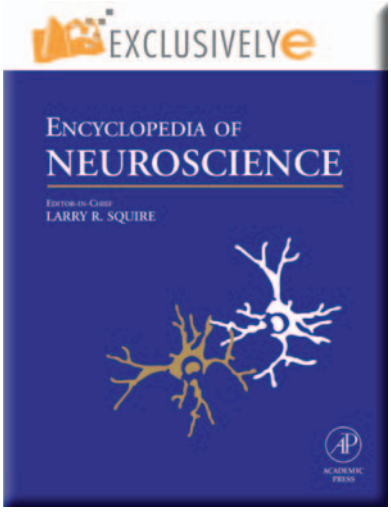

Wang X -J (2009) Attractor Network Models. In: Squire LR (ed.) Encyclopedia of Neuroscience, volume 1, pp. 667-679. Oxford: Academic Press.

Attractor Network Models

X-J Wang, Yale University School of Medicine, New Haven, CT, USA

Introduction

The term attractor is being increasingly used by neurophysiologists to characterize stable, stereotyped spatiotemporal neural circuit dynamics. Examples include readily identifiable rhythmic activity in a central pattern generator; well-organized propagation patterns of neuronal spike firing in cortical circuits *in vivo* or *in vitro*; self-sustained persistent activity during working memory; and neuronal ensemble representation of associative long-term memory. In these examples, the rich and complex neural activity patterns are generated largely through regenerative mechanism(s), and emerge as collective phenomena in recurrent networks. In part, interest in attractor networks arises from our growing appreciation that neural circuits are typically endowed with an abundance of feedback loops and that the attractor theory may provide a conceptual framework and technical tools for understanding such strongly recurrent networks.

The concept of attractors originates from the mathematics of dynamical systems. Given a fixed input, a system consisting of interacting units (e.g., neurons) typically evolves over time toward a stable state. Such a state is called an attractor because a small transient perturbation alters the system only momentarily; afterward, the system converges back to the same state. An example is illustrated in Figure 1, in which a neural network is described by a computational energy function in the space of neural activity patterns and the time evolution of the system corresponds to a movement down hill, in the direction of decreasing the computational energy. Each of the minima of the energy function is thus a stable (attractor) state of the system; a maximum at the top of a valley is an unstable state. Such a depiction is not merely schematic, but can be rendered quantitative for certain neural models.

The characterization stable and stereotyped is sometimes taken to imply that an attractor network is not sensitive to external stimuli and is difficult to reconfigure. On the contrary, as was shown by recent studies, attractor networks are not only responsive to inputs, but may in fact to instrumental to the slow time integration of sensory information in the brain. Moreover, attractors can be created or destroyed by (sustained) inputs; hence, the same network can serve different functions (such as working memory and decision making), depending on the inputs and cognitive control signals. The attractor landscape of a neural circuit is readily modifiable by changes in cellular and synaptic properties, which form the basis of the attractor model for associative learning.

In this article, we first introduce the basic concepts of dynamical systems, attractors, and bistability using simple single-neuron models. Then, we discuss attractor network models for associative long-term memory, working memory, and decision making. Modeling work and experimental evidence are reviewed and open questions are outlined. We show that attractor networks are capable of time integration and memory storage over timescales much longer than the biophysical time constants of fast electrical signals in neurons and synapses. Therefore, strongly recurrent attractor networks are especially relevant to memory and higher cognitive functions.

The Neuron Is a Dynamical System

A passive nerve membrane is a dynamical system described by a simple resistance–capacitance (RC) circuit equation:

$$C_m(dV_m/dt) = -g_L(V_m - E_L) + I_{app} \qquad [1]$$

where V_m is the transmembrane voltage, C_m the capacitance, g_L the leak conductance (the inverse of the input resistance), E_L the leak reversal potential, and I_{app} the injected current. In the absence of an input, the membrane is at the resting state, $V_{ss} = E_L$, say -70 mV. This steady state is stable; if V_m is transiently depolarized or hyperpolarized by a current pulse, after the input offset it will evolve back to V_{ss} exponentially with a time constant $\tau_m = C_m/g_L$ (typically 10–20 ms). Thus, V_{ss} is the simplest example of an attractor. More generally, for any sustained input drive I_{app}, the membrane always has a steady-state $V_{ss} = V_L + I_{app}/g_L$, given by $dV_m/dt = 0$ (i.e., V_{ss} does not change over time). The behavior of this passive membrane changes quantitatively with input current, but remains qualitatively the same; the dynamics is always an exponential time course (determined by τ_m) toward the steady-state V_{ss}, regardless of how high is the current intensity I_{app} or how large is the capacitance C_m or the leak conductance g_L. Moreover, the response to a combination of two stimuli I_1 and I_2 is predicted by a linear sum of the individual responses to I_1 or I_2 presented alone. These characteristics are generally true for a linear dynamical system, such as a differential equation with only linear dependence on V_m.

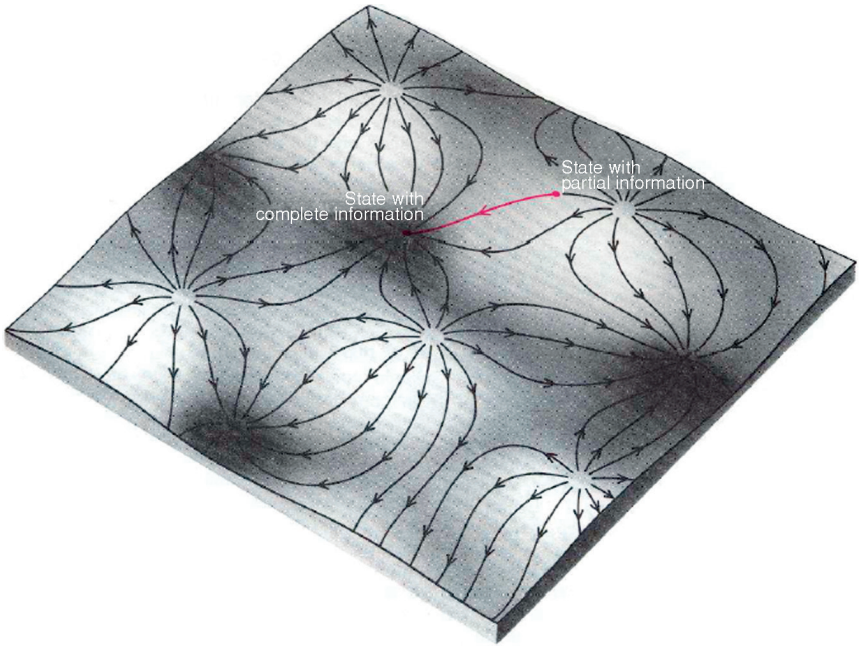

Figure 1 Schematic of an attractor model of neural networks. Computational energy function is depicted as a landscape of hills and valleys plotted against the neural activity states (on the *XY* plane). The synaptic connections and other properties of the circuit, as well as external inputs, determine its contours. The circuit computes by following a path that decreases the computational energy until the path reaches the bottom of a valley, which represents a stable state of the system (an attractor). In an associative memory circuit, the valleys correspond to memories that are stored as associated sets of information (the neural activities). If the circuit is cued to start out with approximate or incomplete information, it follows a path downhill to the nearest valley (red), which contains the complete information. From Tank DW and Hopfield JJ (1987) Collective computation in neuronlike circuits. *Scientific American* 257: 104–114.

More interesting behaviors become possible when nonlinearity is introduced by the inclusion of voltage-gated ionic currents. For instance, if we add in the RC circuit a noninactivating sodium current $I_{NaP} = g_{NaP} m_{NaP}(V_m)(V_m - E_{Na})$, where the conductance exhibits a nonlinear (sigmoid) dependence on V_m, the membrane dynamics becomes

$$C_m(dV_m/dt) = -g_L(V_m - E_L) - g_{NaP} m_{NaP}(V_m)(V_m - E_{Na}) + I_{app} \quad [2]$$

This system is endowed with a self-excitatory mechanism: a higher V_m leads to more I_{NaP}, which in turn produces a larger depolarization. If g_{NaP} is small, the weak positive feedback affects the membrane dynamics only slightly (Figure 2(b), red lines). With a sufficiently large g_{NaP}, the steady state at $V_{Down} \simeq -70$ mV is still stable because at this voltage I_{NaP} is not activated. However, the strong positive feedback gives rise to a second, depolarized plateau potential (at $V_{Up} \simeq -20$ mV) (Figure 2(b), blue lines). Therefore, the membrane is bistable; a brief input can switch the system from one attractor state to another (Figure 2(a)). As a result, a transient stimulus can now be remembered for a long time, in spite of the fact that the system has only a short biophysical time constant (20 ms). Unlike linear systems, in a nonlinear dynamical system gradual changes in a parameter (g_{NaP}) can give rise to an entirely new behavior (bistability).

Attractor states are stable under small perturbations, and switching between the two can be induced only with sufficiently strong inputs. How strong is strong enough? The answer can be found by plotting the total ion current $I_{tot} = I_L + I_{NaP} - I_{app}$ against V_m, called the *I–V* curve (Figure 2(b), top, with $I_{app} = 0$). Obviously, a V_m is a steady state if $I_{tot}(V_m) = 0$ (thus $dV_m/dt = 0$). As seen in Figure 2(b) (blue lines), the two attractors (filled circles) are separated by a third steady state (open circle). The third state is unstable – if V_m deviates slightly from it, the system does not return but converges to one of the two attractors. Indeed, if V_m is slightly smaller, I_{tot} is positive (hyperpolarizing), so V_m decreases toward V_{Down}. Conversely, if V_m is slightly larger, I_{tot} is negative (depolarizing), so V_m increases toward V_{Up}. Therefore, an external input must be strong enough to bring the membrane potential beyond the unstable steady state to switch the system from one attractor state to the other.

It is worth noting that the attractor landscape depends not only on the strength of the feedback mechanism (the value of g_{NaP}) but also sustained inputs. As shown in Figure 2(c), with a fixed g_{NaP} a constant applied current I_{app} shifts the *I–V* curve up

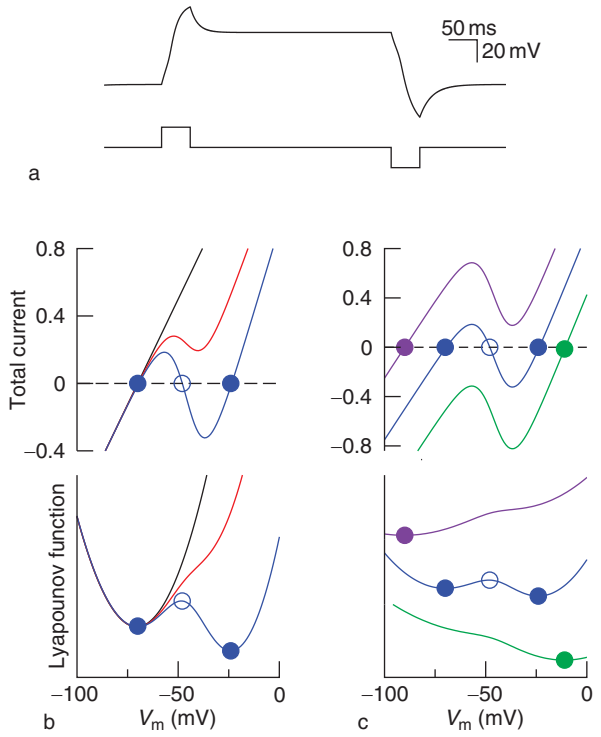

Figure 2 Positive feedback and attractor dynamics in a simple neural membrane model: (a) bistability with $g_{NaP} = 0.015\,\mu S$; (b) *I–V* curve (top) and computational energy function (Lyapounov function) (bottom); (c) bistability modulated by inputs, with $g_{NaP} = 0.015$ fixed, *I–V* curve (top) and computational energy function (Lyapounov function) (bottom). The membrane voltage is described by an RC circuit with the addition of a fast noninactivating sodium current $I_{NaP} = g_{NaP} m_{NaP}(V_m)(V_m - E_{Na})$, where $m_{Nap}(V_m) = 1/(1 + \exp[-(V_m + 45)/5])$ is a sigmoid function of V_m. The interplay between I_{NaP} and membrane depolarization produces an excitatory regenerative process. In (a), the system is initially at rest ($V_{Down} \simeq -70\,mV$). A depolarizing current pulse switches the membrane to a plateau potential ($V_{Up} \simeq -20\,mV$), which persists after the input offset. A second, hyperpolarizing current pulse switches the membrane back to the resting state. In (b), *I–V* curve (upper) is the total current $I_{tot} = I_L + I_{NaP} - I_{app}$ as a function of V_m, with $I_{app} = 0$); a steady state is given by an intersection with $I_{tot} = 0$. In the computational energy function (lower) $U(V_m)$, a steady state corresponds to a maximum (unstable) or a minimum (stable). For $g_{NaP} = 0$ (black, passive membrane) or $g_{NaP} = 0.08$ (red), there is only one steady state ($\simeq -70\,mV$). For $g_{NaP} = 0.015$ (blue), there are three steady states; two are stable (filled circles) and the third is unstable (open circle). In (c), the injected current intensity is varied (blue, violet, and green for $I_{app} = 0$, -0.5, and $0.5\,nA$, respectively). Other parameter values are $C_m = 0.5\,nF$, $g_L = 0.025\,\mu S$, $V_L = -70\,mV$, and $V_{Na} = 55\,mV$. The energy function $U(V_m)$ is defined by rewriting the circuit Eqn. (2) as $dV_m/dt = F(V_m) = -dU/dV_m$; hence, the energy function $U(V_m)$ is the integral of $-F(V_m)$. For instance, with $g_{NaP} = 0$, Eqn. (2) is reduced to Eqn. (1) and can be rewritten as $dV_m/dt = (V_{ss} - V_m)/\tau_m = F(V_m)$. By integrating $-F(V_m)$, we have $U(V_m) = (V_{ss} - V_m)^2/(2\tau_m) + U_0$, with an arbitrary constant U_0. Therefore, $U(V_m)$ is a parabola with V_{ss} at the bottom of the valley of the energy function (**Figure 1(b)**, black). For $g_{NaP} = 0.015$ (**Figure 1(b)**, blue), the energy function $U(V_m)$ displays two valleys, at V_{Down} and V_{Up}, separated by a peak at the unstable steady state. RC, resistance-capacitance; V_m, membrane voltage.

or down. Either a hyperpolarization or depolarization can destroy the bistability phenomenon. This simple example demonstrates that neuronal bistable dynamics can be readily reconfigured by external inputs. This is generally true for neural networks as well and has important computational implications. Attractors need not be steady states. In neurons, a plateau potential is typically not stable as a steady state. Instead, on depolarization the Hodgkin–Huxley-type sodium and potassium currents produce repetitive action potentials, which represent another (oscillatory) type of attractor behavior. The attractor nature of periodic spiking is shown in **Figure 3** using the classic Hodgkin–Huxley model; regardless of the initial states, the membrane system always converges to the same periodic attractor state. If the system is perturbed by a transient stimulus, it resumes the same firing pattern after the stimulus offset, except for a shift of spiking time or the phase of the periodic attractor state. Thus, a periodic attractor is robust (the amplitude and periodicity are not sensitive to transient perturbations). At the same time it is sensitive to phase shift; hence, the clock can be readily reset. This general property of nonlinear oscillators is a key to understanding synchronization among neurons and coherent brain rhythms. It is also a cornerstone of the neurobiology of biological clocks, such as the circadian rhythm, sleep cycle, and central pattern generators for locomotion such as walking, swimming, and breathing. In each of these systems, the rhythm is amenable to being reset by a transient input which leads to a phase shift in time but otherwise does not alter the stereotypical activity pattern in the network.

The oscillatory attractor illustrates that the behavior of a dynamical system (the Hodgkin–Huxley model) is determined by direct observables (the membrane voltage) as well as internal dynamical variables (ion channel-gating variables). The space of all dynamical variables form a phase space of the system, which is typically multidimensional. A neural circuit in the mammalian brain consists of many thousands of cells; its phase space is enormous. The dynamics of such systems can be very complex, exhibiting a wide gamut of spatiotemporal activity patterns. It is generally not possible to define an energy function for such systems. Nevertheless, the concept of computational energy landscape is still helpful for developing intuitions about attractor networks.

Synaptic Plasticity and Associative Memory

Bistable switches or, more generally, multiple attractors can be realized on all scales, from the molecular

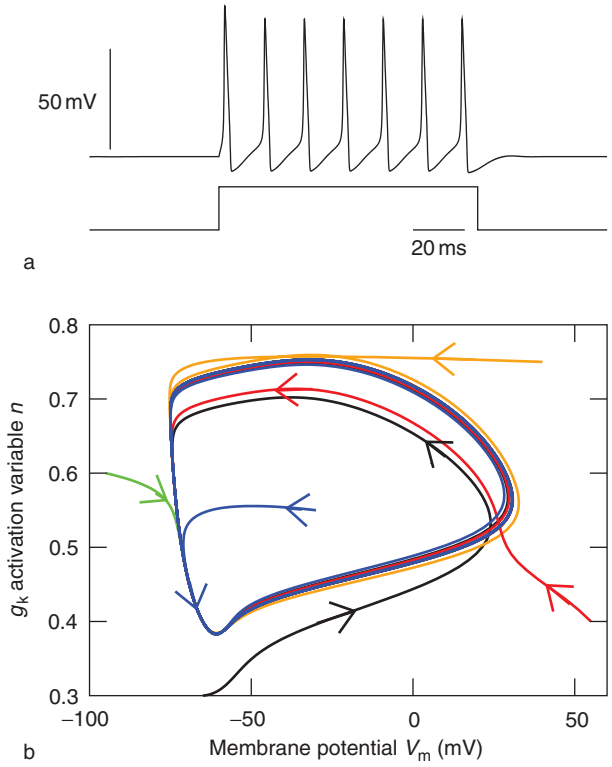

Figure 3 Oscillatory attractor in the original Hodgkin–Huxley model of action potentials: (a) repetitive firing of action potentials with a current pulse $I_{app} = 9\,\mu A/cm^2$; (b) potassium activation variable n plotted against V_m, showing that the oscillation forms a closed trajectory in this projected phase space. The model has a leak current I_L, a fast sodium current I_{Na}, and a noninactivating potassium current I_K. Its dynamics is described by four coupled differential equations (the membrane voltage V_m, the activation and inactivation gating variables m and h for I_{Na}, and the activation variable n for I_K). In (b), the trajectory roughly consists of three portions: the upstroke of an action potential (when both V_m and n increase), the downstroke (when V_m decreases while n keeps increasing and then starts to decrease), and the refractory period (when n continues to decrease while V_m starts to increase). Different colors correspond to five different initial conditions (with different V_m and n values at $t = 0$), in all cases the system dynamically evolves into the oscillatory attractor state. The Hodgkin–Huxley model exhibits bistability between a steady resting state and an oscillatory state, which is not shown for the sake of clarity.

machinery of individual synapses and the electrical activity of single neurons to large neural circuits. It is well known that synapses that form connections between neurons are highly plastic, and experience-dependent synaptic modifications are believed to be a physiological substrate of learning and memory. A single synapse comprises hundreds of proteins that interact with one another in a highly connected signal transduction network; therefore, a synapse is a dynamical system. The biochemical time constants in such a network, and the typical protein lifetimes, range from seconds to hours. So how can synapses store memories that may be retained for many years? One possibility is that the expression of memory maintenance involves changes in the molecular composition of the synapse that are mechanically stable over a long time. Alternatively, memories could be stored in molecular switches, with two or more states that are stable over durations beyond the intrinsic molecular time constants. Recent physiological experiments have yielded evidence for switchlike behavior during synaptic modifications. Molecular studies and modeling have revealed several candidate protein kinases that may exhibit switchlike behavior at single synapses, such as calcium/calmodulin-dependent protein kinase II (CaMKII), protein kinase C (PKC), and mitogen-activated protein kinase (MAPK). Memory switches do not necessarily have an infinite lifetime because an active state may be turned off by subsequent synaptic changes during ongoing neural activity by protein turnover or by molecular fluctuations due to the small synaptic volume (approximately 0.1 fl). Nevertheless, the key point here is that, if individual synapses exhibit stable self-sustained active states (attractors), the lifetime of a memory trace is not directly limited by the biochemical time constants of synaptic signaling pathways.

Whereas synaptic plasticity provides a structural basis for memory formation, the stored information is encoded in a distributed manner by the synaptic connection patterns in a neural circuit. Theoretically, it has been proposed that associative memories are learned through the creation of stable neural activity patterns (attractors). In this view, a memory network has many attractor states, each representing a particular memory item, and has its own basin of attraction within which other states evolve dynamically into the attractor state (**Figure 1**). The stability of attractor states ensures that memory storage is robust against small perturbations. These memory states are imprinted in the network by long-lasting changes of synaptic connections through Hebbian learning, and much theoretical work has been devoted to the analysis of storage capacity (the number of memory items that can be stored and retrieved reliably as a function of the network size). Memories thus established are associative because a partial cue brings the network into the basin of attraction of an attractor with the information content close to that of the sensory cue. Thus, memory retrieval can be done by association between a cue and the corresponding memory item, and the recall process is error-correcting (incomplete information still leads to the correct memory retrieval). This capability for pattern completion is a hallmark of associative memory. At the same time, an attractor model is also capable of pattern separation, in the sense that two slightly different input patterns near the boundary of two basins of attraction may

drive the network to two distinct attractor patterns, leading to the retrieval of two separate memories. To experimentally test the attractor model, many neurons must be simultaneously monitored so that distributed activity patterns in memory circuits can be assessed. One of such circuits is the hippocampus, which is known to be critical to the formation of episodic memory and spatial memory. In rodents, during exploration in a familiar environment, pyramidal place cells in the hippocampus are activated when the animal passes through a specific location, called a place field. This spatial selectivity is characterized by a bell-shaped tuning curve (neuronal firing rate as a function of the animal's location). Place fields of hippocampal cells cover the entire surface of the environment with about an equal density, so that the neural ensemble firing can be decoded to read out the animal's location in space. The hippocampus thus implements a neuronal representation of a spatial map. Moreover, when the animal is exposed to another environment, the place fields of cells undergo great changes (remapping) until a representation of the new environment is established in the hippocampus. Therefore, the hippocampus stores distinct maps, with each map being activated as the rat enters a different environment. It has been proposed that these spatial representations reflect distinct attractor states, based on the observation that activity of place cells is preserved in the dark (without external visual cues). The attractor model has recently been directly tested by combining a clever navigation task design with simultaneous recording from many single cells in the hippocampus. In one experiment, rats were trained to forage in two distinct environments, a square and a circle that differed in color, texture, and shape. This led to the remapping of place fields in the majority of cells recorded in the hippocampal area CA1. Then, in probe trials, the rat commuted between a series of environments of the same color and texture, but the shape morphed gradually between the square and the circle. These environments were chosen randomly from trial to trial. Remarkably, it was found that the place fields of the recorded cells abruptly and coherently changed from squarelike to circlelike (**Figure 4**). This observation provides strong support for the attractor model, which predicts that the hippocampal coding of space should exhibit both pattern completion (minor changes from a familiar environment – a square or a circle – do not alter place fields) and pattern separation (similar inputs intermediate between squarelike and circlelike result in drastically different place fields, due to a switch between the two learned maps).

Persistent neural firing observed in a brain area may not necessarily be generated locally but is a mere reflection of mnemonic activity elsewhere. The area CA1 is commonly viewed as an readout circuit, whereas memories are believed to be stored upstream. It has long been hypothesized that autoassociative memories are formed in the area CA3, which projects to CA1 and is endowed with strong recurrent excitatory connections among pyramidal cells, a prerequisite for the generation of attractor states. Furthermore, recent discoveries have drawn attention to the entorhinal cortex, the major input area for the hippocampus. Cells in the dorsolateral entorhinal cortex fire whenever the rat is on any vertex of a triangular lattice spanning the whole surface of the environment. This grid-cell activity is preserved in the dark, when the visual cues are absent. These findings have led to the proposal that the entorhinal cortex embeds an attractor network with a gridlike representation of the environment and that the hippocampus transforms this periodic firing pattern in the input into a nonperiodic firing pattern that encodes the animal's current spatial position. If so, it remains to be seen whether the hippocampus still needs to operate as an attractor network and whether spatial navigation is subserved by interconnected attractor circuits.

In order for an animal to navigate without reference to external cues, the neural instantiation of a spatial map must be constantly updated by the integration of linear and angular self-motion. Whether an attractor model is capable of carrying out such computations robustly is not well understood. Finally, little is known about precisely how the attractor landscape in the space of neuronal firing patterns is shaped by the details of the animal's training process. Another study, using a somewhat different learning and probe procedure but also with morphed environmental shapes, found that place fields of CA3 and CA1 cells switched more gradually and less coherently. This finding does not necessarily mean that the attractor paradigm is incorrect; we expect from computational work that a different learning history gives rise to different sets of attractor states. Elucidation of the general principles and cellular mechanisms of this learning process in future experiments and modeling will help determine whether the attractor model represents a sound theoretical framework for the neurobiology of learning and memory.

Persistent Activity and Working Memory

Attractor models have also been applied to working memory, our brain's ability to actively hold information online for a brief period of time (seconds). Neurons that maintain working memory must be manifestly active in a sustained manner. How can such persistent activity be generated in the absence

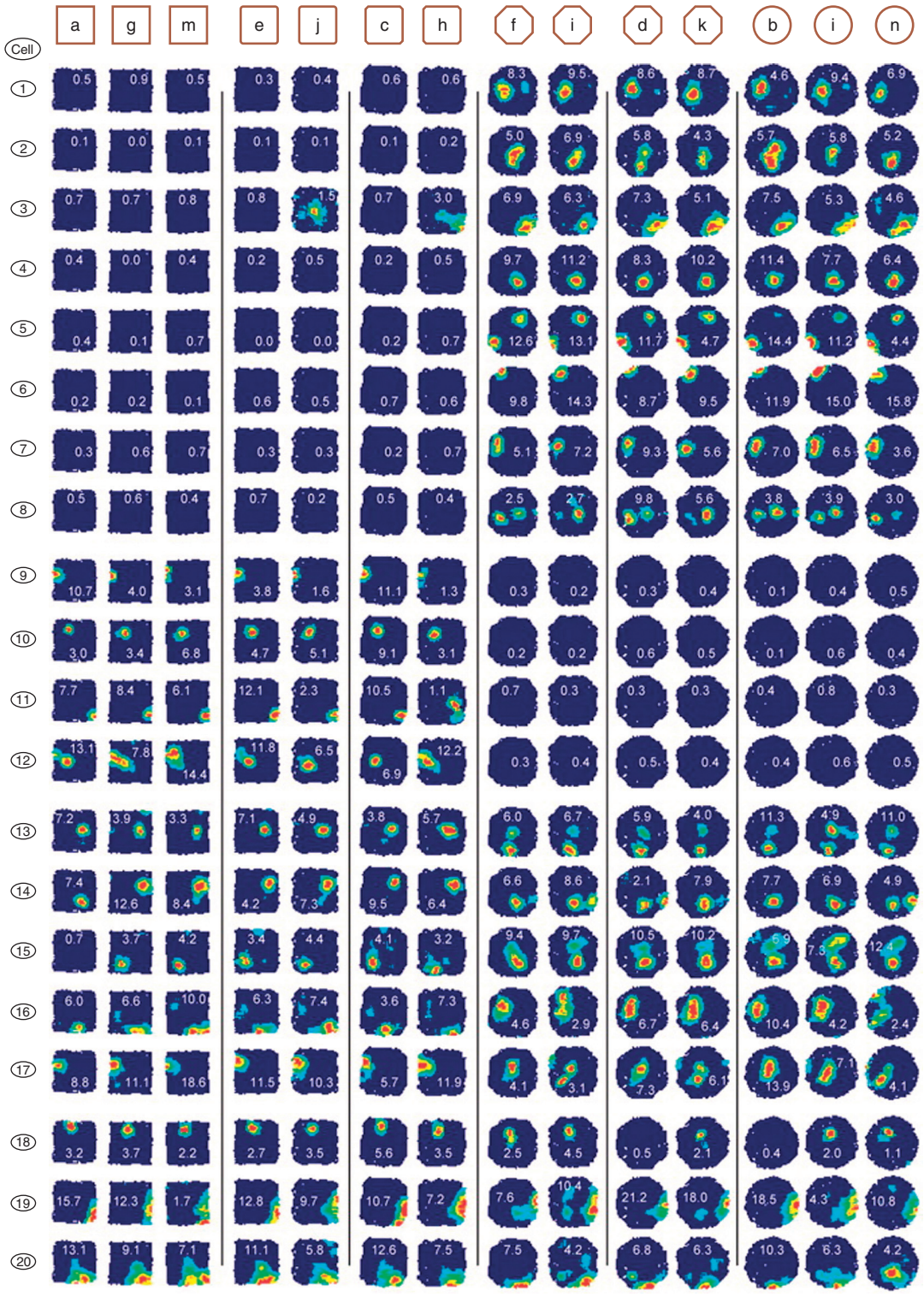

Figure 4 Abrupt and coherent expression of spatial representation of hippocampal neurons. A rat explores a series of environments with morphed shapes (top icons) between a square (the three left-most columns) and a circle (the three right-most columns). Twenty single cells were simultaneously recorded, shown in rows from top to bottom. Warm colors indicate locations (place field) of the rat when the cell's spiking activity is high. Each field is scaled to peak firing rate shown in red. The 17 of 20 simultaneously recorded place cells with different (remapped) firing patterns in the square and the circle, almost all switch from the squarelike to the circlelike pattern between the h and f octagons. Eight cells had fields in the circle but not the square (cells 1–8); four had fields in the square but not the circle (9–12); five fired in both but in different places (13–17); and three did not reach the criterion for remapping (18–20). From Wills TJ, Lever C, Cacucci F, Burgess N, and O'Keefe J (2005) Attractor dynamics in the hippocampal representation of the local environment. *Science* 308: 873–876.

of direct external input? R Lorente de Nó in the 1930s and DO Hebb in the late 1940s proposed that the answer lies in the feedback loop connections. Thus, in a working memory network, every cell receives excitatory drive from both afferent inputs and intrinsic synaptic connections. Inputs activate neurons in a selective cell assembly; the triggered spike activity reverberates through excitatory synaptic circuit, which is enough to sustain an elevated firing when the inputs are withdrawn. This general idea has been made rigorous in attractor models, according to which a working memory circuit exhibits multiple attractor states (each coding a particular memory item) that coexist with a background (resting) state. All the attractor states are self-maintained and relatively stable in the face of small perturbations or noise. Yet memory states can be turned on or switched off by brief external stimuli.

Stimulus-selective neural persistent activity has been observed in awake animals performing delayed-response tasks that depend on working memory. For example, in a delayed match-to-sample task, two visual objects are shown consecutively separated by a time gap of a few seconds, and the subject is asked to judge whether the two stimuli are the same. Or after the presentation of the first object and the delay, an array of visual objects are shown and the subject must indicate by a motor response which of them is identical to the first visual cue. In both cases, the subject's performance relies on the working memory about the first object across the delay. The stored information involves a collection of discrete items. Other delayed-response tasks engage working memory of an analog quantity, such as spatial location or stimulus amplitude. While a monkey was performing such a task, the neurons in the prefrontal, posterior parietal, inferotemporal, and premotor cortices were found to exhibit elevated persistent activity that was selective to stimuli. Three types of mnemonic coding has been observed: (1) object working memory cells are tuned to one or a few of discrete memory items, (2) spatial working memory cells typically exhibit a bell-shaped (Gaussian) tuning function of the spatial location or directional angle, and (3) cells that store parametric working memory of magnitudes (e.g., vibration stimulus frequency) are characterized by a monotonic tuning function of the encoded feature.

These experimental observations lend support to the attractor model inasmuch as stimulus-selective persistent firing patterns are sustained internally in the absence of direct sensory input, are dynamically stable, and are approximately tonic in time (e.g., across a delay). However, experiments show that delay neural activity is often not tonic but exhibits time variations such as ramping up or ramping down. How to account for the heterogeneity and time courses of mnemonic persistent activity represents a challenge to the attractor network model. Moreover, it remains an open question as to what cellular or circuit mechanisms are responsible for the generation of persistent activity. This question is now being addressed using biologically constrained models of persistent activity. Figure 5 shows such a recurrent network model for spatial working memory in which spiking neurons and synapses are calibrated by the known cortical electrophysiology. The key feature is an abundance of recurrent connections (loops) between neurons, according to a Mexican-hat-type architecture – localized recurrent excitation between pyramidal cells with similar preference to spatial cues and broader inhibition mediated by interneurons (Figure 5(a)). In a simulation of a delayed oculomotor task (Figure 5(b)), the network is initially in a resting state in which all cells fire spontaneously at low rates. A transient input drives a subpopulation of cells to fire at high rates. As a result, they send recruited excitation to one another via horizontal connections. This internal excitation is large enough to sustain elevated activity, so the firing pattern persists after the stimulus is withdrawn. Synaptic inhibition ensures that the activity does not spread to the rest of the network, and persistent activity has a bell shape (bump attractor). At the end of a mnemonic delay period, the cue information can be retrieved by reading out the peak location of the persistent activity pattern; and the network is reset to the resting state. In different trials, a cue can be presented at different locations. Each cue triggers a persistent firing pattern of the same bell shape but with the peak at a different location (Figure 5(c)). A spatial working memory network thus displays a continuous family of bump attractors.

Biophysically realistic models have specific predictions about the circuit properties required for the generation of stimulus-selective persistent activity. In particular, it was found that a network with strong recurrent loops is prone to instability if excitation (positive feedback) is fast compared to negative feedback, as is expected for a nonlinear dynamical system in general. This is the case when excitation is mediated by the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPARs), which are approximately two to three times faster than inhibition mediated by γ-aminobutyric acid A ($GABA_A$) receptors (time constant, 5–10 ms). The interplay between AMPARs and $GABA_A$ receptors in a excitatory-inhibitory loop naturally produces fast network oscillations. In a working memory model, the large amount of recurrent connections, needed for the generation of persistent activity, often leads to excessive oscillations that are detrimental to network stability.

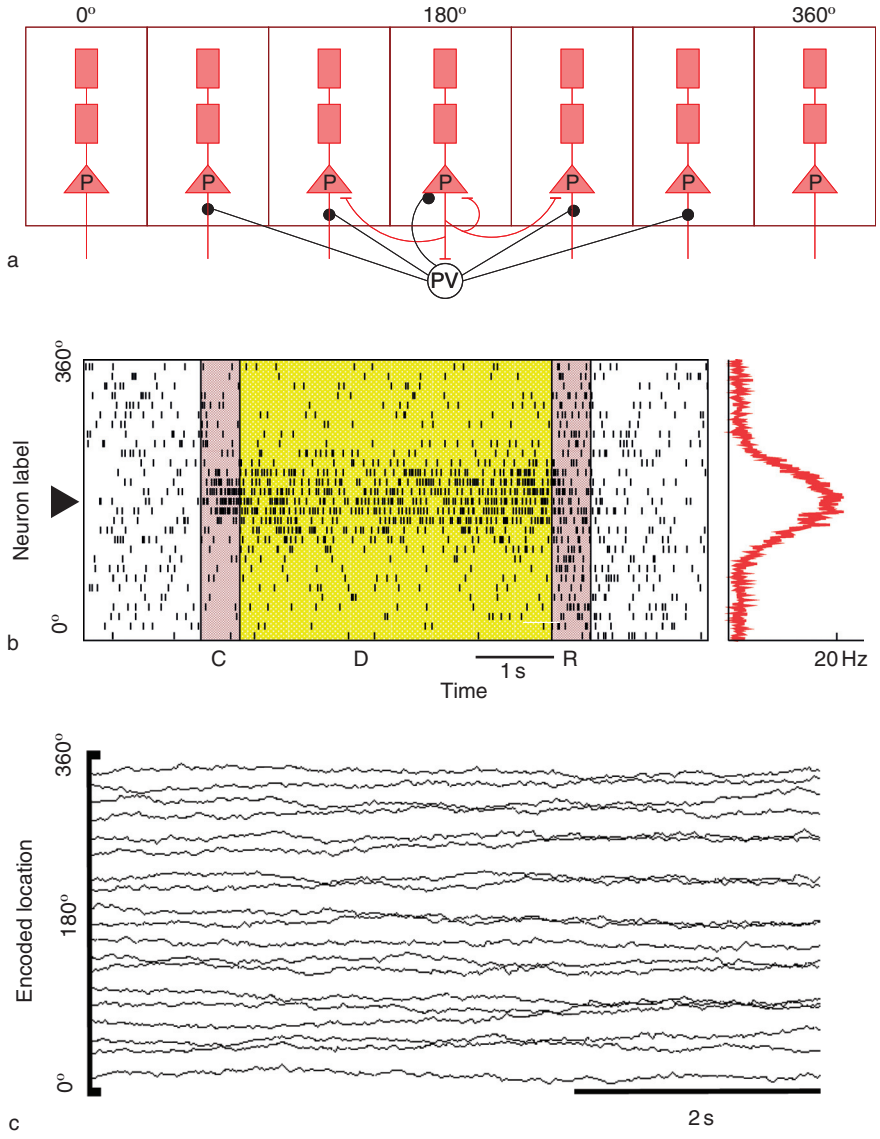

Figure 5 Bump attractor model for spatial working memory: (a) Mexican-hat-type connectivity of the model circuit; (b) spatiotemporal activity pattern of the pyramidal cell population in a simulation of delayed oculomotor response task; (c) temporal evolution of the peak location of mnemonic persistent activity pattern in 20 trials with transient stimuli at different locations. In (a), pyramidal (P) neurons are arranged according to their preferred cues (0–360°). Recurrent excitatory connections are strong between cells with similar cue preference and decrease with the difference in their preferred cues. Local excitation is counteracted by broad synaptic inhibition mediated by paravalbumin (PV) expressing inhibitory interneurons. In (b), each dot is an action potential; the elevated and localized neural activity is triggered by the cue stimulus at 180° and persists during the delay period. On the right is shown the spatial pattern, in which the average firing rate during the delay period is plotted vs. the neuron's preferred cue. The memory of the spatial cue is stored by the peak location of this bell-shaped persistent activity pattern (bump attractor). In (c), the remembered cue (defined by the peak location of the network activity pattern) as a function of time during a 6 s delay period, for a number of trials, each with a different initial spatial cue. R, response period. Panel (b) From Compte A, Brunel N, Goldman-Rakic PS, and Wang X-J (2000) Synaptic mechanisms and network dynamics underlying visuospatial working memory in a cortical network model. *Cerebral Cortex* 10: 910–923; panel (c) from Renart A, Song P, and Wang X-J (2003) Robust spatial working memory in a heterogeneous network model with homeostatic synaptic scaling. *Neuron* 38: 473–485.

Working memory function can be rendered stable if excitatory reverberation is slow, that is, contributed by the *N*-methyl-D-aspartate (NMDA) receptors (time constant 50–100 ms) at recurrent synapses. Thus, the model predicts a critical contribution of NMDA receptors to working memory. Other processes with time constants of hundreds of milliseconds, such as short-term synaptic facilitation or intrinsic ion channels in single cells, may also contribute to reverberatory dynamics underlying working memory.

On the other hand, the feedback mechanism cannot be too slow. An alternative to the attractor model is the scenario in which persistent activity actually is not stable but represents slowly decaying afterdischarges mediated by some intracellular mechanisms such as second-messenger pathways or kinetics of protein kinases. However, this scenario predicts that triggering inputs must be long lasting (lasting for seconds), which is incompatible with physiological experiments in which working memory states have been shown to be switchable quickly by brief external inputs (a few hundreds of a millisecond). The recurrent (attractor) network mechanism achieves the stability and long persistence time of memory storage, as well as rapid flexible memory encoding and erasure, that are behaviorally desirable.

Winner-Take-All and Vector Averaging

Attractor dynamics naturally instantiate winner-take-all (WTA); when several neural pools, each selective for a different sensory stimulus, are simultaneously activated by the presence of several stimuli, competition through recurrent attractor dynamics may ultimately lead to an output pattern with a high firing activity in one neural pool while all the other neural pools are suppressed. WTA could subserve a number of neural processes, such as categorical decision making. In a continuous neural network, such as the ring model shown in Figure 6, WTA operation is not always categorical but depends on the similarity between simultaneously presented stimuli. For instance, if two inputs are very close to one another, instead of WTA the network's response is expected to be a vector average of the two individual responses to the stimuli presented alone. Moreover, in a working memory network, WTA often involves different stimuli that occur at different times; this interaction across temporal gaps is possible because an earlier input can trigger persistent activity that interacts with responses to later inputs.

These points are illustrated in Figure 6, which shows the interaction of a remembered stimulus and a distractor in the continuous (bump) attractor network model of spatial working memory (same as in Figure 5). Significantly, the to-be-remembered stimulus (located at θ_S) and the distractor (θ_D) have the same amplitude but presented at different times, one at the beginning of a trial and the other during the delay. As we intuitively expect, the impact of a distractor depends on its strength (saliency). If the stimulation amplitude is sufficiently large, the distractor is powerful enough to overcome the intrinsic dynamics of the recurrent circuit, and the network is always perturbed to a location close to the intervening stimulus (Figure 6(a), top; red curve in Figure 6(b)). In this case, the network can be reset by every new transient stimulus and keeps a memory of the last stimulus in the form of a refreshed selective persistent activity state. On the other hand, if external input is moderately strong relative to recurrent synaptic drive, a distractor triggers only a transient response and the memory of the initial cue is preserved (Figure 6(b), bottom). The resistance to distractors can be understood by the fact that, in a memory delay period, active neurons recruit inhibitions which project to the rest of the network. Consequently, those cells not encoding the initial cue are less excitable than when they are in the resting state and hence are less responsive to distracting stimuli presented during the delay. Moreover, WTA takes place only when the spatial locations of the initial cue and the later distractor are sufficiently distant from one another (more than 90°) (blue curve in Figure 6(b)). When they are closer, the activated neural pools (selective for θ_S and θ_D) overlap one another and the peak location of the resulting population activity is roughly the vector average of θ_S and θ_D (blue curve in Figure 6(b)). Hence, the same network can perform both WTA and vector-averaging computations, depending on the similarity of stimuli.

Time Integration and Categorical Decision Making

Cortical areas that are engaged in working memory – such as the prefrontal cortex – are also involved in other cognitive functions such as decision making, selective attention, and behavioral control. This suggests that microcircuit organization in these areas is equipped with the necessary properties to subserve both the internal representation of information and dynamical computations of cognitive types. As it turns out, models originally developed for working memory can account for decision-making processes as well. An example is shown in Figure 7 from model simulations of a visual motion-discrimination experiment. In this two-alternative forced-choice task, monkeys are trained to make a judgment about the direction of motion (say, left or right) in a stochastic random dot display and to report the perceived direction with a saccadic eye movement. A percentage of dots (called motion strength) move coherently in the same direction, so the task can be made easy or difficult by varying the motion strength (from close to 100 to 0%) from trial to trial. While a monkey is performing the task, single-unit recordings revealed that neurons in the posterior parietal cortex and prefrontal cortex exhibit firing activity correlated with the animal's perceptual choice. For example, in a trial

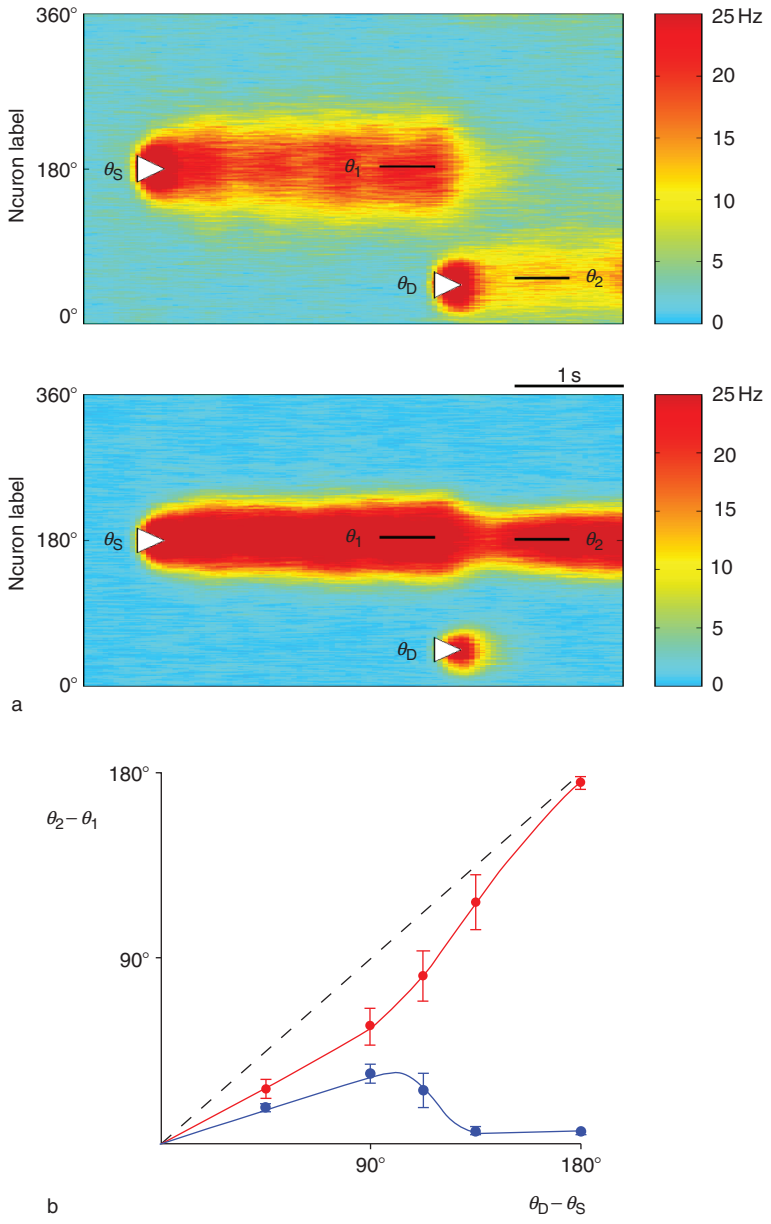

Figure 6 Winner-take-all and vector averaging in the continuous attractor network model of spatial working memory (Figure 5): (a) two sample simulations; (b) dependence of network distraction on the distance between the cue and distractor and on the stimulation intensity. In (a), the network spatiotemporal firing pattern is plotted in a color-coded map. A cue is presented initially at θ_S, triggering a tuned persistent activity. After a 2.5 s delay, a distractor stimulus is presented at θ_D, with the same intensity and duration as the cue stimulus. The peak location of the population activity pattern is computed just before the distractor (θ_1) and 500 ms after the distractor (θ_2). The network is completely distracted by a strong stimulus (top), but is resistant to a distractor of moderate intensity (bottom). In (b), the distracted angle $\theta_2 - \theta_1$ is plotted vs. the distraction angle θ_D–θ_S for several distractor cues. The dashed line indicates perfect distraction (as in (a), top), whereas points on the *x*-axis imply the complete absence of distraction (as in (a), bottom). The red curve indicates large input intensity; the blue curve indicates moderate input intensity. In the latter case, the network exhibits a switch from winner-take-all (when $\theta_D - \theta_S$ is large) to vector averaging (when θ_D–θ_S is small). Adapted from Compte A, Brunel N, Goldman-Rakic PS, and Wang X-J (2000) Synaptic mechanisms and network dynamics underlying visuospatial working memory in a cortical network model. *Cerebral Cortex* 10: 910–923.

in which the motion strength is low (say, 6.4%), if the stimulus direction is left but the monkey's choice is right, the response is incorrect. In that case, cells selective for right display a higher activity than those selective for left. Hence, the neural activity signals the animal's perceptual decision rather than the actual sensory stimulus. This experiment can be simulated using the same model designed for working

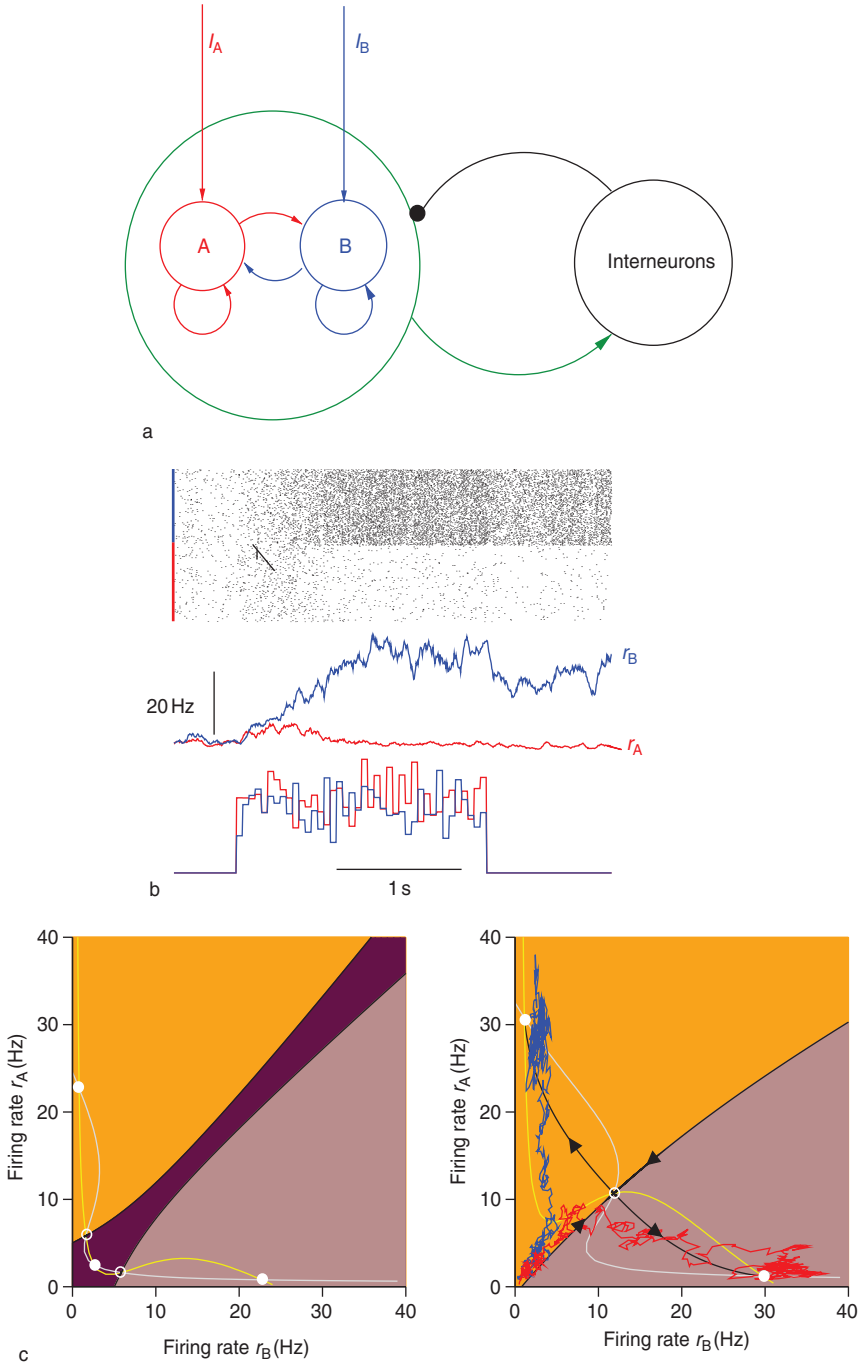

Figure 7 A spiking neuron circuit model for two-alternative forced-choice tasks: (a) model scheme; (b) network simulation with $c' = (I_A - I_B)/(I_A + I_B) = 6.4\%$; (c) decision dynamics shown in the two-dimensional plane where firing rates r_A and r_B are plotted against one another. In (a), there are two groups of spiking pyramidal cells, each of which is selective to one of the two directions (A = left, B = right) of random moving dots in a visual motion-discrimination experiment. Within each pyramidal neural group, there are strong recurrent excitatory connections which can sustain persistent activity triggered by a transient preferred stimulus. The two neural groups compete through feedback inhibition from interneurons. In (b), the population firing rates r_A and r_B exhibit an initial slow ramping (time integration) followed by eventual divergence (categorical choice). In this sample trial, although the input A is larger, the decision is B (an erroneous choice). As shown in (c), in the absence of stimulation (left), there are three attractor states (filled circles) and two unstable steady states (open circles); colored regions are the basins of attractions (maroon for the resting state; orange and brown for the persistent activity states). With a stimulus of $c' = 6.4\%$ in favor of choice A (right), the decision trajectory is shown for two trials (correct trial in blue; error trial in red). Panels (a,b) from Wang X-J (2002) Probabilistic decision making by slow reverberation in neocortical circuits. *Neuron* 36: 955–968; panel (c) from a simulation by KF Wong using the model published in Wong KF and Wang X-J (2006) A recurrent network mechanism for time integration in perceptual decisions. *Journal of Neuroscience* 26: 1314–1328.

memory. The only difference between a working memory simulation and a decision simulation is that in a delayed-response task only one stimulus is presented but for a perceptual discrimination task conflicting sensory inputs are fed into competing neural subpopulations in the circuit. This is schematically depicted in Figure 7(a), in which the relative difference in the inputs $c' = (I_A - I_B)/(I_A + I_B)$ mimics the motion strength in the visual motion discrimination experiment. Figure 7(b) shows a simulation with $c' = 6.4\%$. At the stimulus onset, the firing rates of the two competing neural populations, r_A and r_B, initially ramp up together for hundreds of milliseconds before diverging from one another when one increases while the other declines. The perceptual choice is decided based on which of the two neural populations wins the competition. Therefore, consistent with the physiological observations from the monkey experiment, the decision process proceeds in two steps. Sensory data are first integrated over time in a graded fashion, which in the model is instantiated by the NMDA receptor-dependent slow reverberation. This is followed by WTA competition produced by synaptic inhibition, leading to a categorical (binary) choice.

Figure 7(c) shows attractor dynamics in the decision space, where r_A is plotted against r_B. In the absence of stimulation (Figure 7(c), left), three attractors coexist (filled circles): a spontaneous state (when both r_A and r_B are low), and two persistent activity states (with high r_A and low r_B , or vice versa). On the presentation of a stimulus (Figure 7(c), right), the attractor landscape is altered and the spontaneous steady state disappears, so the system is forced to evolve toward one of the two active states which represent perceptual decisions (A or B). In this graph, the sensory evidence is in favor of the choice A (with $c' = 6.4\,\%$), so the attractor A has a larger basin of attraction (orange) than that of the attractor B (brown). The system is initially in the spontaneous state which now falls in the basin of attraction A and evolves toward decision state A in a correct trial (blue). However, at low c', the bias is not strong, and noise can induce the system's trajectory to travel across the boundary of the two attraction basins, in which case the system eventually evolves to decision state B in an error trial (red). The crossing of a boundary between attraction basins is slow, which explains why the reaction times are longer in error trials than in correct trials, as was observed in the monkey experiment. After the offset of the stimulus, the system's configuration reverts to that in Figure 7(c) (left). Because a persistently active state is self-sustained, the perceptual choice (A or B) can be stored in working memory for later use, to guide behavior. In this way, the attractor dynamics model offers an unified account for working memory and decision-making computations.

Concluding Remarks

In summary, the language of attractors is natural for describing the electrical activity of neurons and neural circuits. It provides a plausible theoretical framework for both short-term working memory and long-term associative memory. Significantly, nonlinearity due to feedback loops makes it possible that graded changes of a cellular or synaptic parameter lead to the emergence of qualitatively different behaviors (e.g., with or without persistent activity). The functional implications of this insight are potentially far-reaching because it suggests that cortical areas dedicated to distinct functions (e.g., sensory processing vs. working memory) may share a same canonical cortical circuit layout but with subtle differences in the cellular/molecular makeup, connectivity properties, and neuromodulatory influence.

Qualitatively speaking, working memory requires neurons to convert a transient input pulse into a sustained persistent activity, such as a time integral of the stimulus. Similarly, in perceptual decisions, approximate linear ramping activity, at a rate proportional to input strength, can also be conceptualized as time integration. It is worth noting, however, that a genuine integrator implies that, after a transient input is turned off, the activity is persistent at a firing rate proportional to the input strength, spanning a continuous range. This is not the case in Figure 7, in which after the stimulus offset the neural activity is binary (representing one of the two categorical choices), independent of the input motion strength. This is what has been observed in posterior parietal neurons, and it is the kind of neural signals needed to accomplish categorical choices. Integration by neural circuits over long timescales represents an important topic of active research.

An open question concerning working memory is whether persistent activity is primarily generated by local circuit dynamics, a single-cell property, or a large-scale network composed of several brain regions. For both working memory and long-term memory, a major challenge is to elucidate the biological substrates that underlie the robustness of continuous attractors and integrators. Furthermore, there is a dichotomy between rapidly switchable attractors, on the one hand, and an intracellular signaling network with

multiple time constants, on the other hand. The interplay between cellular processes and collective network dynamics will turn out to be an exciting topic in future research.

Acknowledgments

I thank K-F Wong for making Figure 7(c). This work is supported by the NIH grant MH62349.

See also: Hodgkin–Huxley Models; Neural Integrator Models; Prefrontal Cortex: Structure and Anatomy; Prefrontal Cortex; Short Term and Working Memory; Visual Associative Memory; Working Memory: Capacity Limitations.

Further Reading

Amari S (1977) Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics* 27: 77–87.

Amit DJ (1992) *Modelling Brain Function: The World of Attractor Neural Networks*. Cambridge, UK: Cambridge University. Press.

Ben-Yishai R, Bar-Or RL, and Sompolinsky H (1995) Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences of the United States of America* 92: 3844–3848.

Bhalla US and Iyengar R (1999) Emergent properties of networks of biological signaling pathways. *Science* 283: 381–387.

Brunel N (2005) Network models of memory. In: Chow CC, Gutkin B, Hansel D, Meunier C, and Dalibard J (eds.) *Methods and Models in Neurophysics*, pp. 407–476. Amsterdam: Elsevier.

Camperi M and Wang XJ (1998) A model of visuospatial working memory in prefrontal cortex: Recurrent network and cellular bistability. *Journal of Computational Neuroscience* 5: 383–405.

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America* 79: 2554–2558.

Machens CK, Romo R, and Brody CD (2005) Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science* 307: 1121–1124.

McNaughton BL, Battaglia FP, Jensen O, Moser EI, and Moser MB (2006) Path integration and the neural basis of the "cognitive map." *Nature Reviews Neuroscience* 7: 663–678.

Miller P, Zhabotinsky A, Lisman J, and Wang X-J (2005) The stability of a stochastic CaMKII switch: Dependence on the number of molecules and protein turn over. *PLoS* 3: 705–717.

O'Connor DH, Wittenberg GM, and Wang SS (2005) Graded bidirectional synaptic plasticity is composed of switch-like unitary events. *Proceedings of the National Academy of Sciences of the United States of America* 102: 9679–9684.

Seung HS (1996) How the brain keeps the eyes still? *Proceedings of the National Academy of Sciences of the United States of America* 93: 13339–13344.

Strogatz SH (1994) *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Reading, MA: Addision-Wesley.

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

Wang X-J (2001) Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neuroscience* 24: 455–463.

Wang X-J (2002) Probabilistic decision making by slow reverberation in neocortical circuits. *Neuron* 36: 955–968.

Zhang K (1996) Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *Journal of Neuroscience* 16: 2112–2126.