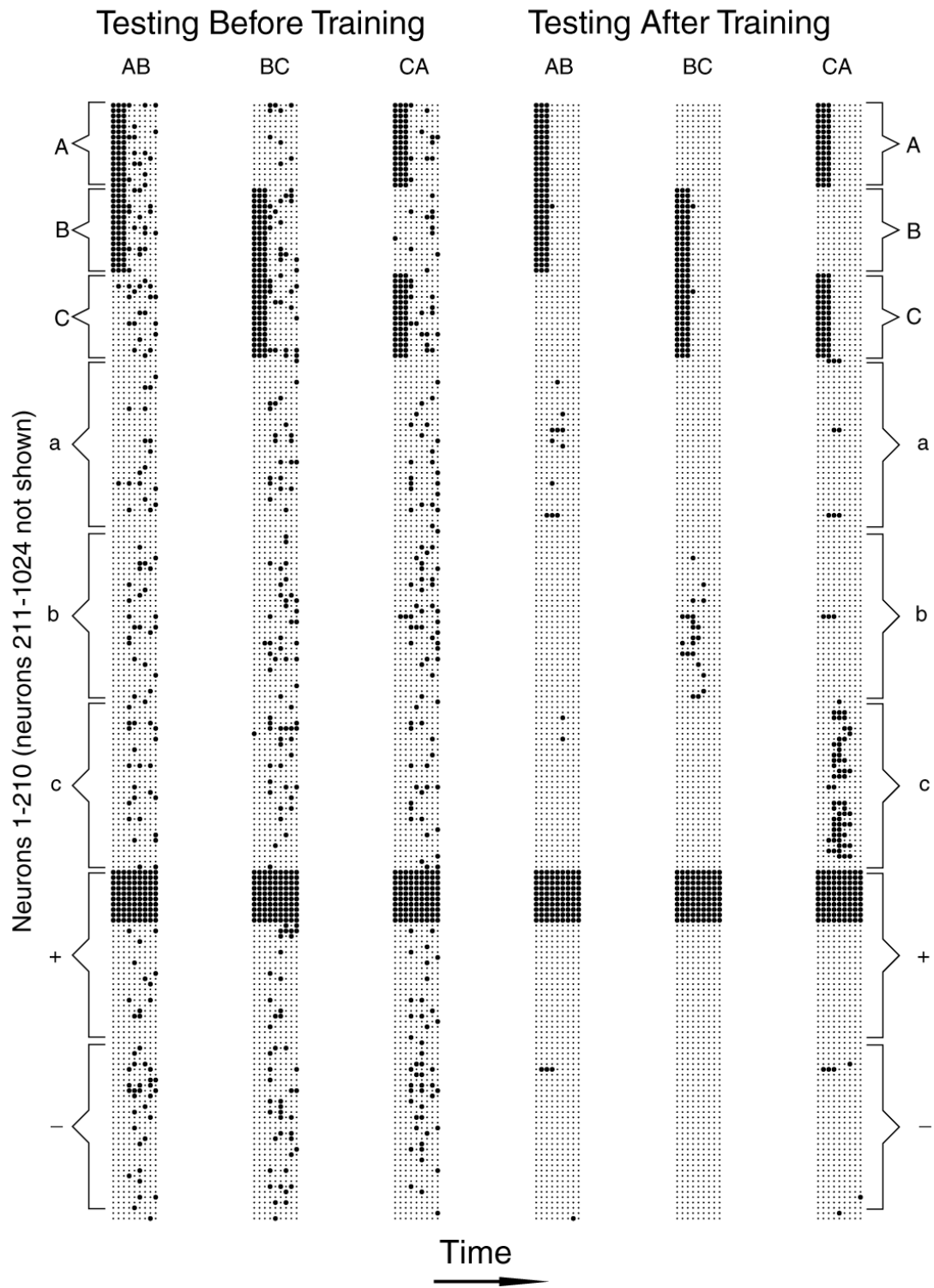


Figure 14.



**Figure 15.**

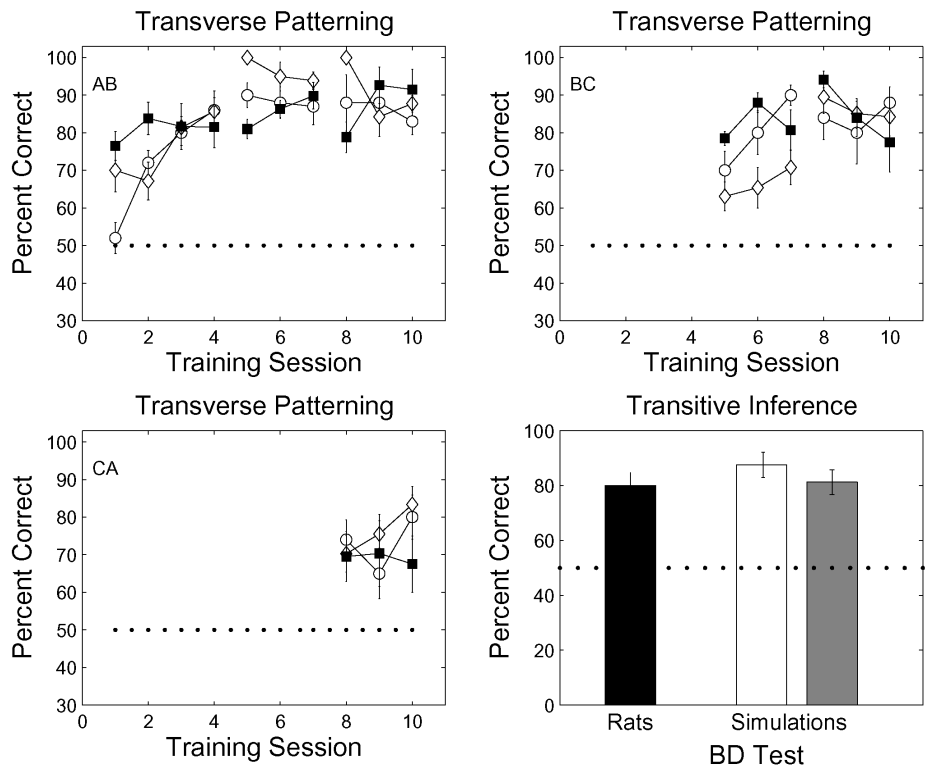


Figure 16.

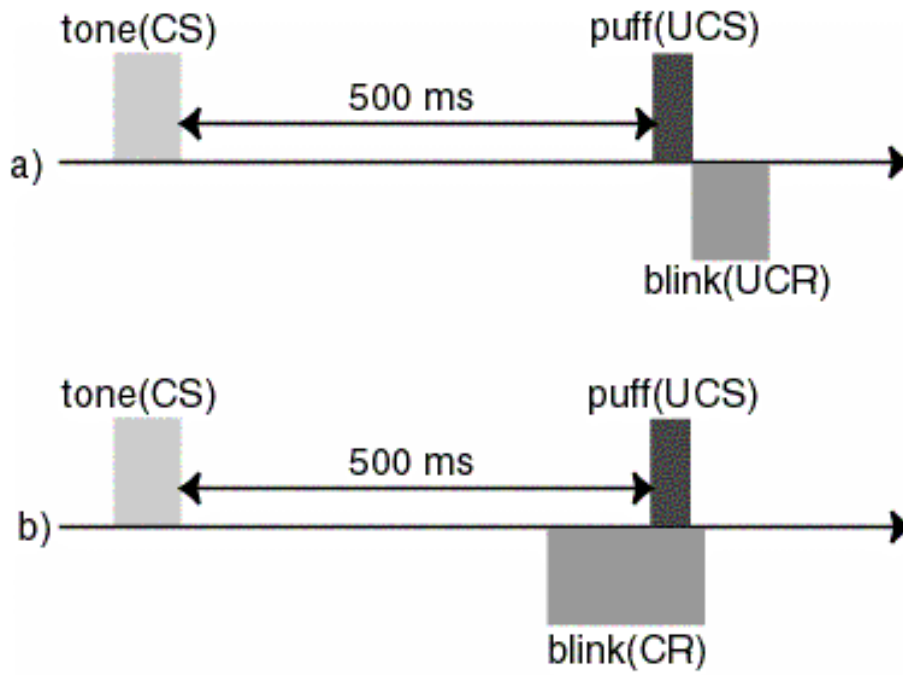
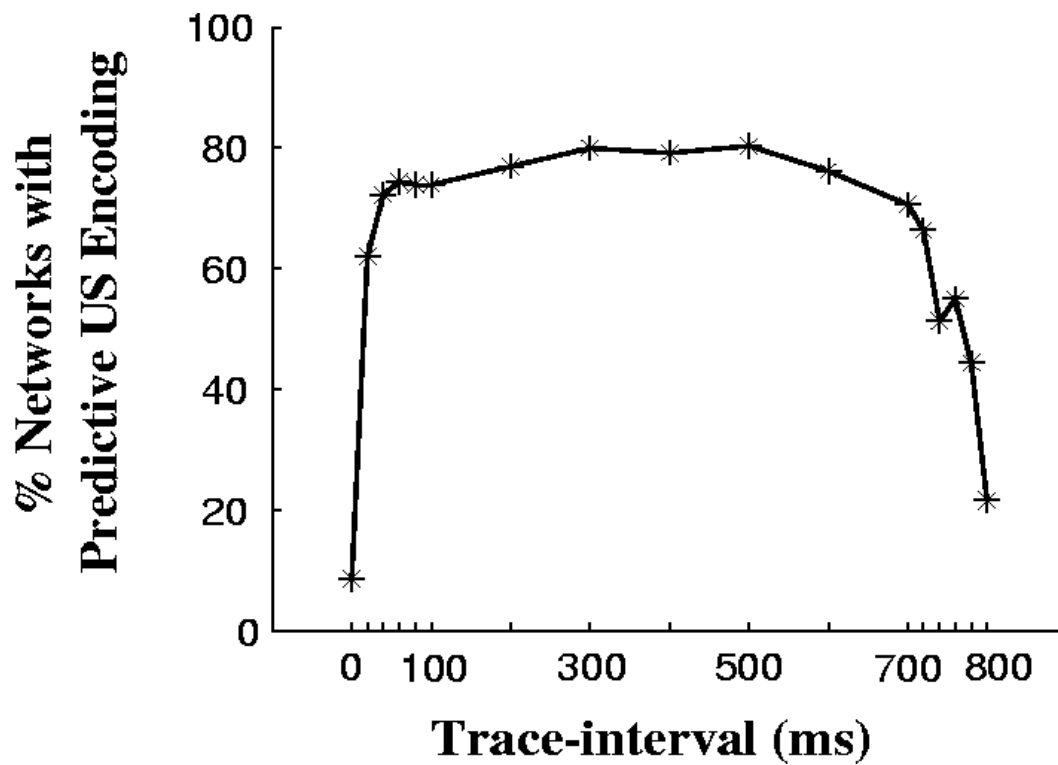


Figure 17.



**Figure 1.** (a) The EC and the DG inputs are collapsed into a single powerful external input. Most excitation is recurrent, but feedback inhibition helps maintain activity levels. (b) Recurrent excitatory connectivity is sparse and random.

**Figure 2.** Tight packing can lower the complexity of a representation and enhance sequence length memory capacity. (A) Four representations (●) in a 3-d space (cube) are transformed into four representations in a 2-d space (a face of the cube). (B) A neuron corresponds to a dimension of the cube. Initially, three neurons are needed for the four representations scattered around the eight corners of the cube. However, only two neurons are needed for the representation after the transformation, and thus the representation is simplified and memory capacity is enhanced.

**Figure 3.** Recoding of predictable sequences increases representational similarity and thus creates on-line temporal compression. The development of local context firings enhances the similarity of successive patterns in CA3 (e.g., A, A', A'', A''') even though the input (e.g., W, X, Y, Z) might be quite different. Figure modified from Levy (1989).

**Figure 4.** Recoding enhances similarity between successive CA3 states. A simulation with an orthogonal input sequence is recoded into a CA3 state space sequence of overlapping patterns. All neurons receive recurrent input, but only neurons 1-520 are externally activated. The input sequence consists of 13 orthogonal patterns of 40 activated neurons each with a dwell time (stutter) of three time-steps. Note that the firing pattern for a typical, solely recurrent neuron changes from a somewhat random pattern in panel 3 to a local context pattern of firing in panel 4; that is, after training, individual neurons can be described across a trial as first off for a long period – then on repeatedly for a short period – then off for the rest of the time. External pattern

1 is coded by selectively activating neurons 1-40. External pattern 2 consists of selective activation of neurons 41-80 and so on for each of the 13 input patterns. The simulation used  $n = 4096$ ;  $a \approx 5\%$ ; no quantal failures;  $\mu = 0.01$ ;  $\alpha = 0.7165$ ;  $K_0 = 0.964$ ;  $K_{FF} = 0.018$ ;  $K_{FB} = 0.053$ , and  $\lambda = 0.5$ . One hundred training trials were used. A big dot stands for a firing ( $Z_j = 1$ ) and a small dot stands for a non-firing ( $Z_j = 0$ ). Time goes from left to right and neurons go from top to bottom.

**Figure 5.** Training-induced time shift of the recurrent neuron encoding. This histogram quantifies the across trial backward cascade of recurrently activated neurons although a small fraction of neurons shift forward (later), most neurons shift backward (earlier). Shift for each neuron is defined as the change in onset time of its initial firing on training trial 5 vs. trial 250. The median shift is -34 time-steps. However, by only comparing firing-onset shifts, this figure does not reflect the full skewing of place fields since local context lengths are also changing across training trials. The external sequence used a stimulus dwell time of nine with  $\alpha \approx e^{-1/9}$  (see Mitman et al. (2003) for further details).

**Figure 6.** Example of temporal compression in a model using integrate-and-fire neurons (August & Levy 1999). Panel A illustrates the last training trial of a circular sequence. There are 1,000 neurons in the simulation although only 500 are illustrated. All of the 100 externally activated neurons (1-100) are plotted. Panel B shows spontaneous, compressed replay. During replay testing, inhibition is decreased so that activity levels rise and a small amount of external random activity is applied as a model of slow wave sleep.

**Figure 7.** The training experience produces a modest backward cascade of externally activated neurons. On-line compression is achieved by the development of local context firing, and

prediction (forecasting) is achieved by the earlier onset of external firing, a result of the backward cascade. Firing of the first 1000 neurons of the 4096 total neurons in a network during the fifth (upper figure) and 250th (lower figure) training trials. The external inputs to the network include neurons 1 through 960. Neurons 961 through 1000 are recurrently activated if they fire. The inputs are a sequence of noisy orthogonal patterns, each with a dwell time of nine time-steps (this figure uses dashes instead of dots to indicate firing). The rectangular blocks along the diagonal in the upper figure are the externally activated neurons. These rectangles are not solid because of input noise that randomly deselects external activation. See Mitman et al. (2003) for details.

**Figure 8.** Training on the trace conditioning task leads to earlier and more prolonged firing of the UCS neurons. As can be seen here, the recoding characterizations of backward cascade and enhanced context length are the successful recoding result as in a trace conditioning paradigm. Each large dot is a cell firing. These simulations are competitive versions of the model running 1000 neurons with 10% connectivity and 10% activity. The NMDA-R-like synaptic modification rule is used with  $\alpha = 0.7165$  (corresponding to a time-step size of approximately 33 ms); synaptic modification rate is 0.05. Neurons are reordered for the purposes of this figure. CS and UCS are 3 time-steps (100 ms). Trace interval is time-step 4 to time-step 23. The trace exists for 20 time-steps (667 ms). See Levy et al. (2005) for details.

**Figure 9.** Randomness of initial firing state,  $Z(0)$ , facilitates learning of transverse patterning (TP). Performance begins improving once approximately 50% of the  $Z(0)$  neurons are randomized from trial to trial. The fraction of the simulations that successfully learned TP is plotted as a function of the number of randomly chosen neurons in each  $Z(0)$  state of a simulation. See Shon et al. (2002) for details.

**Figure 10.** Adding synaptic failures improves performance at lower activity levels. In simulations without synaptic failures, performance is poor when fractional activity is below 11% and only reaches the behavioral/experimental criterion at 13.5%. However, setting the synaptic failure rate at 50%, while keeping other parameters the same, improves learning for activity levels from 5.5% to 10% and produces criterion performance across the activity range of 6.5 to 9.5%. Each plotted point represents the fraction of 20 simulations that successfully learned transverse patterning. The number of externally activated neurons per time-step ( $m_e$ ) accounts for 30% of total activity, regardless of relative activity setting. The number of neurons,  $n$ , is 8192. See Sullivan & Levy (2004) for details.

**Figure 11.** Random recoding helps transitive inference (TI) learning. Synaptic failures are necessary for solving the (TI) problem at low activity levels, and there is a restricted range of external activations for best performance. At 0% synaptic failures with 5% activity, the performance is never better than 40%, but failure rates of 40% and 50% produce performance of 80% or better for a range of external activations. Note that if external activity is too high (> 40%) or too low (< 25%), synaptic failures are not able to recover the behaviorally observed levels of performance. The fraction of simulations that learn is plotted as a function of fractional external input ( $\sum X_i(t)$  divided by number of  $\sum Z_i(t)$ ). As relative external activation is changed, the overall activity level is kept approximately constant. Most simulation parameters were fixed ( $n=8192$ ,  $\theta=0.5$ ,  $\mu=0.05$ ,  $c=10\%$ ,  $w_0=0.45$ ,  $\alpha=0.8669$ ,  $\lambda=0.5$ ), but  $K_{FB}$ ,  $K_{FF}$ , and  $K_0$  must be adjusted to keep total activity around 5% when external activity is changed. For each simulation, good performance is defined as 80% or more correct responses at the end of training for all comparisons AB, BC, CD, DE, BD, and AE. Twenty simulations were trained and tested for each data point.

**Figure 12.** Frequency distribution comparing the number of correct transitivity (BD) responses in the inference test for seven human subjects and fourteen simulations of the computer model. Despite small sample sizes, the human experiment and the computer model show remarkably similar response distributions. For both the human experiment and the computer simulations, there were four premise pairs of five atomic stimuli learned via the staged training paradigm. For the human experiment, the stimuli were five distinct Kanji characters described elsewhere (experiment 1 of Greene et al., 2001). For the computer model, orthogonal blocks of neurons coded each stimulus of a pair.

**Figure 13.** Sequences to be learned in two environments: **a.** The looping path input sequence: this sequence of forty patterns has an identical pair of subsequences (6-10 and 26-30). Learning trials consist of the complete 40-pattern sequence and then resetting the network with a noisy input. **b.** Two overlapping sequences can be learned individually. The two twelve-pattern sequences share a common subsequence of three patterns ( $\alpha, \beta, \gamma$ ). During learning, these two sequences were randomly presented with a noisy pattern between the end of one trial and the beginning of the next one.

**Figure 14.** Training is necessary for reliable correct decision patterns. Illustrated here are firing patterns for test trials given before and after training. The first three vertical strips show a subset of cell firing patterns during testing before training when stimulus AB, BC, or CA, and 10 neurons of the positive input pattern are presented respectively. Note that the firing patterns of decision neurons (indicated by a, b, and c) are more random prior to training. After training, the network predicts the a coding over b coding when AB is the input, the b coding over c coding when BC is the input, and the c coding over a coding when AC is the input.

**Figure 15.** Reproducing the variation of learned transverse patterning and learned transitive inference. The network reproduces the learning curves and SEM's of the transverse patterning behavioral experiments (Alvarado & Rudy 1992, 1995). There are three phases of Training (I, II, and III), corresponding to sessions 1-4, 5-7, and 8-10 respectively, where a session is a block of 30 trials. There are three pairs of stimuli that form each overlapping problem set. A new problem set is added at the beginning of each training phase. Thus, problem set {A,B} where A is correct and B is incorrect is always part of training; problem set {B,C} where B is correct and C is incorrect is present in phases II and III; and problem set {C,A} where C is correct and A is incorrect is present only in phase III. Open circles (Alvarado & Rudy 1992) and open diamonds (Alvarado & Rudy 1995) are behavioral data and filled squares are the simulation results. Note the agreement between our simulations and the behavioral data. Error bars represent the SEM over seven simulations and seven rats (Alvarado & Rudy 1995). In the bar graph on the lower right, the same parameterization of the model also reproduces the behavioral data of the transitive inference problem. Proportion correct for the transitivity testing (BD) is shown here. Note again that the simulations not only reproduce the percent correct on the critical BD test, but they also reproduce the SEM's of the rat data for the same sample size of eight.

**Figure 16.** The trace conditioning paradigm. A 500 ms trace interval both (a) before and (b) after training. In this paradigm, a tone is played and then turned off. The stimulus free interval that follows tone offset is called the trace interval. At the end of the trace interval (e.g., 500 ms), a puff of air is delivered to the rabbit's eye. The rabbit has successfully learned if it can "blink" (draw across its nictitating membrane) in time to intercept the puff of air. Blinking too soon or too late is a failed trial. Thus, such an air puff trace paradigm is an escape task. Paradigms using electric shock are not.

**Figure 17.** Performance of the network simulations is best for trace-interval durations of 100-600 ms (Rodriguez & Levy 2001). Average performance across training is plotted for different trace-interval durations. Each point is the percentage of 10 networks with a predictive US encoding averaged across 750 training trials. Extra data points in the 1-100 ms range show a fast change in performance for short trace intervals, whereas the 700-800 ms range shows a more gradual change. The graph has an inverted U shape comparable to behavioral data (Gormezano et al. 1983).