Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 18 (2005) 1163–1171
Neural Networks
www.elsevier.com/locate/neunet

2005 Special issue

Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model

Adam Johnson [a], A. David Redish [b,*]

[a] Center for Cognitive Sciences and Graduate Program in Neuroscience, University of Minnesota, MN 55455, USA
[b] Department of Neuroscience, University of Minnesota, MN 55455, USA

Abstract

Temporal difference reinforcement learning (TDRL) algorithms, hypothesized to partially explain basal ganglia functionality, learn more slowly than real animals. Modified TDRL algorithms (e.g. the Dyna-Q family) learn faster than standard TDRL by practicing experienced sequences offline. We suggest that the replay phenomenon, in which ensembles of hippocampal neurons replay previously experienced firing sequences during subsequent rest and sleep, may provide practice sequences to improve the speed of TDRL learning, even within a single session. We test the plausibility of this hypothesis in a computational model of a multiple-T choice-task. Rats show two learning rates on this task: a fast decrease in errors and a slow development of a stereotyped path. Adding developing replay to the model accelerates learning the correct path, but slows down the stereotyping of that path. These models provide testable predictions relating the effects of hippocampal inactivation as well as hippocampal replay on this task.

1. Introduction

Models of reinforcement learning are based on the simple premise of trying to optimize reward receipt. Temporal difference models maximize reward receipt by learning the *value* of achieving a state of the world (Daw, 2003; Sutton & Barto, 1998). For episodic tasks with a finite horizon, the value of a state can be defined as the expected reward achievable from that state. If the agent can identify the consequences of its actions on the state of the world, then it can predict the value of those subsequent states. Maximizing reward then requires selecting the action, which leads to the state with maximal estimated value.

It is possible to learn the estimated value as the agent experiences the world, taking actions based on the current value estimation, and updating the estimation through observations. In these algorithms, the agent maintains an estimate $V(s)$, and takes an action based on the expected value of the consequences of that action (the expected reward and the expected value of the state achieved via the action). Upon taking an action, thereby reaching state $s(t+1)$, the agent is able to observe the actual reward $r(t+1)$, as well as its estimate of $V(s(t+1))$. From these observations, the agent can update its value estimate of the previous state $V(s(t))$. If more reward is received than expected or if the state the agent finds itself in is more valuable than expected, then the estimate of $V(s(t))$ needs to be increased. If, in contrast, less reward is received or if the state the agent finds itself in is less valuable than expected, then the estimate of $V(s(t))$ needs to be decreased. Through experience, the value estimate begins to approach the real value function (Sutton & Barto, 1998).

These algorithms learn very slowly over many episodes (requiring thousands or tens of thousands of episodes, Sutton & Barto, 1998). A variety of methods have been developed to speed up the learning process. One of the most common is *practice*, which allows the agent to use previously recorded experience or a model to simulate experience following the episode and update the value and policy function accordingly. In the Dyna-Q family of algorithms (Sutton, 1990; Sutton & Barto, 1998), a state and action are selected from previous experience (that is, from memory) and used to update the value function between episodes. Dyna-Q thus requires a replay of the agent's recent experience during periods of rest between episodes.

In this paper, we suggest that the well-observed replay of recent memories in hippocampus during periods of rest and sleep (Lee & Wilson, 2002; Louie & Wilson, 2001; Nádasdy, Hirase, Czurkó, Csicsvari, & Buzsáki, 1999; Pavlides & Winson, 1989; Skaggs & McNaughton, 1996; Wilson & McNaughton, 1994) may provide a mechanism implementing a

* Corresponding author.
E-mail addresses: john5726@umn.edu (A. Johnson), redish@ahc.umn.edu (A.D. Redish).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.08.009

practice component of TDRL algorithms. Early observations showed that hippocampal pyramidal cells that were active during a task remained active during sleep following the task performance (Pavlides & Winson, 1989) and that place cell activity was highly correlated between cells with overlapping place fields and reduced in cells without overlapping place fields (Wilson & McNaughton, 1994). Later studies gave shape to the neural activity: place cell firing is ordered—during sleep episodes place cells fire in the order they were encountered by the rat during task performance (Lee & Wilson, 2002; Louie & Wilson, 2001; Nádasdy et al., 1999; Skaggs, McNaughton, Wilson, & Barnes, 1996). This has been termed *route replay* [see Redish, 1999, for review]. Single unit activity in sleep or rest periods shows increased spike probability associated with sharp-wave ripple events (Buzsáki, 1989; Kudrimoti, Barnes, & McNaughton, 1999; O'Keefe & Nadel, 1978). Further evidence showed that during slow wave sleep, replay occurs during these sharp-wave ripple events (Kudrimoti et al., 1999; Lee & Wilson, 2002). Because place cell activity shows increases in temporal structure following task performance, hippocampal activity during replay has been suggested to be associated with memory consolidation processes (Buzsáki, 1989; Lee & Wilson, 2002; Louie & Wilson, 2001; Marr, 1971; McNaughton, 1983; Pavlides & Winson, 1989; Redish, 1999; Redish & Touretzky, 1998; Skaggs et al., 1996; Wilson & McNaughton, 1994).

Hippocampal replay has been examined on simple runway tasks (Lee & Wilson, 2002; Louie & Wilson, 2001; Nádasdy et al., 1999; Skaggs et al., 1996; Wilson & McNaughton, 1994). To gauge the effect of replay on behavior in an explicit choice-task, we examined a model of TDRL learning a multiple-T task (described by Schmitzer-Torbert & Redish, 2002, 2004; Fig. 1(A)). In order to receive the reward, the animal had to successfully navigate a sequence of T choices. Although the task used by Schmitzer-Torbert and Redish (2002, 2004) formed a loop and animals were not removed from the multiple-T maze between laps, they tended to run the maze episodically, pausing for a long time (mean 27 s) at the second feeder before running another lap quickly (mean 16 s) (Schmitzer-Torbert and Redish, unpublished observations). Rats running the multiple-T task showed two differentiable learning rates: a fast decrease in the number of errors (the number of incorrect choices per lap) and a slow increase in the regularity of the path on each lap (path stereotypy) (Schmitzer-Torbert & Redish, 2002, 2004).

2. Methods

2.1. TDRL model

The TDRL component of the model was based on a standard SARSA Q-learning algorithm over a continuous state, discrete action space (Sutton & Barto, 1998). More specifically, the model consisted of a continuous state-space TD model using function approximation with radial basis functions. State-action value function approximation was accomplished by

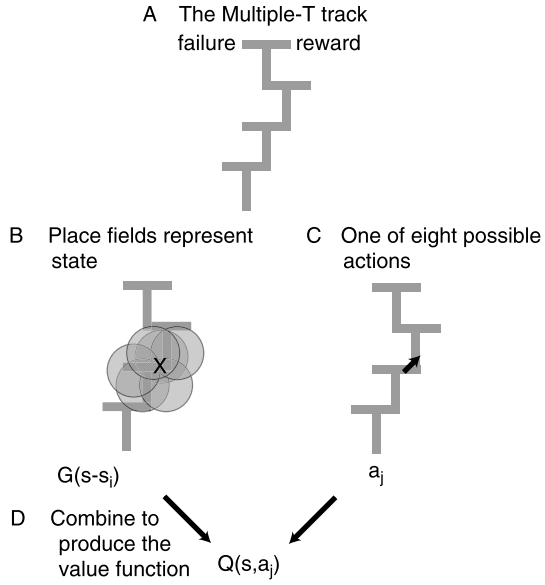

Fig. 1. Summary of the multiple-T task and model. The task consisted of a sequence of four T choices with reward available on one arm of the final T. The model used radial basis functions to compute the state-action value $Q(s, a)$ over a continuous state, discrete action space. Each action was selected using *softmax*. (A) One example track as used in the model. (B) Place fields represent state through a distributed encoding. x indicates the position of the animal, active units are shown. (D) State (B) and one of the eight possible actions (C) were each associated with a value function $Q(s, a_j)$.

associating a set of *place* neurons assigned to n_s static positions randomly distributed over the state-space with each of n_a actions to produce a scalar value (or activity), $Q(s_i, a_j)$, the 'quality' or 'value' of being in state s_i and taking action a_j. Action directions were distributed evenly over the interval $[0, 2\pi]$. The value of any state-action in the continuous state-space was thus given by the sum of values weighted by their similarity to the current state s:

$$Q(s, a_j) = \sum_{i=1}^{n_s} Q(s_i, a_j) G(s - s_i) \tag{1}$$

where s_i is the *static* position of the ith neuron in state-space and $G(s-s_i)$ is a Gaussian kernel with standard deviation σ. It should be noted that G effectively implements a *place code* with place field centers located at s_i.

Following the standard discrete time TD model, the prediction error signal δ was

$$\delta = r + \gamma Q(s(t+1), a(t+1)) - Q(s(t), a(t)). \tag{2}$$

which could then be used to update $Q(s_i, a_j)$ by

$$Q(s_i, a_j) \leftarrow Q(s_i, a_j) + \eta \delta G(s(t) - s_i) \langle a(t) \cdot a_j \rangle_+ \tag{3}$$

where η is a learning rate parameter and $\langle a \cdot a_j \rangle_+$ the positive component of the dot product of the last action a and each possible action a_j. Thus, $G(s-s_i)$ measured the distance between the actual state and the state contributing to $Q(s_i, a_j)$ and $\langle a \cdot a_j \rangle_+$ measured the distance between the actual action and the action contributing to $Q(s_i, a_j)$.

Action selection followed the standard softmax method,

$$P(a(t)) = a_j = \frac{e^{\beta Q(s(t),a_j)}}{\sum_j e^{\beta Q(s(t),a_j)}} \quad (4)$$

where β is a temperature parameter and $Q(s, a_j)$ is the value of each action a_j at the current position s.

2.2. The practice signal

Replay was implemented probabilistically by allowing cellular activity to propagate across a learned transition matrix $W_{\text{transition}}$. The transition matrix was initialized to 0 at the start of the simulation for each new session (i.e. each day) but carried over across laps (i.e. within each session). On each active timestep, the transition matrix was updated via:

$$W_{\text{transition}} \leftarrow W_{\text{transition}} + \zeta \arctan(\bar{P}^T(t)[\bar{P}(t) - \bar{P}(t-1)]) \quad (5)$$

where $\bar{P}(t)$ indicates the vector of current place cell activity ($\bar{P}(t) = G(s(t) - s_i) \forall$ cells $i, 1 \le i \le n_s$), and ζ is a learning rate. This is similar to a discretized time version of the weight matrix used by Blum and Abbott (1996) or an approximation of the learning that occurs through the combination of spike-time-dependent-plasticity and phase-precession (Redish & Touretzky, 1998).

At the completion of each lap, n_r replays were attempted. Each replay attempt began with one randomly selected place cell. The activity of that selected cell was set to 1, and it propagated activity to other cells via the learned transition matrix. The next active cell was found using a winner-take-all selection method, selecting the cell with the largest activity produced by the outputs of the active cell. If the activity of the next active cell was greater than threshold, $r_{\text{threshold}}$, it was counted as part of the replay, its activity was set to 1, and it was allowed to propagate its activity onward. This process was repeated until either the activity did not reach threshold or the replay reached the end of the maze. At each step of each replay, the state-action value function $Q(s, a)$ was updated according to the equations given above except that the action was estimated as the one most similar to the change in state.

The values for the parameters used in the model are given in Table 1.

Table 1
Parameters used in the model

Parameter		
n_a	8	Number of actions
n_s	430	Number of place cells
σ	2.5	Place field width
β	1.0	Softmax temperature
Stepsize	0.8	Movement length
γ	0.99	Discounting factor
η	0.6	Value function learning rate
n_r	8	Number of replays attempted per lap
$r_{\text{threshold}}$	$n_s/50$	Replay activity threshold
ζ	1.0	Transition matrix learning rate
Tracksize	(38,38)	
Trackwidth	2	

2.3. Data analysis methods

Twenty-four sessions were simulated with replay and without replay. Each session consisted of eighty laps. If the agent did not complete the maze (i.e. reaching 'failure' or 'reward' in Fig. 1(A)) within 2000 steps, the agent was removed from the lap and no reward was given, that is, the lap ended in 'failure'. The minimum number of steps an agent would require to get to reward depended on the specific configuration of the maze, but was approximately 200 steps.

2.3.1. Errors

Errors were defined as entry into an incorrect arm at each T choice point. Because repeated (or extended) entries into an incorrect arm were only counted once, a maximum of four errors could occur on a maze with four T choices.

2.3.2. Path-stereotypy

Path stereotypy was defined as the correlation of spatial path through the maze for each lap following the methods outlined by Schmitzer-Torbert and Redish (2002). Briefly, this included calculation of a path correlation matrix that summarized the correlation between the pixelated path for each lap with each other lap. The linearized path correlation at lap i, identified as *path stereotypy* hereafter, was defined as the mean correlation of lap i with all other laps.

2.3.3. Amount of replay

Replay was quantified by counting the length of each cascade as the number of place cells activated following each lap. This number was summed over each of the n_r replays for each lap.

3. Results

The basic results of the model are summarized in Figs. 2 and 3. The simulation provided four primary results: (1) The temporal difference model learned the correct path through the maze. (2) Errors decreased with experience. (3) Path stereotypy increased with experience. (4) Replay developed with experience.

3.1. The model successfully learned to navigate the multiple-T maze

Both models (with and without replay) learned the multiple-T and showed a decrease in the number of steps required to complete each lap and an increase in the average reward per lap over a training session (Fig. 2). While the model with replay shows a slightly faster initial decrease in the number of steps required for each lap, the model without replay reaches a slightly lower steady state (beyond 50 laps). The model with replay acquired more reward on average than the model without replay.

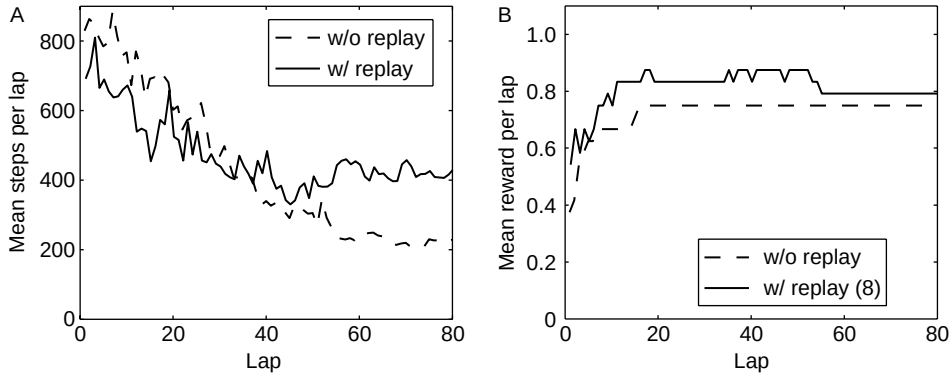

Fig. 2. Comparison of learning in TDRL models with and without developing replay-based practice. (A) Both models show a decrease in the number of steps to complete each lap. The number of steps required to achieve the optimal path depended on the configuration of the maze, but was approximately 200 steps. (B) The model with replay typically acquires more reward than the standard model. For agents that completed a lap, chance would be 50%, however, because agents were removed after 2000 steps, chance is actually much lower than 50%.

3.2. Errors decrease with experience

Similar to animal behavior, the number of errors decreased early in the simulated training session. However, it should be noted that the decrease in errors in the simulation was slower than what is observed in animal behavior (compare Schmitzer-Torbert & Redish, 2002).

3.3. Path stereotypy increased with experience

The path stereotypy increased over the simulated training session. In combination with the improved decrease in errors, this indicates a trend toward a more efficient path through the maze. However, the increase in path stereotypy was still slower than what was observed in animal behavior (compare Schmitzer-Torbert & Redish, 2002, 2004).

3.4. Replay developed with experience

Simulations showed slow development of route replay (Fig. 4). Because replay develops probabilistically in this model, increasing the number of replays at the completion of each lap, n_r, should magnify differences between the models.

4. Discussion

Temporal difference models of navigation learn mazes very slowly (much more slowly than real animals). Algorithms developed to speed up TD learning models include indirect practice occurring via replay of recent memories between experiential episodes. We suggest that the replay observed during hippocampal sleep states may occur in periods of awake rest, such as occur between laps on the multiple-T maze, and that this replay may provide the needed indirect practice to speed up the TD algorithms. In a computational model of the multiple-T task, adding replay sped up the agent's learning of the correct choices, but because it also produced more variability, it slowed down the development of the stereotyped path. From this model and these hypotheses, we can make a number of testable predictions.

4.1. Predictions

4.1.1. Route replay should occur during awake behavior

The model presented above requires replay in the pauses taken between laps as the animal learns the multiple-T. Extensive evidence has shown that route replay occurs during

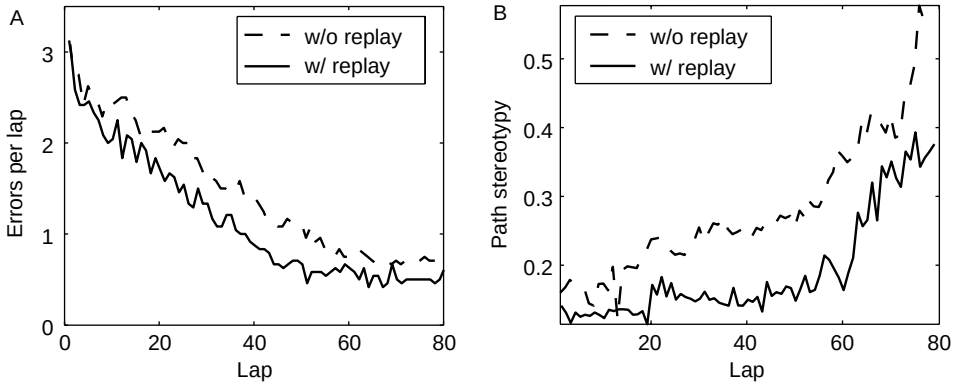

Fig. 3. Comparison of TDRL models with and without developing replay-based practice over 24 trials. (A) The model with replay shows a faster decrease in the number of errors per lap than the standard model. (B) The model with replay shows a slower onset of path stereotypy than the standard model.

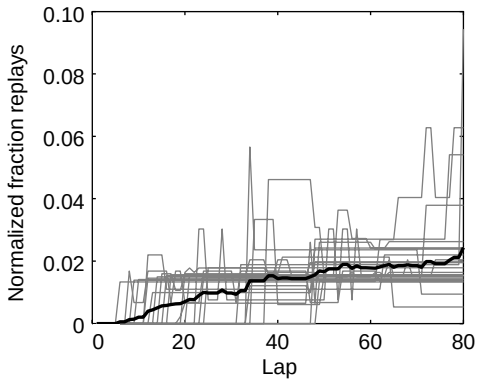

Fig. 4. Replay developed over the course of the session. Replay strength was measured as the length of the replayed sequence. Plotted are the proportions of replay strength occurring at each lap (length of replay on each lap divided by total length of replay over entire session). On early laps, the replay tended to fail early, thus producing short replays, while on later laps, replay tended to proceed robustly, thus, producing longer replays.

sleep states in between experimental sessions (Lee & Wilson, 2002; Louie & Wilson, 2001; Nádasdy et al., 1999; Skaggs et al., 1996). Sharp waves do occur during periods of rest and immobility (Buzsáki, Leung, & Vanderwolf, 1983; O'Keefe & Nadel, 1978). Although they have not been explicitly studied separately, some researchers have included sharp waves occurring during awake rest states in their analyses and do not report differences between replay during those awake states and replay during sleep states (Kudrimoti et al., 1999; Wilson & McNaughton, 1994).

4.1.2. Route replay should develop during a session

In the model as presented, the role of replay is to allow the system to practice the task as presented and thus, indirectly learn it more quickly. This means that replay should develop over the course of behavior. As can be seen in Fig. 4, in the model, replay develops over the course of the session.

Current theories suggest that sharp waves are indicative of activity cascades derived originally from recurrent connections in CA3 (Csicsvari, Hirase, Mamiya, & Buzsáki, 2000; Shen & McNaughton, 1996; Ylinen et al., 1995) and that the replay component follows asymmetries that arise through learning (Nádasdy et al., 1999; Redish, 1999; Redish & Touretzky, 1998; Skaggs & McNaughton, 1996). These hypotheses predict that sharp-wave ripple activity should develop with a time-course similar to that of place field expansion. With experience running along a path, place fields stretch backwards along the direction of travel (Lee, Rao, & Knierim, 2004; Mehta, Barnes, & McNaughton, 1997; Mehta, Quirk, & Wilson, 2000). The phenomena of asymmetric plasticity (Bi & Poo, 2001; Levy & Steward, 1983) and phase precession (O'Keefe & Recce, 1993; Skaggs et al., 1996) combine to produce an expansion of place fields (Blum & Abbott, 1996; Redish & Touretzky, 1998). These effects are dependent on NMDA integrity (Ekstrom, Meltzer, McNaughton, & Barnes, 2001; Shen, Barnes, McNaughton, Skaggs, & Weaver, 1997). The time-course of this increase in sharp-wave ripple activity should be similar to that of place field expansion. The ultimate test of this proposal would be to directly examine sharp-wave activity, the development of place field asymmetry, and the sequential firing of place cells during sharp waves (i.e. replay) on the multiple-T maze under normal and NMDA-impaired conditions.

4.1.3. Hippocampal inactivation should increase the number of errors made but should facilitate the development of path stereotypy

In the model, hippocampal replay provides an indirect practice signal which enables faster learning. As such, it speeds up the gross decision-making, thus, reducing errors faster (Fig. 3(A)). However, because the replay overlearns the actual experience, it slows the development of an optimized fine-motor performance path (Fig. 3(B)). These predictions can be tested by removing the hippocampal contribution through direct inactivation studies. Hippocampal inactivation should increase errors during learning (i.e. slow the correction of errors), but should speed up the development of path stereotypy.

4.2. Suggested anatomical instantiation

Most models of temporal difference reinforcement learning (TDRL) suggest that its underlying anatomical instantiation includes the basal ganglia (Barto, 1995; Dayan & Balleine, 2002; Doya, 1999). The strongest support for temporal difference learning comes from data relating dopamine and the value-prediction error signal δ (Montague, Dayan, & Sejnowski, 1996; Schultz, 1998; Waelti, Dickinson, & Schultz, 2001). Recent fMRI experiments have also found that the hemodynamic (BOLD) signal in the striatum correlates with changes in the predicted value function in humans making economic choices (O'Doherty, Dayan, Critchley, & Dolan, 2003; Seymour et al., 2004; Tanaka et al., 2004). Striatal recordings in rats and primates suggest that striatal cells represent the association of stimuli and other state parameters (such as location of the animal) and specific actions and sequences of actions (Graybiel, Aosaki, Flaherty, & Kimura, 1994; Itoh et al., 2003; Jog, Kubota, Connolly, Hillegaart, & Graybiel, 1999; Kermadi, Jurquet, Arzi, & Joseph, 1993; Kimura, 1986; Schmitzer-Torbert & Redish, 2004).

This is in contrast to hippocampal data in which hippocampal pyramidal cells only show sensitivity to actions when those actions are an important component of the decision process (Redish, 1999), that is, when the actions are part of the 'state' variable. For most rat navigation tasks, the key state variable is the location of the animal and thus, the usually-observed largest correlate of pyramidal cell activity is the location of the animal (the 'place field' of a 'place cell', O'Keefe & Dostrovsky, 1971; Redish, 1999, for review). The cases in which non-spatial correlates have been seen in rat hippocampal recordings are all cases in which the non-spatial correlates are important state variables in the decision process. For example, Eichenbaum, Kuperstein, Fagan, and Nagode (1987) found hippocampal firing tuned to groups of odors when animals were forced to make a decision based on group

membership of the odors (see Cohen & Eichenbaum, 1993, for review). Similarly, hippocampal cells are sensitive to match and non-match of samples when animals were forced to make a decision based on that match versus non-match (Cohen & Eichenbaum, 1993; Otto & Eichenbaum, 1992; Sakurai, 1990). This may also explain the cases in which prospective and retrospective encoding have been seen in hippocampal recordings (Ferbinteanu & Shapiro, 2003; Wood, Dudchenko, Robitsek, & Eichenbaum, 2000). In Wood et al. (2000), rats alternated between sides on a looped single-T-maze, thus, requiring a memory of which action had been previously taken—the previous action was a necessary part of the state variable in order to correctly decide on the subsequent course of action. In Ferbinteanu and Shapiro (2003), rewarded actions were switched in blocks, and thus, the blocks (and the current action) were critical components of the decision process.

In summary, we suggest that the hippocampus provides the distributed state component to the equation and the striatum associates that state component with efferent copy from motor and parietal cortices to calculate the $Q(s, a)$ variable. We suggest that hippocampal replay occurring during awake, rest periods may supply a practice signal to speed up learning in the striatum, analogous to the Dyna-Q family of algorithms in the TDRL literature (Sutton & Barto, 1998). Furthermore, the mechanisms used for the practice signal are derived from experimental studies on development of hippocampal cellular assemblies (Bi & Poo, 2001; Ekstrom et al., 2001; Harris, 2003; Mehta et al., 1997). While we use a random initiation of replay similar to other models (Redish, 1999; Redish & Touretzky, 1998; Shen & McNaughton, 1996), the biological source of replay remains unclear (Csicsvari, Hirase, Czurkó, & Buzsáki, 1999a, b). In the model as presented here, the action in the TDRL component was calculated from the change in state during the replay, however, replay also occurs in cortex (Qin, McNaughton, Skaggs, & Barnes, 1997) and is coordinated with hippocampal replay (Hoffman & McNaughton, 2002). It is possible that coordinated action signals could be supplied to the striatum from those cortical areas.

Consistent with this model, the subiculum provides a strong efferent projection to the ventral striatum (Heimer, Alheid, & Zaborszky, 1985; McGeorge & Faull, 1989; Witter & Groenewegen, 1990) and may be a means by which the state information reaches the striatum. Recent work by Pennartz, Lee, Verheul, Lipa, Barnes and McNaughton (2004) suggests that replay information occurring during sharp-wave ripples does propagate to the ventral striatum. It remains unknown how these signals operate in the ventral striatum.

This instantiation (hippocampus provides state information, while striatum associates the state with action) is similar to that of Mogenson, Jones, and Yim (1980), who first suggested that the ventral striatum serves as the motor output of the limbic system, and to the more specific models of Brown and Sharp (1995), and Foster, Morris, and Dayan (2000), both of whom used hippocampus to provide a state component as input to a temporal difference reinforcement learning model. Both Brown and Sharp (1995), and Foster et al. (2000) suggested that the TDRL component likely included the ventral striatum.

Recently, Ascoli and Samsonovich (2004) proposed that replay during sleep states could be used to find a shortest path through a sequence by, essentially, learning to short-cut loops within the sequence. Our hypothesis differs from that of Ascoli and Samsonovich in that we hypothesize an online process (occurring during awake states, within a session) and in that our replay only learns to emphasize already experienced paths. While we did not explicitly examine the ability of our model to learn short-cuts, most TDRL algorithms that implement practice or planning quickly modify their paths in order to adapt to new changes in the environment (Sutton & Barto, 1998). The two hypotheses (ours and Ascoli and Samsonovich's) are not incompatible, and may well both occur using the same sharp-wave replay mechanism.

4.3. A transfer of control between learning systems

Hippocampal replay improves performance early, but impairs performance late (Fig. 2). One possible solution to this situation is to have two interacting systems, one of which includes replay and the other of which does not. Experimental evidence supports the hypothesis of a transfer of control from a hippocampal-dependent system to a dorsal-striatal-dependent system with learning (Packard & McGaugh, 1996; Poldrack & Packard, 2003) as well as changes within the striatal system (from posterior to anterior striatum, Hikosaka et al., 1999; Yin & Knowlton, 2004; and from ventral to dorsal striatum, Ito, Dalley, Howes, Robbins, & Everitt, 2000; Ito, Dalley, Robbins, & Everitt, 2002). It may be that behavioral control changes from the quickly-learning, 'place' system to a slowly-learning 'response' system (Cohen & Eichenbaum, 1993; Hikosaka et al., 1999; O'Keefe & Nadel, 1978; Packard & Knowlton, 2002; Packard & McGaugh, 1996; Poldrack & Packard, 2003; Redish, 1999; Restle, 1957; Tolman, Ritchie, & Kalish, 1946) and reflects a transition from a highly variable, quickly-learning system (dependent on replay) to a finer-tuned, slowly-learning system (without replay).

Although similarities between temporal difference learning and stimulus-response patterns of behavior have been noted, a direct mapping from one to the other does not exist (Dayan & Balleine, 2002). Standard forms of temporal difference learning store only a single value function and are insufficient for describing many forms of animal behavior (Redish, Johnson, Jensen, & Jackson, 2005). The practice signal provides a method for quickly changing the value function and, consequently, behavior. The practice signal outlined here is a member of the family of planning signals used in many artificial intelligence algorithms for fast adaptation, particularly in novel obstacle and wayfinding problems (Sutton & Barto, 1998), and supplies the behavioral flexibility required for devaluation and other autonomous behaviors (Dickinson, 1985).

4.4. Open questions

4.4.1. What behavior is replayed?

The model presented here neglects the potential influences of cognitive or top-down processes to structure replay and instead uses experimentally observable cellular mechanisms to structure replay. This leaves open a critical question: if relatively simple cellular mechanisms mediate replay, at what rate does synaptic modification occur? The rate of synaptic change dictates the content of route replay and the practice signal; very fast synaptic modifications will result in the replay of recent episodes, while slow synaptic modifications will result in a more general averaged replay. Theories of hippocampal function suggest that synaptic modifications within the hippocampus, particularly within the CA3 recurrent collaterals, are very fast (Levy, 1996; Marr, 1971). Alternatively, slow modifications may result in a split in replay at T choice points, unless the network acts to maintain a consistent representation (as in an attractor network Doboli, Minai, & Best, 2000; Redish & Touretzky, 1997; Samsonovich & McNaughton, 1997; Tsodyks, 1999). Currently, experimental examinations of route replay have employed only highly practiced behaviors (Kudrimoti et al., 1999; Lee & Wilson, 2002; Louie & Wilson, 2001; Skaggs & McNaughton, 1996; Wilson & McNaughton, 1994) and leave open the question of whether route replay reflects recent behavioral episodes or more global averaged behavior.

Clearly, the content of the replayed practice signal directly influences the development of the value function and behavioral plasticity. The synaptic modification rates used in the current model were quite slow, providing a generalized average replay and relatively small changes in the value function. Faster synaptic modification rates lead to greater temporal specificity in replay and larger changes in the value function. In sum, the degree of plasticity within the value function directly affects the level of behavioral flexibility of the modeled animal.

4.4.2. Is replay prioritized?

Current models of replay (such as the one here) have generally hypothesized the initiation of replay as random (e.g. Redish & Touretzky, 1998), or related to the ratio of experience of the agent (e.g. Shen & McNaughton, 1996). However, computer science models of TDRL have shown that a prioritized-sweep provides for a significant improvement in learning speed (Sutton & Barto, 1998). Experimental studies of hippocampal replay have only focused on simple, hippocampal-independent, runway tasks (i.e. not requiring choices). It is not yet known whether the replayed patterns are selected at random or whether they are prioritized in some way.

The TDRL literature suggests that replay should be prioritized with a preference for state-transitions in which there was a large change in the value function, that is experiences in which the value-prediction error signal δ was large (Sutton & Barto, 1998). If replay were prioritized in this way, replay should begin nearest to the source of reward and follow the greatest change in the δ signal. Given the hypothesis that phasic dopamine carries the δ signal (Barto, 1995; Montague et al., 1996; Schultz, 1998; Schultz, Dayan, & Montague, 1997), this would also predict that dopamine should have an effect on prioritizing states for replay. Past research has shown that dopamine enhances early long term potentiation in CA1 (Otmakhova & Lisman, 1996, 1999) and that dopamine agonists enhance the stability of hippocampal pyramidal cell place fields while dopamine antagonists destabilize them (Kentros, Agnihotri, Streater, Hawkins, & Kandel, 2004). While it remains unclear whether phasic dopamine, corresponding to a δ signal, is the basis for these observed changes in place cell stability, one interesting possibility is that this dopamine signal is responsible for place field modulation by behaviorally relevant learning signals such as those seen by Moita, Rosis, Zhou, LeDoux, and Blair (2003).

5. Conclusions

Off-line replay of recently experienced states is known to speed up slow temporal difference learning algorithms. We suggest that hippocampal replay occurring during immobile awake states may provide just such an indirect practice signal. In a computer model of behavior on a multiple-T maze, developing asymmetric connections provide for replay of recent memories and speed up learning. While this replay speeds up the learning of correct choices on the mulitple-T maze, it produces a more variable path and slows the development of behavioral stereotypy. These predictions are directly testable with current technologies.

Acknowledgements

We thank Jadin Jackson, Zeb Kurth-Nelson, Beth Masimore, Neil Schmitzer-Torbert, and Giuseppe Cortese for helpful discussions and comments on the manuscript. This work was supported by NIH (MH68029) and by fellowships from 3M and from the Center for Cognitive Sciences (grant number T32HD007151).

References

Ascoli, G. A., & Samsonovich, V. (2004). Connectionist model of the hippocampus suggesting a new link between episodic memory and spatial navigation. *Society for Neuroscience Abstracts, Program number* , 66714.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 215–232). Cambridge MA: MIT Press.

Bi, G., & Poo, M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. *Annual Review of Neuroscience*, *24*(1), 139–166.

Blum, K. I., & Abbott, F. (1996). A model of spatial map formation in the hippocampus of the rat. *Neural Computation*, *8*(1), 85–93.

Brown, M. A., & Sharp, E. (1995). Simulation of spatial learning in the Morris water maze by a neural network model of the hippocampal formation and nucleus accumbens. *Hippocampus*, *5*(3), 171–188.

Buzsáki, G. (1989). Two-stage model of memory trace formation: A role for "noisy" brain states. *Neuroscience*, *31*(3), 551–570.

Buzsáki, G., Leung, L. W., & Vanderwolf, H. (1983). Cellular bases of hippocampal EEG in the behaving rat. *Brain Research*, *287*(2), 139–171.

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

Csicsvari, J., Hirase, H., Czurkó, A., & Buzsaki, G. (1999a). Fast network oscillations in the hippocampal CA1 region of the behaving rat. *Journal of Neuroscience*, *19*(RC20), 1–4.

Csicsvari, J., Hirase, H., Czurkó, A., & Buzsaki, G. (1999b). Oscillatory coupling of hippocampal pyramidal cells and interneurons in the behaving rat. *Journal of Neuroscience*, *1*, 274–287.

Csicsvari, J., Hirase, H., Mamiya, A., & Buzsaki, G. (2000). Ensemble patterns of hippocampal CA3-CA1 neurons during sharp wave-associated population events. *Neuron*, *28*, 585–594.

N. D. Daw. Reinforcement learning models of the dopamine system and their behavioral implications. PhD thesis, Carnegie Mellon University, 2003.

Dayan, P., & Balleine, W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, *36*, 285–298.

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society, London B*, *308*, 67–78.

Doboli, S., Minai, A. A., & Best, J. (2000). Latent attractors: a model for context-dependent place representations in the hippocampus. *Neural Computation*, *12*(5), 1009–1043.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia, and the cerebral cortex? *Neural networks*, *12*, 961–974.

Eichenbaum, H., Kuperstein, M., & Fagan, A. (1987). Cue-sampling and goal-approach correlates of hippocampal unit activity in rats performing an odor-discrimination task. *Journal of Neuroscience*, *7*(3), 716–732.

Ekstrom, A. D., Meltzer, J., McNaughton, B. L., & Barnes, A. (2001). NMDA receptor antagonism blocks experience-dependent expansion of hippocampal "place fields". *Neuron*, *31*, 631–638.

Ferbinteanu, J., & Shapiro, M. L. (2003). Prospective and retrospective memory coding in the hippocampus. *Neuron*, *40*(6), 1227–1239.

Foster, D. J., & Morris, R. G. M. (2000). A model of hippocampally dependent navigation using the temporal difference learning rule. *Hippocampus*, *10*, 1–6.

Graybiel, A. M., Aosaki, T., & Flaherty, A. W. (1994). The basal ganglia and adaptive motor control. *Science*, *265*(5180), 1826–1831.

Harris, K. D., Csicsvari, J., Hirase, H., Dragoi, G., & Buzsaki, G. (2003). Organization of cell assemblies in the hippocampus. *Nature*, *424*(6948), 552–556.

Heimer, L., & Alheid, G. F. (1985). Basal ganglia. In G. Paxinos (Ed.), *The Rat Nervous System*1st ed. (pp. 37–86). Orlando FL: Academic Press.

Hikosaka, O., Nakahara, H., Rand, M. K., Sakai, K., Lu, X., Nakamura, K., & Miyachi, S. (1999). Parallel neural networks for learning sequential procedures. *Trends in Neurosciences*, *22*(10), 464–471.

Hoffman, K. L., & McNaughton, L. (2002). Coordinated Reactivation of Distributed Memory Traces in Primate Neocortex. *Science*, *297*(5589), 2070–2073.

Ito, R., Dalley, J. W., Howes, S. R., Robbins, T. W., & Everitt, J. (2000). Dissociation in conditioned dopamine release in the nucleus accumbens core and shell in response to cocaine cues and during cocaine-seeking behavior in rats. *Journal of Neuroscience*, *20*(19), 7489–7495.

Ito, R., Dalley, J. W., Robbins, T. W., & Everitt, J. (2002). Dopamine release in the dorsal striatum during cocaine-seeking behavior under the control of a drug-associated cue. *Journal of Neuroscience*, *22*(14), 6247–6253.

Itoh, H., Nakahara, H., Hikosaka, O., Kawagoe, R., & Takikawa, Y. (2003). Correlation of primate caudate neural activity and saccade parameters in reward-oriented behavior. *Journal of Neurophysiology*, *89*(4), 1774–1783.

Jog, M. S., Kubota, Y., Connolly, C. I., Hillegaart, V., & Graybiel, M. (1999). Building neural representations of habits. *Science*, *286*, 1746–1749.

Kentros, Clifford G., Agnihotri, Naveen T., Streater, Samantha, Hawkins, Robert D., & Kandel, Eric R. (2004). Increased attention to spatial context increases both place field stability and spatial memory. *Neuron*, *42*, 283–295.

Kermadi, I., Jurquet, Y., Arzi, M., & Joseph, P. (1993). Neural activity in the caudate nucleus of monkeys during spatial sequencing. *Experimental Brain Research*, *94*, 352–356.

Kimura, M. (1986). The role of primate putamen neurons in the association of sensory stimuli with movement. *Neuroscience Research*, *3*(5), 436–443.

Kudrimoti, H. S., Barnes, C. A., & McNaughton, L. (1999). Reactivation of hippocampal cell assemblies: Effects of behavioral state, experience, and EEG dynamics. *Journal of Neuroscience*, *19*(10), 4090–4101.

Lee, A. K., & Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, *36*, 1183–1194.

Lee, I., Rao, G., & Knierim, J. (2004). A double dissociation between hippocampal subfields: Differential time course of ca3 and ca1 place cells for processing changed environments. *Neuron*, *42*, 803–815.

Levy, W. B. (1996). A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus*, *6*(6), 579–591.

Levy, W. B., & Steward, O. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, *8*(4), 791–797.

Louie, K., & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during Rapid Eye Movement sleep. *Neuron*, *29*, 145–156.

Marr, D. (1971). Simple memory: A theory of archicortex. *Philosophical Transactions of the Royal Society of London*, *262*(841), 23–81.

McGeorge, A. J., & Faull, L. (1989). The organization of the projection from the cerebral cortex to the striatum in the rat. *Neuroscience*, *29*(3), 503–537.

McNaughton, B. L. (1983). Associative properties of hippocampal long term potentiation. In W. Seifert (Ed.), *Neurobiology of the Hippocampus* (pp. 433–447). New York, NY: Academic Press.

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

Mehta, M. R., Quirk, M. C., & Wilson, A. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, *25*, 707–715.

Mogenson, G. J., Jones, D. L., & Yim, Y. (1980). From motivation to action: Functional interface between the limbic system and the motor system. *Progress in Neurobiology*, *14*, 69–97.

Moita, M. A., Rosis, S., Zhou, Y., LeDoux, J. E., & Blair, T. (2003). Hippocampal place cells acquire location-specific responses to the conditioned stimulus during auditory fear conditioning. *Neuron*, *37*(3), 485–497.

Montague, P. R., Dayan, P., & Sejnowski, J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*(5), 1936–1947.

Nádasdy, Z., Hirase, H., Czurkó, A., Csicsvari, J., & Buzsáki, G. (1999). Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience*, *19*(2), 9497–9507.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.

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

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

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

Otmakhova, N. A., & Lisman, J. E. (1999). Dopamine selectively inhibits the direct cortical pathway to the CA1 hippocampal region. *Journal of Neuroscience*, *19*(4), 1437–1445.

Otmakhova, N. A., & Lisman, J. E. (1996). D1/D5 Dopamine Receptor Activation Increases the Magnitude of Early Long-Term Potentiation at CA1 Hippocampal Synapses. *Journal of Neuroscience*, *16*(23), 7478–7486.

Otto, T., & Eichenbaum, H. (1992). Neuronal activity in the hippocampus during delayed non-match to sample performance in rats: Evidence for hippocampal processing in recognition memory. *Hippocampus*, *2*(3), 323–334.

Packard, M. G., & Knowlton, J. (2002). Learning and memory functions of the basal ganglia. *Annual Reviews Neuroscience*, *25*(1), 563–593.

Packard, M. G., & McGaugh, L. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory*, *65*, 65–72.

Pavlides, 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*(8), 2907–2918.

Pennartz, C. M. A., Lee, E., Verheul, J., Lipa, P., Barnes, C. A., & McNaughton, L. (2004). The ventral striatum in off-line processing: ensemble reactivation during sleep and modulation by hippocampal ripples. *Journal of Neuroscience*, *24*(29), 6446–6456.

Poldrack, R. A., & Packard, G. (2003). Competition among multiple memory systems: Converging evidence from animal and human studies. *Neuropsychologia*, *41*, 245–251.

Qin, Y. L., McNaughton, B. L., Skaggs, E., & Barnes, A. (1997). Memory reprocessing in corticocortical and hippocampocortical neuronal ensembles. *Philosophical Transactions of the Royal Society, London*, *B352*(1360), 1525–1533.

Redish, A. D. (1999). *Beyond the Cognitive Map: From Place Cells to Episodic Memory*. Cambridge MA: MIT Press.

Redish, A. D., Johnson, A., & Jensen, S. (2005). Latent learning requires multiple value functions with temporal difference reinforcement learning. *Computational Neurosciences Abstracts* .

Redish, A. D., & Touretzky, S. (1997). Cognitive maps beyond the hippocampus. *Hippocampus*, *7*(1), 15–35.

Redish, A. D., & Touretzky, S. (1998). The role of the hippocampus in solving the Morris water maze. *Neural Computation*, *10*(1), 73–111.

Restle, F. (1957). Discrimination of cues in mazes: A resolution of the 'place-vs-response' question. *Psychological Review*, *64*, 217–228.

Sakurai, Y. (1990). Hippocampal cells have behavioral correlates during the performance of an auditory working memory task in the rat. *Behavioral Neuroscience*, *104*(2), 253–263.

Samsonovich, A. V., & McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, *17*(15), 5900–5920.

Schmitzer-Torbert, N. C., & Redish, A. D. (2002). Development of path stereotypy in a single day in rats on a multiple-T maze. *Archives Italiennes de Biologie*, *140*, 295–301.

Schmitzer-Torbert, N. C., & Redish, A. D. (2004). Neuronal activity in the rodent dorsal striatum in sequential navigation: Separation of spatial and reward responses on the multiple-T task. *Journal of Neurophysiology*, *91*(5), 2259–2272.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1–27.

Schultz, W., & Dayan, P. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., & Frackowlak, S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, *429*, 664–667.

Shen, B., & McNaughton, B. L. (1996). Modeling the spontaneous reactivation of experience-specific hippocampal cell assembles during sleep. *Hippocampus*, *6*(6), 685–693.

Shen, J., Barnes, C. A., McNaughton, B. L., Skaggs, W. E., & Weaver, L. (1997). The effect of aging on experience-dependent plasticity of hippocampal place cells. *Journal of Neuroscience*, *17*(17), 6769–6782.

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, A. (1996). Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus*, *6*(2), 149–173.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. *In Proceedings of the Seventh International Conference on Machine Learning*, 216–224.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An introduction*. Cambridge MA: MIT Press.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, *7*, 887–893.

Tolman, E. C., Ritchie, B. F., & Kalish, D. (1946). Studies in spatial learning. II. Place learning versus response learning. *Journal of Experimental Psychology*, *36*, 221–229.

Tsodyks, M. (1999). Attractor network models of spatial maps in hippocampus. *Hippocampus*, *9*(4), 481–489.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature* . doi10.1038/35083500.

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

Witter, M. P., & Groenewegen, H. J. (1990). The subiculum: Cytoarchitectonically a simple structure but hodologically complex. In J. Storm-Mathisen, J. Zimmer, & O.P Ottersen, *Understanding the Brain through the Hippocampus. Progress in Brain Research* (vol. 83). New York: Elsevier.

Wood, E. R., Dudchenko, P. A., & Robitsek, R. J. (2000). Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron*, *27*, 623–633.

Yin, H. H., & Knowlton, B. J. (2004). Contributions of Striatal Subregions to Place and Response Learning. *Learn. Mem.*, *11*(4), 459–463.

Ylinen, A., Bragin, A., Nadasdy, Z., Jando, G., & Buzsaki, G. (1995). Sharp wave-associated high-frequency oscillation (200 Hz) in the intact hippocampus: Network and intracellular mechanisms. *Journal of Neuroscience*, *15*(1), 30–46.