Cortical hierarchies, sleep, and the extraction of knowledge from memory

Bruce L. McNaughton

Canadian Centre for Behavioural Neuroscience, Department of Neuroscience, University of Lethbridge, Lethbridge AB, Canada

ABSTRACT

Odin the Allfather had in his service two great ravens. These ravens’ names were Hugin (Thought) and Munin (Memory) and every morning at dawn they would fly off over Midgard (the world) in search of news and information to learn more about humans and their activities. At sundown, they would return to Odin where they would perch one on each of Odin’s shoulders, and whisper into his ears all that they had seen and heard.

Experience, stored in the brain as memory, is the raw material for intelligence and thought. It has been suggested that at sundown (i.e., during sleep) the brain adjusts its own synaptic matrix to enable adaptive responses to future events by a process of gradient descent optimization, involving repeated reactivations of recent and older memories and gradual adjustment of the synaptic weights. Memory retrieval, thought, and the generation of adaptive behavioral responses involve globally coordinated trajectories through the neuronal state-space, mediated by appropriate synaptic linkages. Artificial neural networks designed to implement even the most rudimentary forms of memory and knowledge extraction and adaptive behavior incorporate massively and symmetrically interconnected nodes; yet, in the cerebral cortex, the probability of a synaptic connection between any two arbitrarily chosen cells is on the order of $10^{-6}$, i.e., so close to zero that a naive modeler might neglect this parameter altogether. The probability of a symmetric connection is even smaller ($10^{-12}$). How then, are thought and memory even possible? The solution appears to have been in the evolution of a modular, hierarchical cortical architecture, in which the modules are internally highly connected but only weakly interconnected with other modules. Appropriate inter-modular linkages are mediated indirectly via common linkages with higher level modules collectively known as association cortex. The hippocampal formation in the temporal lobe is the highest level of association cortex. It generates sequentially coupled patterns unique to the location and content of experience, but which do not contain the actual stored data. Rather, the patterns serve as pointers or ‘links’ to the data. Spontaneous reactivation of these linking patterns during sleep may enable the retrieval of recent sequences of experience stored in the lower levels of the cortex and the gradual extraction of knowledge from them. In this essay I explore these ideas, their implications, and the neuroscientific evidence for them.

© 2009 Published by Elsevier B.V.
1. Introduction

The idea that the brain makes use of ‘off-line’ periods, such as sleep, to ‘sort-out’ and consolidate memories has a long history, probably as long as humans have wondered about the meaning of their dreams. Computationally principled suggestions for why such post-experience reprocessing should be necessary, however, have a much more modern history [1, 15,16], and were at least partly inspired by the phenomenon of temporally limited retrograde amnesia, the so-called Ribot gradient, following damage to the hippocampus and surrounding cortex. Such damage leaves humans cognitively normal except for a loss of recently acquired memories and a profound inability to form new ones. What follows is a review of the theoretical considerations for why the brain might require an active reprocessing of memories during periods when it is relatively ‘disconnected’ from external input, and a summary of current understanding of the phenomenon based on neurophysiological investigations in animals.

The fundamental necessity for off-line reprocessing can be understood from the most basic model for associative memory: a network of neurons containing a primary (e.g., sensory) input that determines the output pattern, and an association input pathway that is exhaustively (i.e., all-to-all) connected via Hebbian, associatively modifiable synapses. In the simplest scheme, the modifiable synapses have binary weights that are initially 0 and convert permanently to 1 following the principle of association outlined by Hebb [6]: correlated pre- and post-synaptic activity. Input vectors (also binary) are of a fixed length, and retrieval of the paired-associate of a given pattern (or pattern fragment) on the association input is accomplished by summing the net synaptic current to each cell (forming a dot product of the input with the synaptic matrix) and performing integer division by the number of active input axons. The latter operation, which captures the all-or-none nature of neuronal impulse communication, ensures that only those output neurons fire which contain a maximum proportion of already potentiated synapses in the current input pattern. Marr proposed that the division is accomplished by inhibitory interneurons which shunt the membrane resistance of the spike generating region, thus approximating a division according to Ohm’s law. The physiological and anatomical properties of at least one known class of inhibitory interneurons are surprisingly consistent with this hypothesis [20]. Similar fundamental principles apply to the recurrent networks which can implement auto-association and pattern completion, as well as a simple form of sequence encoding.

The three primary factors that determine the storage capacity (in terms of the number of arbitrary patterns) of such networks are well understood [16,5,28]: connectivity density, coding sparsity, and ‘orthogonality’. Optimizing for the latter two constraints involves redundancy reduction and the generation of efficient feature detectors. It is the first of these constraints that provides a basis for understanding both the modular and hierarchical organization of cortical association areas, and the fundamental necessity for a memory consolidation phase involving off-line reprocessing. To be clear about the basis of this constraint, however, it is necessary to make a theoretical distinction between two types of modifiable synapse: physical synapses and virtual synapses. Physical synapses can be identified at the ultrastructural level by a characteristic, morphological connection between two neurons, consisting of distinct pre- and post-synaptic elements. The weight \( w_{ij} \) of a physical synapse from neuron \( j \) onto neuron \( i \) may be zero or non-zero (up to some maximum value). In cerebral cortex, the weights of synapses involved in storage of information are almost universally positive valued (“excitatory”); this is not the case everywhere in the brain. Non-zero physical synapses almost universally propagate information in one direction. A ‘virtual synapse’ has a weight of zero and no physically identifiable contact between pre- and post-synaptic elements. It can be converted to a physical synapse with non-zero weight with some probability that depends on complex factors, of which physical proximity and correlated activity of the pre- and post-synaptic elements are of main concern for a theory of association. Because the pre- and post-synaptic elements are themselves dynamic structures, exhibiting possible growth and retraction, the probability of conversion from virtual to physical is also dynamic: the elements have to be at the right place at the right time. Physical synapses can also be converted to virtual synapses, possibly as a result of anti-correlated activity between the two neurons, or by correlated activity in the wrong temporal order (i.e., post-synaptic activity preceeding pre-synaptic). The weights of physical synapses can be positively or negatively modified on a rapid time-scale (seconds to minutes), whereas the modification of a virtual synapse is stochastic. The pre- and post-synaptic elements need to be in close proximity at the time of a conjunction of activity of the two neurons. Hence, rearrangement of physical synaptic connectivity is slow on average (hours to days), and typically requires multiple conjunction events before one happens to occur at a time when the physical processes of the neurons involved happen to be appropriately aligned.
Fig. 1. Hierarchical organization of cortical modules. In the cerebral cortex there are about $10^{10}$ cells but each can send and receive only about $10^4$ physical synapses. The connection matrix is therefore extremely sparse on average, much too sparse to enable rapid, arbitrary associations by modifying existing synaptic connections. This is illustrated diagrammatically in (a), which is a hypothetical $n \times n$ matrix in which the neurons are numbered $1, \ldots, n$ and the non-zero elements, represented by black dots, are physical connections. Creating modules of about $10^5$ cells, would enable complete, within-module connectivity (black squares), at the expense of not being able to form associations among the items stored in different modules (b). A possible solution is a hierarchical arrangement of densely connected modules, linked vertically by sparse, reciprocal, rapidly modifiable connections which create pointers to items stored in lower level modules, and horizontally by connections that can be gradually rearranged to capture long-term statistical regularities in the overall mnemonic experience. The topology of such a hypothetical hierarchy is illustrated in (d) (where the solid, curved arrows represent modifiable within-module physical connectivity, the large dashes with arrows represent modifiable physical connections and the small dashes with arrows represent rearrangeable configurations of virtual and physical connections). The corresponding physical connection matrix is illustrated in (c) (where the shading density represents physical connection density). According to this view, episodic memory (memory for specific experiences) is encoded by rapid indirect associations among modules mediated by the vertical connections, and semantic memory (knowledge) is encoded by reconfiguring the horizontal connections. The latter encoding is assumed to involve a slow, gradient descent type of adaptation process that can occur during wakefulness if the stored events are repeated many times but more likely involves spontaneous retrieval of the pointers during sleep, leading to 'virtual' repetitions of the stored experiences. In the mammalian brain, the highest level module can be identified as the hippocampal formation, loss of which seriously impairs the acquisition of new memories and semantic knowledge, but leaves existing knowledge remarkably intact.

2. Modular and hierarchical organization of the mammalian cortex, and the memory indexing theory

The mammalian cortical physical connection matrix ($W$) is so sparse on average that making rapid linkages between arbitrary pairs $(i, j)$ of neurons by modifying existing physical synapses would seem impossible (Fig. 1a). Yet making such arbitrary linkages rapidly and extracting knowledge about the world from these stored experiences seems to be a major hallmark of intelligence. How does the brain accomplish this? Each cortical principal cell sends and receives about $10^4$ physical connections (synapses). Thus, organizing the cortex into modules of about 10000 neurons would permit all-to-all, symmetric connectivity, but at the expense of there not being enough connections left over to support associations among items stored in different modules (Fig. 1b). Evolution appears to have devised a solution by creating a hierarchical modular architecture (Fig. 1c, d). Modules at different hierarchical levels are linked by fixed bidirectional connections whose weights are rapidly modifiable (dashed arrows). Modules at the same hierarchical levels are linked by sparse connections (dotted arrows) that can be physically rearranged (a biologically slow process). Anatomically, the topmost module can be identified with the hippocampal formation of the mammalian brain. A rather small amount of information ascending from each lower level module is sufficient to create a unique pattern in the next higher module. If two or more lower level modules receive the same top–down input, and store it with their current internal representation via the dense, modifiable intrinsic connections (curved arrows), then the top–down pattern can later serve as an indirect associative link, or 'index' for the items stored in the lower level modules. These top–down connections thus permit a global pattern stored in the cortex to be completed from a fragment of the initial input without the need for exhaustive connectivity. According to this theory, the horizontal connections play a different role. As first suggested by Marr [15], repeated reactivation of stored items while new
inputs are excluded may allow the horizontal connections to reorganize in a manner that optimizes the cortex to respond adaptively to future inputs, i.e., to recode the data in a smoothly generalized categorical structure. Thus, “knowledge” may be extracted from experience. This general theory provides an explanation for a central fact of neuropsychology: damage to the topmost module (the hippocampal formation) results in a severe impairment of acquisition of new memories and semantic knowledge, but leaves the brain’s existing knowledge base and cognitive function surprisingly intact. The theory leads to a major prediction: during sleep, when external input is largely gated out, there should be a process of spontaneous retrieval of recently stored items, and such retrieval should lead to the gradual rearrangement of the cortical horizontal (and possibly within module) connections.

The concept of a module in the indirect association theory is left intentionally vague, but incorporates the idea of a set of cells that share common or related inputs from lower levels, have substantially more synaptic interactions with each other than with cells of other modules, and subserve a common function. It is assumed that the intrinsic connections are Hebb-modifiable such that each module is capable of storing information within its input domain auto-associatively. These assumptions could apply to different sensory processing areas (e.g., somatosensory, visual, etc.), but, in the present context, are better suited to submodules within an area such as the cortical ‘barrel’ fields in rodent S1 (Fig. 2a), or the distinct cell clusters found in layer II of the cortical association areas (Fig. 2b). Such modules range from about 100 to about 300 microns in diameter and encompass approximately the theoretically appropriate numbers of neurons. A particularly cogent example is the demonstration by Tanifuji and colleagues (e.g., [29]) that images of complex objects elicit neural activity in sets of “feature columns” in area TE of the temporal cortex (Fig. 2c). It is often possible to parse an image into components that activate different subsets of the original set. Fiber tracing studies are revealing rich and specific patterns of horizontal connections among modules in higher association areas (Fig. 2d).

3. Sparsity of coding maximizes event memory capacity

What one means by capacity depends both on whether one is talking about the number of separate memories or the maximum information stored, and on what sort of criterion one is willing to accept for the accuracy of recall. In either case, for a completely connected recurrent neural network model of binary-state neurons with binary synapses that undergo permanent modification from 0 to 1 according to Hebbian association of pre- and post-synaptic conjunction, a simple argument sets an upper bound on capacity: the information content (entropy) of the synaptic weight distribution is maximal when exactly half of the binary synapses have been “potentiated” (from $W = 0$ to $W = 1$). This conclusion follows from the fact that the number of possible configurations (i.e., combinations) of $n$ binary items is greatest when half are 0 and half...
Fig. 3. The hippocampus generates output codes unique to location and content of experience, which are propagated back down the cortical modular hierarchy. Hippocampal neurons fire in an interesting, spatially selective fashion. For example, when a rat forages for food in a walled box, say 75 × 75 cm, with specific patterns of visual stimuli on the walls, a small fraction (about 20%) of cells will fire in one or a few specific locations. Changing the location of the box, say to a different room, completely changes which cells are active; however, changing the sensory features of the box without changing its locations does not change which cells are active or where in the box they fire, but it does dramatically change the relative firing rates of these cells. In the left columns of (a) and (b), the color maps represent the spatial firing rate maps for eight different hippocampal neurons, each map scaled to the maximal firing rate of the cell. The middle panels show the corresponding rate maps for the same cells plotted at the same scale after either a change in sensory features of the box (constant place-variable cues) or a change in the location of the box (constant cues-variable place). For these particular cells, firing rates were drastically reduced (rates of other cells, not shown, increased); however, rescaling each of the central maps to their specific maxima shows that in the constant place condition, the location of firing was unchanged, whereas, in the constant cue-variable place condition, firing locations relative to the box were radically altered. This implies that the firing rate vectors for any given location in the constant place-variable cue condition are correlated, whereas they are uncorrelated in the opposite condition ((c) and (d)). This effect can be represented by a diagram in which each cell is ordered according to the location of its firing maximum. A hill of activity in this space represents the current hippocampal output vector. Fluctuations in the hill are not noise but reflect the influence of external inputs to the animal while it is at a given location. (Data in (a) and (b) from [10].)

\[ P(W_{ij} = 0) = (1 - \alpha^2)^m \]  

The upper limit for saturation is thus found by setting \( P = 0.5 \) and solving for \( m \),

\[ m_{\text{max}} \leq \ln(0.5) / \ln(1 - \alpha^2) \]  

Notice that the number of units in the system (\( n \)) does not enter into this calculation, but does set a lower limit on \( \alpha \).

The foregoing considerations suggest that, in the brain, there is a premium on minimizing the number of units that are involved in the representation of each stored event. This general principal is known as 'sparse coding', and its utility for maximizing event storage has been widely recognized. Clearly, if a representation of an experience is to be made sparse without loss of detail, then the encoding scheme must be "efficient". Making a neural code efficient means classification, removal of redundancy, or feature extraction, all of which are synonymous with extraction of knowledge.

There appears to be a general principle in the brain that encoding becomes increasingly sparse as one ascends the hierarchy of sensory association areas. At the highest level, in some parts of the hippocampus, \( \alpha \) is on the order of 0.005 or less. Conversely, in output structures like primary motor cortex, where there is a premium on high resolution and smooth generalization, but where rapid information storage is probably minimal, coding is almost fully distributed (\( \alpha \sim 0.5 \)).
4. Hippocampus generates sparse linking codes unique to location and content of experience

The hippocampus and adjacent structures represent an evolutionarily older part of the cortex, often referred to as “archicortex”, or “paleocortex” to distinguish them from its more recently evolved regions, which are called “neocortex”. From the perspective of connectivity, however, the hippocampal formation represents the highest level of association cortex, receiving highly processed, multimodal information indirectly from the entire neocortex. Its position in the hierarchy of cortical association areas is indicated by the top of element of Fig. 1d. It is generally believed that hippocampal outflow to neocortex may bind together components of each unique memory that are stored in neocortex in distributed form (e.g., [16,2,18,23,8]). Hippocampal neurons fire in spatially specific patterns known as ‘place fields’, which are approximately Gaussian ‘hot-spots’ of elevated firing in an environment and could serve as radial basis functions for location if integrated by downstream neurons. At any moment, the output vector of the hippocampus is very sparse, involving activity in only a few percent of neurons. The preponderance of evidence is that the place field of a given cell is determined primarily by a complex intrinsic mechanism that takes into account which cells were active at a given time, and the distance and direction that the animal has moved since that time. This operation is known as ‘path integration’ (see [22] for review). More recently, it has been shown that the amplitude of place fields is strongly modulated by both the external events currently taking place and by what internal activity is ongoing in other brain structures (e.g., current plans, behavioral schemata, or items stored in working memory, etc.). See Fig. 3.

5. Memory and knowledge are represented in deep layers of cortex, context in superficial layers

Interestingly, although the activity of the hippocampal network itself is very sparse, a data compression operation appears to take place at an output relay known as the subiculum. Here, high dimensional, sparse vectors seem to be compressed into lower dimensional, non-sparse outputs, prior to distribution to the hierarchically lower levels of association cortex. There is evidence to support the speculation that this compressed code is then ‘unzipped’ in the superficial layers of the cortex. Hippocampal output appears to modulate activity in the superficial layers of the neocortex in a manner that reflects the spatial context of a given memory episode [3]. In contrast, the deep layers of the cortex appear to represent information in a context-independent manner. For example, in rat parietal cortex, information about egocentric movement is represented in a context-independent manner in the deep layers, whereas superficial parietal cortex represents conjunctions between egocentric movement and spatial context. Likewise, deep gustatory cortex represents pure taste information, whereas in superficial gustatory cortex, taste and spatial context are conjoined. This sort of laminar architecture could support a mechanism for associative retrieval of context-free data using linking codes reflecting context (i.e., hippocampal outflow).

6. Sequence storage, the synaptic symmetry problem, and its remarkable solution

Episodes of experience unfold in time. It is therefore unlikely that memories are fixed patterns of neural activity, but rather specific trajectories through the neuronal state-space. Donald Hebb [6] speculated that the temporal aspect of memories was due to the formation of what he called “phase sequences”, which would be collections of neurons representing discrete items (“cell assemblies”) linked together by asymmetric connections formed as a consequence of the fundamental asymmetry in his proposed mechanism for synaptic association; viz. that, for synaptic strengthening to occur, the presynaptic neuron must fire just before the post-synaptic neuron. More recent thought in the field of sequence encoding theory identifies an important constraint on this idea. For effective sequence encoding, the neural representations of the items in the sequence must be relatively uncorrelated, or else the connections among the cells encoding the sequence will tend to become symmetric and, during recall, the network will ‘get stuck’ in a unitary ‘cell assembly’, rather than progressing through a ‘phase sequence’ of cell assemblies. The time-scale over which the neural patterns must be uncorrelated depends on the biophysical dynamics of the association mechanism, whose time-constant appears to be on the order of a few tens of milliseconds; however, experience seldom changes substantially over such a rapid time-scale. For example, in the rat, a typical hippocampal place field occupies a space of at least 20–30 cm, and rats typically run about 20 cm/sec. In 100 msec, therefore, the rat covers less than 1/10 of the place field, which means that population activity patterns separated by 100 msec will be very highly correlated and sequence retrieval would be difficult at best.

The hippocampus exhibits a global oscillation of its net cellular activity in the range of about 6–8 Hz, while animals are actively interacting with their environments and this shows up as a corresponding rhythm in the gross electrical potential known as the theta rhythm; however, the behavior of individual neurons is more complex. Typically it takes about 10–20 cycles of the theta rhythm for a rat to traverse a place field of a given cell, which occupies about 25 cm. As the animal passes through the field, that cell systematically varies its firing phase relative to the net population oscillation as a function of spatial location (Fig. 4). The first spikes occur late in the theta cycle, but, as the animal traverses the field, firing shifts earlier in phase so that, by the time the rat leaves the field, the firing has ‘precessed’ in phase through almost 360°, but never more. As explained in Fig. 4, this phase precession allows the index codes generated by the hippocampus for sequences of experiences to become asymmetrically linked in such a way that activating the set of neurons representing the beginning of the sequence would lead to recall of the complete sequence. One prediction of the generation of such asymmetrical memory structures is that, if a rat runs repeatedly through a sequence of locations, one would observe activity representing a given location to begin to appear before the rat actually gets there. This has been verified experimentally [17].
7. Coherent reactivation of recent memory traces in hippocampus and cortex

What does it mean to say that a system expresses spontaneous memory reactivation? In the most general terms, it must mean that if, at some time, external input to the system drives it into some region of its state-space, and this event results in an enhanced probability of the system revisiting that region at a future time, in the absence of the external input, then the system exhibits memory reactivation. One useful and experimentally measurable index of the state of the brain is simply a list of the impulse frequency of each neuron over an arbitrary (short) time interval. Actually, it is currently technically possible to study only a small subspace of the brain's state-space, i.e., a sample of less than a few hundred cells. Suppose we record from $N$ cells while the animal sleeps ($S_1$). We then allow it to perform some arbitrary behavior in some arbitrary test environment ($B$). After some time, we remove the animal from the test environment allow it to go to sleep again ($S_2$). Throughout this time we will observe that the $N$-dimensional state vector changes from moment to moment ($Fig. 5$). We can define the state-space occupancy distribution as the relative amount of time the system spends in each of its possible states. If knowledge of the occupancy distribution during the behavioral epoch provides more information about the distribution in the subsequent sleep than it does about the distribution in prior sleep, we can say that memory trace reactivation is occurring during the second sleep episode. Practically, this can be measured by computing, for each epoch, the $N \times N$ correlation matrix for the firing rate time series of the $N$ cells, and assessing how much additional variance in the correlations during $S_2$ is 'explained' by the observed values in $B$, after taking into consideration $S_1$ [9]. Using this approach, referred to as "explained variance", a large number of studies lead jointly to the conclusion that memory reactivation occurs during sleep, that it occurs in widespread regions of the brain, including hippocampus, neocortex and some subcortical structures, and that it is coherent over multiple modules or regions. By coherent, I mean that the sub-patterns retrieved in different areas are components of the same original global pattern. Perhaps contrary to intuition, this playback does not occur during periods of sleep associated with dreams ("REM" sleep) but during periods called slow-wave sleep, which are not associated with reports of dream activity when subjects are wakened from them.

8. Memory playback on fast forward and speculations about the speed of thought

The 'explained variance' measure of memory reactivation is robust because it is completely insensitive to the temporal order of the state vectors, and also relatively insensitive to possible differences between waking and sleep in how quickly the vectors may change over time. It turns out that this insensitivity is largely responsible for the early success of the method in detecting memory trace reactivation. Early attempts to detect the reactivation of memory sequences in the brain used a method called 'template matching' (see Figs. 5a and 6a). In this method, a segment of multi-neuronal firing sequences during waking behavior is selected and convolved with the corresponding data from subsequent sleep. The expectation is that, if the behavioral sequence is faithfully replayed, then one should detect a high incidence of matches [11,27]. Template matching is highly sensitive to both permutations of temporal order of recorded state vectors and fluctuations in playback time-scale (Figs. 5b–d, 6) and these properties underlie the limited success of early attempts to detect sequence reactivation during sleep.

It turns out that memory replay in the cortex during sleep is a process with complex dynamics; however, the following analogy may provide a basic understanding. Suppose that each day we were to make a video record the day's events. We create multiple copies of the video file and then subdivide each copy into segments of random length by a Poisson process, such that the length distribution is exponential with a mean of about 10 sec. We add these clips to the pool of clips remaining from previous days, and then select segments by random draw and play them back on 5–10× fast forward. If the
Fig. 5. Modern neurophysiological methods enable recording signals from up to about 200 brain cells in behaving animals, via specially designed surgically implanted probes. The data are reduced to the spike time-series for each cell, which can be represented by an $N \times T$ matrix ($Q$), where $N$ is the number of cells and $T$ is time. In (a), spikes from about 50 cells are represented over about 10 seconds. Time can be binned into intervals and the spike events summed over intervals to give firing rate column vectors, which can be thought of as representing points in an $N$-dimensional state-space. Column wise correlations reflect proximity in state-space. The row-wise correlation matrix is a reflection of the state-space occupancy distribution, or the relative amount of time the system spends in particular states. A series of column vectors (sometimes called a template) reflects a trajectory through state-space; however a bundle of similar trajectories through state-space (b) may encompass a range of velocities, which may be governed by a combination of external inputs to the system or the asymmetry of the intrinsic synaptic matrix. In general, velocity in neural state-space can be quantified as the rate of decorrelation of successive state vectors. Parts (c) and (d) illustrate the fact that the latter measure, on average, is substantially different during wakeful behavior and during slow-wave sleep. In (c) is shown representative EEG traces, $Q$ matrix for 20 seconds of activity from 65 simultaneously recorded cells, and movement trace (bottom of each panel), from a rat hippocampus during behavior and slow-wave sleep. The column vector autocorrelograms for neocortex and hippocampus show that the mean state vector velocity changes more quickly in sleep. ((c) and (d) adapted from [24].)

Fig. 6. Trajectories through state-space that occur during a repeated behavior (in this case, the rat was repeatedly performing a learned spatial sequence task) reoccur during slow-wave sleep at about 7 times the speed at which they occurred during the behavior. This is illustrated in two independent ways. (a) (Left) An average template from one segment of the sequence, showing firing rate (spikes/bin) from multiple cells ($y$ axis) sorted by time of peak firing ($x$ axis). (Middle) Examples of good matches to the template from sleep (bottom) and task (top) periods. Bin size for the task and sleep examples are 100 ms and 14 ms respectively, the latter representing a compression factor of 7. (Right) The two graphs show histograms of match strength ($z$ scores) between templates and “target” data within sequentially ordered 14-s windows. $Z$ scores were derived via random shuffling of template columns. Color indicates the number of matches. In the bottom graph, target bin size is a factor of 7 smaller than in the template. (b) Sorted cross-correlations from simultaneously recorded cell pairs. Each row in each subpanel shows the cross-correlation between a single pair of cells, scaled so that peak and valley range from zero to one. The rows are sorted according to the temporal offset of the maximum peak during the behavior. In addition, only cell pairs showing a peak $z$ score exceeding 11 during the task were included. Red indicates the highest coincidence rate and blue, the lowest. The time axis during sleep epochs is magnified 6×. Note that the temporal lags of the peaks for behavior and sleep match up when the time base for sleep is compressed, indicating that the state-space trajectory is faster. (For interpretation of colors in this figure, the reader is referred to the web version of this article.)
clips are not replaced in the pool after replay, then we will observe that the replay of the current day’s memories dominates the activity in the first hour or so of sleep, but that the frequency of retrieval of this pool of memories decays approximately exponentially. This analogy may not be a perfect, biophysically accurate, description of the replay dynamics, but it captures the main ideas: during the first hour or so of sleep after learning, the brain retraces short segments of the learned state trajectories at high speed, interspersed with brief (100–200 msec) pauses, during which the cortex is essentially silent. The replay periods have been termed “up-states” whereas the silent pauses are called “down-states”. The trajectories from one up-state to the next are uncorrelated. After about an hour, the replay of the most recent experiences becomes statistically undetectable. The brain continues to exhibit the fluctuations between up- and down-states, but we are not able to read the up-state content experimentally. We presume that the latter content relates to older memories that were not experimentally recorded, and is hence unreadable.

What I have described is a process by which recent memories are replayed randomly interspersed with items already stored in memory. This may be recognized by those familiar with artificial learning theory as a prerequisite for adding new data into an existing network that has already been optimized for some corpus of knowledge, without creating what McCloskey and Cohen [19] called “catastrophic interference”. Catastrophic interference occurs when an optimized synaptic matrix is forced to acquire new data all at once, instead of by a process of gradual, gradient descent learning, during which new and old exemplars of the population from which the experiences are being sampled are interleaved. Forcing a network to make large weight changes in response to a new input essentially destroys the existing optimization of the weight distribution. It was this consideration that led McClelland et al. [18] to propose that the hippocampus and neocortex constitute complementary learning systems: the former captures raw data (memories) rapidly whereas the latter extracts knowledge about the world from these stored experiences. The dynamics of off-line memory trace reactivation are completely consistent with this conjecture. Presumably, the utility of shuffling the new and old memories during playback is to provide the interleaving necessary to perform smooth gradient descent learning, whereas the utility of high-speed playback optimizes the number of training trials that can occur in a limited time period.

But what allows memories to be played back at high speed? The short answer is that we do not really know. One simple idea is that, during learning, the state vector changes at a rate controlled by how fast the inputs from the external world change, whereas during replay, the system is free to change at a rate controlled by its internal parameters; the synaptic strengths, the asymmetry of the connection matrix, and various spike propagation delays. There is some preliminary evidence that, as the probability of a memory being replayed declines, its playback speed decreases, which would be consistent with the synaptic strength argument; however, this cannot be the complete answer, because it is already known that, during REM sleep, the brain state changes at about the speed it does during behavior [24], yet both REM and slow-wave sleep activity is spontaneous. So some other global parameter must play an important role in regulating playback speed. A likely candidate is changes in global neuromodulatory substances such as acetylcholine, one of whose main effects is to increase cortical inhibitory tonus [21]. Activity of subcortical neurons that release acetylcholine into the neocortex and hippocampus is at a minimum during slow-wave sleep, when the brain changes state most quickly. Acetylcholine is switched on again during REM sleep, when the rate of change is slowed down to real-time scale. Thus, it is possible that enhanced playback speed is regulated to some degree by cortical disinhibition, which might facilitate the transitions among quasi-stable attractor states (cell assemblies) linked by asymmetric connections.

It is tempting to speculate that the speed of thought itself may be related to the measurable brain-state trajectory speed, which would mean that factors that regulate the speed of thought in health and disease, at least in animal models, are now amenable to direct experimental study. There is indeed some evidence that defective cortical inhibition may underlie some of the disrupted normal thought process associated with schizophrenia and perhaps some of the enhanced creativity that sometimes accompanies this condition. On the other hand, there is increased cortical inhibition in aging [12] and decreased sequence reactivation, which has been associated with impaired learning in aged rats [4].

Acknowledgement

This work was supported by an Alberta Heritage Foundation for Medical Research Polaris Award.

References