OF THE EVOLUTION OF THE BRAIN

Alessandro Treves[§,¶] and Yasser Roudi[§]

[§] *SISSA, Cognitive Neuroscience sector, Trieste, Italy*
[¶] *NTNU, Center for the Biology of Memory, Trondheim, Norway*

Photo: width 7.5cm height 11cm

Contents

1. Introduction and summary

We review the common themes, the network models and the mathematical formalism underlying our recent studies about different stages in the evolution of the human brain. The first pair of studies both deal with radical changes in neuronal circuitry presumed to have occurred at the transition from early reptilians to mammals, introduced in sect. 2: the lamination of sensory cortex (sect. 4) and the differentiation into sub-fields of the mammalian hippocampus (sect. 5). In neither case the qualitative structural change seems to be accompanied by an equally dramatic functional change in the operation of those circuits. The last study, introduced in sect. 6, deals, instead, with the neuronal dynamics that might underlie the faculty for language in the human frontal lobes, a qualitatively new functional capacity that is not apparently associated with any new structural feature. These studies therefore all discuss the evolution of cortical networks in terms of their computations, quantified by simulating simplified formal models. All such models can be conceived as variants of a basic autoassociative neural network model, and their storage capacity, even when not formally analyzed, plays an important role in the results. We thus sketch, in sects. 3 and 7, the formalism that leads to storage capacity calculations, particularly in view of the fact that all three studies dwell on the interrelationship between qualitative and quantitative change, and all would benefit from more detailed mathematical analysis. Moreover, all studies include, as a necessary ingredient of the relevant computational mechanism, a simple feature of pyramidal cell biophysics: firing rate adaptation; a feature which to be treated properly requires extending the thermodynamics formalism into a full analysis of network dynamics. Overall, our approach is that while there is not necessarily a coupling between structural and functional phase transitions, understanding both at the mechanistic neural network level is a necessary step to understand the evolution of the organ of thought.

2. The phase transition that made us mammals

Mammals originate from the therapsids, one order among the first amniotes, or early reptiles, as they are commonly referred to. They are estimated to have radiated away from other early reptilian lineages, including the anapsids (the progenitors of modern turtles) and diapsids (out of which other modern reptilians, as well as birds, derive) some three hundred million years ago [25]. Perhaps mammals emerged as a fully differentiated class out of the third-to-last of the great extinctions, in the Triassic period. The changes in the organization of the nervous system, that mark the transition from proto-reptilian ancestors to early mammals, can be reconstructed only indirectly. Along with supporting arguments from the examination of endocasts (the inside of fossil skulls; [54]) and of presumed behavioural patterns [103], the main line of evidence is the comparative anatomy of present day species [32]. Among a variety of *quantitative* changes in the relative development of different structures, changes that have been extended, accelerated and diversified during the entire course of mammalian evolution [40], two major *qualitative* changes stand out in the forebrain, two new features that, once established, characterize the cortex of mammals as distinct from that of reptilians and birds. Both these changes involve the introduction of a new “input” layer of granule cells.

In the first case, it is the medial pallium (the medial part of the upper surface of each cerebral hemisphere, as it bulges out of the forebrain) that reorganizes into the modern-day mammalian hippocampus. The crucial step is the detachment of the most medial portion, that loses both its continuity with the rest of the cortex at the hippocampal sulcus, and its projections to dorso-lateral cortex [99]. The rest of the medial cortex becomes Ammon’s horn, and retains the distinctly cortical pyramidal cells, while the detached cortex becomes the dentate gyrus, with its population of granule cells, that project now, as a sort of pre-processing stage, to the pyramidal cells of field CA3 [8]. In the second case, it is the dorsal pallium (the central part of the upper surface) that reorganizes internally, to become the cerebral neocortex. Aside from special cases, most mammalian neocortices display the characteristic isocortical pattern of lamination, or organization into distinct layers of cells (traditionally classified as 6, in some cases with sublayers). The crucial step, here, appears to be the emergence, particularly evident in primary sensory cortices, of a layer of non-pyramidal cells (called spiny stellate cells, or granules) inserted in between the pyramidal cells of the infragranular and supragranular layers. This is layer IV, where the main ascending inputs to cortex terminate [33].

2.1. An information-theoretical advantage in the hippocampus

What is the evolutionary advantage, for mammals, brought about by these changes? In the case of the hippocampus, attempts to account for its remarkable internal organization have been based, since the seminal paper by David Marr [70], on the computational analysis of the role of the hippocampus in memory. The hippocampus is important for spatial memory also in birds. A reasonable hypothesis is that the "invention" of the dentate gyrus enhances its capability, in mammals, to serve as a memory store. Following the approach outlined by David Marr, it was proposed 12 years ago [90] that the new input to CA3 pyramidal cells from the mossy fibers (the axons of the dentate granule cells) serves to create memory representations in CA3 richer in information content than they could have been otherwise. The crucial prediction of this proposal was that the inactivation of the mossy fiber synapses should impair the formation of new hippocampal dependent memories, but *not* the retrieval of previously stored ones. This prediction has recently been supported [61] at the behavioural level in mice, while neurophysiological experiments are in progress with rats. If validated, this hypothesis suggests that indeed a quantitative, information-theoretical advantage may have favored a qualitative change, such as the insertion of the dentate gyrus in the hippocampal circuitry. This possibility raises the issue of whether also the insertion of layer IV in the isocortex might be accounted for in quantitative, information-theoretical terms, an issue discussed in section 4. At the same time, the DG argument does not itself address the CA3-CA1 differentiation, which is equally prominent in the mammalian hippocampus. Section 5 will review a computational approach to this problem, and mention fresh experimental results that are shading an entirely new light on it.

2.2. An information-theoretical hypothesis about layers and maps

It has long been hypothesized that isocortical lamination appeared together with fine topography in cortical sensory maps [6], pointing at a close relationship between the two phenomena. All of the cortex, which develops from the upper half of the neural tube of the embryo, has been proposed to have been, originally, sensory, with the motor cortex differentiating from the somatosensory portion [34, 65]. In early mammals, the main part of the cortex was devoted to the olfactory system, which is not topographic, and whose piriform cortex has never acquired isocortical lamination [45]. The rest of the cortex was largely allocated to the somatosensory, visual and auditory system, perhaps with just one topographic area, or map, each [32]. Each sensory map thus received its inputs directly from a corresponding portion of the thalamus, as opposed to the network of cortico-cortical connections which has been greatly expanded [2, 22] by the evolution of multiple, hierarchically organized cortical areas in each sensory system [56,59]. In the thalamus, a distinction has been drawn [55] between its matrix and core nuclei. The matrix, the originally prevalent system, projects diffusely to the upper cortical layers; while the core nuclei, which specialize and become dominant in more advanced species [38], project with topographic precision to layer IV, although their axons contact, there, also the dendrites of pyramidal cells whose somata lie in the upper and deep layers.

The crucial aspect of fine topography in sensory cortices is the precise correspondence between the location of a cortical neuron and the location, on the array of sensory receptors, where a stimulus can best activate that neuron. Simple visual and somatosensory cortices thus comprise 2D maps of the retina and of the body surface, while auditory cortices map sound frequency in 1 dimension, and what is mapped in the other dimension is not quite clear [78]. Some of the parameters characterizing a stimulus, those reflected in the position of the receptors it activates, are therefore represented continuously on the cortical sheet. We define them as providing *positional* information. Other parameters, which contribute to identify the stimulus, are not explicitly mapped on the cortex. For example, the exact nature of a tactile stimulus at a fixed spot on the skin, whether it is punctuate or transient or vibrating, and to what extent, are reflected in the exact pattern of activated receptors, and of activated neurons in the cortex, but not directly in the position on the cortical sheet. We define these parameters as providing *identity* information. Advanced cortices, like the primary visual cortex of primates, include complications due to the attempt to map additional parameters on the sheet, like ocular dominance or orientation, in addition to position on the retina. This leads to the formation of so-called columns, or wrapped dimensions, and to the differentiation of layer IV in multiple sublayers. They should be regarded as specializations, which likely came much after the basic cortical lamination scheme had been laid out. The sensory cortices of early mammals therefore received from the thalamus, and had to analyse, information about sensory stimuli of two basic kinds: positional or *where* information, I_p, and identity or *what* information, I_i. These two kinds differ also in the extent to which cortex can contribute to the analysis of the stimulus. Positional information is already represented explicitly on the receptor array, and then in the thalamus, and each relay stage can only degrade it. At best, the cortex can try to maintain the spatial resolution with which the position of a stimulus is specified by the activation of thalamic neurons: if these code it inaccurately, there is no way the cortex can reconstruct it any better, because any other position would be just as plausible. The identity of a stimulus, however, may be coded inaccurately by the thalamus, with considerable noise, occlusion and variability, and the cortex can reconstruct it from such partial information. This is made possible by the storage of previous sensory events in terms of distributed efficacy modifications in synaptic systems, in particular on the recurrent collaterals connecting pyramidal cells in sensory cortex. Neural network models of autoassociative memories [53, 70] have demonstrated how simple "Hebbian" rules modelling associative synaptic plasticity can induce weight changes that lead to the formation of dynamical attractors [9]. Once an attractor has been formed, a partial cue corresponding *e.g.* to a noisy or occluded version of a stimulus can take the recurrent network within its basin of attraction, and hence lead to a pattern of activation of cortical neurons, which represents the stored identity of the original stimulus. Thus by exploiting dishomogeneities in the input statistics - some patterns of activity, those that have been stored, are more "plausible" than others - the cortex can reconstruct the identity of stimuli, over and beyond the partial information provided by the thalamus. This analysis of current sensory experience in the light of previous experience is hypothesized here to be the generic function of the cortex, which thus blends perception with memory [101]. Specialized to the olfactory sense, this function does not seem to require new cortical machinery to be carried out efficiently. A novel circuitry may instead be advantageous, when the generic function is specialized to topographic sensory systems, which have to relay both where and what information, I_p and I_i. We take up the validation of this possibility after considering in the next section a fully defined model, which exemplifies the mathematical structures underlying our arguments.

3. Maps and patterns of threshold-linear units

The notion of autoassociative networks refers to a family of neuronal architectures, which in the simplest way can be thought of as one of the three main building blocks of cortical networks [82]. The two others are Pattern Associators and Competitive networks. By an autoassociative network we refer to a recurrent neuronal network with plastic connections. As briefly mentioned previously, associatively modifiable synapses, which might be modeled by a simple Hebbian plasticity mechanism, together with massive recurrent connections give a network of neurons the ability to function as a content addressable memory device.

In the past two decades, physicists have studied various models of autoassociative memory using different model neurons and different “learning rules” to implement Hebbian learning [12, 53]. Mathematical methods have been adapted from statistical and spin glass physics for the purpose of analyzing these neuronal networks [7]. Although most of these investigation have been made on very abstract and simplified models, they have provided us with a good understanding of the general properties of associative memories, *e.g.* storage capacity and retrieval dynamics. Basically, the methods borrowed from physics are based on the assumption of the existence of a Hamiltonian describing the dynamics of the system. The condition of the existence of a Hamiltonian imposes the important constraint of symmetric interactions on the network; this may be taken to be a good first approximation, but obviously it is not satisfied in the cortex. Actually, cortical networks belong to a subclass of asymmetrically wired networks, in which connections are not just asymmetric but, in addition, nearby neurons are more likely to make synapses with each other. This kind of more realistic models in terms of connectivity is what we want to briefly introduce in this section. To sketch the analytical treatment, we use an improved version of the Self-consistent signal-to-noise analysis [85].

We thus introduce and analyze an autoassociative network which is comprised of threshold-linear units and includes a geometrical organization of neuronal connectivity, meant as a simplistic model of the type of organization of connections observed in the cortex.

3.1. A model with geometry in its connections

Consider a network of N units. The level of activity of unit i is a variable $v_i \geq 0$, which corresponds to the firing rate of the neuron. We assume that each unit receives $C \ll N$ inputs from the other units in the network. The specific covariance 'Hebbian' learning rule we consider prescribes that the synaptic weight between units i and j be given as:

$$J_{ij} = \frac{1}{Ca^2} \sum_{\mu=1}^{p} c_{ij} \left(\eta_i^\mu - a\right) \left(\eta_j^\mu - a\right) , \tag{3.1}$$

where η_i^μ represents the activity of unit i in pattern μ, c_{ij} is a binary variable and is equal to 1 if there is a connection running from neuron j to the neuron i and 0 otherwise. Each η_i^μ is taken to be a quenched variable, drawn independently from a distribution $p(\eta)$, with the constraints $\eta \geq 0$, $\langle\eta\rangle = \langle\eta^2\rangle = a$, where $\langle\rangle$ stands for the average over the distribution of η. Here we concentrate on the binary coding scheme $p(\eta) = a\delta(\eta - 1) + (1 - a)\delta(\eta)$, but the calculation can be carried out for any probability distribution. As in one of the first extensions of the Hopfield model [98], we thus allow for the mean activity a of the patterns to differ from the value $a = 1/2$ of the original model [87]. We further assume that the input (local field) to unit i takes the form

$$h_i = \sum_{j \neq i} J_{ij} v_i + b \left(\frac{1}{N} \sum_j v_j \right) , \tag{3.2}$$

where the first term enables the memories encoded in the weights to determine the dynamics; the second term is unrelated to the memory patterns, but is designed to regulate the activity of the network, so that at any moment in time $x \equiv \frac{1}{N}\sum_i v_i = \frac{1}{N}\sum_i v_i^2 = a$. The activity of each unit is determined by its input through a threshold-linear function

$$v_i = F[v_i] = g(h_i - T_{thr})\Theta(h_i - T_{thr}) \tag{3.3}$$

where T_{thr} is a threshold below which the input elicits no output, g is a gain parameter, and $\Theta(...)$ the Heaviside step function. The exact details of the updating rule are not specified further, here, because they do not affect the steady states of the dynamics, and we take "fast noise" levels to be vanishingly small, $T \to 0$. Discussions about the biological plausibility of this model for networks of pyramidal cells can be found in [11, 88], and will not be repeated here.

In order to analyze this network, we first define a set of order parameters $\{m_i^\mu\}$, with $\mu = 1 \ldots p; i = 1 \ldots N$, which we call the *local overlaps*, as follows:

$$m_i^\mu = \frac{1}{C} \sum_j c_{ij} (\eta_j^\mu / a - 1) v_j, \qquad (3.4)$$

If we rewrite the local field h_i defined above in terms of these order parameters we have:

$$h_i = \sum_\mu (\eta_i^\mu / a - 1)\, m_i^\mu - c_{ii} \alpha (1/a - 1) v_i \qquad (3.5)$$

in which $\alpha = p/C$ is the storage load. We will use this identity for the local field in the next section.

3.2. Retrieval states

A pattern μ is said to be retrieved if $\sum_i m_i^\mu = O(C)$. Without loss of generality, we suppose that the first pattern is the retrieved one and therefore $m_i^\nu \ll m_i^1$ for $\nu \neq 1$ and any i. When one pattern is retrieved, the local field to each unit can be decomposed into two terms. One is the *signal*, which is in the direction of keeping the network in a state with large overlap with the retrieved pattern. The second term, which we call *noise*, does the opposite. The idea is to calculate these terms as a function of the local overlap with the retrieved pattern. In other words we wish to express the r.h.s of 3.5 solely as a function of $m_i = m_i^1$ and η_i^1. If we are able to do so then we can calculate the activity of each unit as a function of m_i and by using it in the definition of local overlaps, we will be able to find a self consistent equation for the local overlap with the first pattern.

To proceed further, we define two more order parameters z_i^μ and γ through the equality below:

$$\sum_{\nu \neq 1, \mu} (\eta_i^\nu / a - 1) m_i^\nu = z_i^\mu + \gamma_i v_i \qquad (3.6)$$

with this, we can write the activity of the network as:

$$v_i = F[(\eta_i^1/a - 1) m_i^1 + (\eta_i^2/a - 1) m_i^\mu + z_i^\mu + \gamma_i v_i - c_{ii} \alpha (1/a - 1) + b(x) - T_{thr}] \qquad (3.7)$$

from which v_i can be found self consistently:

$$v_i = G[(\eta_i^1/a - 1)m_i^1 + (\eta_i^2/a - 1)m_i^\mu + z_i^\mu + b(x) - T_{thr}] \tag{3.8}$$

Assuming that $\Gamma_i = \gamma_i - c_{ii}\alpha(1/a - 1) < 1/g$ [1] the function $G[x]$ is for a threshold-linear unit:

$$G[x] = \frac{g}{1 - g\Gamma} x\Theta(x) \tag{3.9}$$

now we expand the r.h.s. of the above equation for v_i up to the linear term in m_i^μ and insert the result in 3.4, to get:

$$m_i^\mu = L_i^\mu + \sum_j K_{ij}^\mu m_j^\mu \tag{3.10}$$

where:

$$\begin{aligned} L_i^\mu &= \frac{1}{C}\sum_j c_{ij}(\eta_j^\mu/a - 1)G[(\eta_j^1/a - 1)m_j^1 + \widehat{z_j^\mu} + b(x) - T_{thr}] && (3.11)\\ K_{ij}^\mu &= \frac{c_{ij}}{C}(\eta_j^\mu/a - 1)^2 G'[(\eta_j^1/a - 1)m_j^1 + \widehat{z_j^\mu} + b(x) - T_{thr}]. && (3.12)\end{aligned}$$

For the above equation, the solution for m_i^μ can be approximated as:

$$m_i^\mu = \frac{1}{C}R_{ii}^\mu(\eta_i^\mu/a - 1)G^\mu[i] + \frac{1}{C}\sum_{j\neq i} R_{ij}^\mu(\eta_j^\mu/a - 1)G^\mu[j] \tag{3.13}$$

where R_{ij}^μ is defined as:

$$R_{ij}^\mu = c_{ij} + \sum_l K_{il}^\mu c_{lj} + \sum_{lt} K_{il}^\mu K_{lt}^\mu c_{tj} + \ldots \tag{3.14}$$

in which we have used the notation $G^\mu[i] = G[(\eta_i^1/a - 1)m_i^1 + z_i^\mu + b(x) - T_{thr}]$.

Now that we have the local overlaps with the non-condensed patterns as a function of m_i^1, we can write the noise also as a function of it:

$$\begin{aligned} \sum_{\nu\neq\mu,1}(\eta_i^\mu - 1)m_i^\nu &= \frac{1}{C}\sum_{\nu\neq\mu,1} R_{ii}^\nu(\eta_i^\nu/a - 1)^2 G^\nu[i] && (3.15)\\ &+ \frac{1}{C}\sum_{j\neq i,\nu\neq\mu,1} R_{ij}^\nu(\eta_i^\mu/a - 1)(\eta_j^\mu/a - 1)G^\mu[j] \end{aligned}$$

[1] We shall see later that this is a reasonable assumption at least when one deals with diluted networks or very low storage loads.

for the first sum in the r.h.s of Eq.3.15 above, using the independence of different patterns and assuming that z_i^μ does not depend on η_i^μ, one can write:

$$\begin{aligned} \frac{1}{C}\sum_{\nu\neq\mu,1} R^{\nu}_{ii}(\eta^{\nu}_i/a-1)^2 G^{\nu}[i] &= \alpha\langle R^{\nu}_{ii}(\eta^{\nu}_i/a-1)^2 G^{\nu}[i]\rangle \\ &= \alpha\langle R^{\nu}_{ii}(\eta^{\nu}_i/a-1)^2\rangle v_i \end{aligned} \tag{3.16}$$

and as a result of this we have:

$$\gamma_i = \alpha < R^{\nu}_{ii}(\eta^{\nu}_i/a-1)^2 >= \alpha T_0\langle R^{\nu}_{ii}\rangle, T_0 = 1/a-1 \tag{3.17}$$

and therefore:

$$\Gamma_i = \frac{\alpha T_0^2}{C}\sum_j c_{ij}c_{ji}\langle G'[j]\rangle. \tag{3.18}$$

The second term is a bit tricky. For this term, by replacing the sum with the average we get zero mean, but for its deviation we have:

$$\rho^2 = \frac{\alpha}{C}(1/a-1)\sum_j c_{ij}\langle {R^{\mu}_{ij}}^2(\eta^{\mu}_j/a-1)^2 G^{\mu}[j]^2\rangle \tag{3.19}$$

which is, actually, the standard deviation of the noise. Now we replace the second term, in the noise sum corresponding to z_i^μ, with a Gaussian random variable with mean zero and standard deviation ρ, and take it into account in our fixed point equations by averaging the equations over this Gaussian measure.

Having done so, with some mathematical manipulations and considering the assumption that $C/N \ll 1$ we derive the following fixed point equations:

$$\begin{aligned} \psi_{ij} &= \frac{gT_0}{C}\sum_k c_{ik}c_{kj}\langle\int^+ Dz(1-g\Gamma_k)^{-1}+\ldots\rangle \\ \Gamma_i &= \alpha T_0\psi_{ii} \\ \rho_i^2 &= \frac{\alpha g^2 T_0^2}{C}\sum_j \left(c_{ij}+2c_{ij}\psi_{ij}+\psi_{ij}^2\right)\times \\ & \quad \langle\int^+ Dz\left((\frac{\eta_j}{a}-1)m_j+b(x)-T_{thr}-\rho_j z\right)^2(1-g\Gamma_j)^{-2}\rangle \end{aligned} \tag{3.20}$$

where $Dz = dz\frac{e^{-z^2/2}}{\sqrt{2\pi}}$ and the superscript $+$ indicates that the integration has to be carried out in the range where $(\frac{\eta_i}{a}-1)m_i + b(x) - T > \rho_i z$. Using the definitions of m_i and x we can get the following for their corresponding fixed point equations:

$$\begin{aligned} m_i &= \frac{g}{C}\sum_j c_{ij}(\eta_j/a-1) \times \\ & \quad \int^{+} Dz\left((\frac{\eta_j}{a}-1)m_j+b(x)-T_{thr}-\rho_j z\right)(1-g\Gamma_j)^{-1} \qquad (3.21) \\ x &= \frac{g}{N}\sum_j \langle \int^{+} Dz\left((\frac{\eta_j}{a}-1)m_j+b(x)-T_{thr}-\rho_j z\right)(1-g\Gamma_j)^{-1}\rangle . \end{aligned}$$

3.3. The network without structure

Assume that the c_{ij}'s are randomly generated with probability $Pr\{c_{ij}=1\}=C/N$. When $C/N\to 0$ the network is said to be in the highly diluted regime and the case of $C/N=1$ corresponds to the fully connected network. Of course in these cases where the connectivity is randomly drawn from a non-geometric probability distribution, the order parameters become uniform in space and solutions have no spatial dependence. It can be shown that for the network without geometry the mean-field equations read:

$$\begin{aligned} \Omega &= \frac{gT_0}{1-g\Gamma} < \int^{+} Dz > \\ \psi &= \frac{C}{N}\left(\Omega+\Omega^2+\Omega^3+\ldots\right) \\ \Gamma &= \alpha T_0 \psi \\ \rho^2 &= \alpha\left(\frac{gT_0}{(1-g\Gamma)}\right)^2\left(1+2\psi+\frac{N}{C}\psi^2\right) \times \qquad (3.22) \\ & \quad \langle \int^{+} Dz\left((\frac{\eta}{a}-1)m+b(x)-T_{thr}-\rho z\right)^2\rangle \\ m &= \frac{g}{1-g\Gamma}\langle \int^{+} Dz(\eta/a-1)\left((\frac{\eta}{a}-1)m+b(x)-T_{thr}-\rho z\right)\rangle \\ x &= \frac{g}{1-g\Gamma}\langle \int^{+} Dz\left((\frac{\eta}{a}-1)m+b(x)-T_{thr}-\rho z\right)\rangle . \end{aligned}$$

It is worthnoting that the contribution of the activity reverberating in the loops of the network is measured by the order parameter ψ. Also Γ essentially measures the effect of the activity of each unit on itself, after it has reverberated through the network. The fact that these order parameter disappears when $C/N=0$ reflects that when one considers a highly diluted network, the number of loops becomes negligible, and they do not contribute to network dynamics. This also makes the inequality $\Gamma_i < 1/g$ a valid assumption, and this effect becomes negligible when one deals with an extremely diluted network.

We can then define the new variables $r = m/\rho$ and $w = [b\,(x) - m - T_{thr}]/\rho$ and the following integrals, which are functions of r and w, as in [87]:

$$\begin{aligned}
A_2 &= \frac{1}{rT_0}\langle(\frac{\eta}{a}-1)\int^{+} Dz(w+\frac{r\eta}{a}-z)\rangle \\
A_1 &= A_2 - \langle\int^{+} Dz\rangle \qquad (3.23)\\
A_3 &= \langle\int^{+} Dz(w+\frac{v\eta}{a}-z)^2\rangle.
\end{aligned}$$

By using this notation the mean-field equations can be reduced to:

$$E_1(r,w) = A_2^2 - \left(1+\frac{C}{N}\left(\frac{(2-\Omega)\Omega}{(1-\Omega)^2}\right)\right)\alpha A_3 = 0 \qquad (3.24)$$

$$E_2(r,w) = (\frac{1}{gT_0} - \alpha\frac{C\Omega}{N(1-\Omega)}) - A_2 = 0 \qquad (3.25)$$

which extend and interpolate the results of [88] to finite values of C/N.

The first equation above appears as a closed curve in the (w, r), plane, which shrinks in size when one increases α and then disappears; whereas the second equation is an almost straight curve, which for a certain range of g intersects twice with the closed curve above. Since for a given value of α such that the first equation is satisfied, there always exists a value for g that satisfies the second equation, the storage capacity is the value of α for which the closed curve shrinks to a point. We treat g as a free parameter, because it can be easily changed in a network by mechanisms like multiplicative inhibition, if required in order to approach the optimal storage load.

In the limit of extreme dilution, *i.e.* $C/N \to 0$, Ω does not contribute to the equation for the storage capacity. The result of calculating the storage capacity as a function of the sparseness of the coding is shown in Fig.2 (the full curve). For other values of C/N the contribution from Ω should be taken into account, which for small $C/N \neq 0$ results in deviations from the storage capacity of a highly diluted network. An example is illustrated in in Fig.2 for $C/N = 0.05$. It is clear that, at least for small a, a network with 5% connectivity can be considered as highly diluted, in the sense that for sparse patterns of activity, the effect of loops – what produces the difference between A_2 and A_1 – becomes unimportant.

An equivalent approach to study such a network is to use the replica method, from spin glass physics. If one considers a fully connected network with symmetric connections, then the dynamics can be described by a Hamiltonian. Using this Hamiltonian it is possible to calculate the partition function, and therefore the mean field equations, *e.g.* for the fully connected version of this model. Then one sends the order parameter corresponding to ψ for the fully connected network to zero, to obtain the extremely diluted limit. This was basically the way the threshold linear network was first solved. One can look at [87, 89] for details of the calculations.

3.4. *Appearance of bumps of activity*

If we consider a network with a low connectivity level which is spatially organized, there can exist solutions of the fixed point equations that are spatially non-uniform. This is what one might call pattern formation. An interesting case, in one dimension, is a network with a Gaussian connectivity probability distribution:

$$Pr\{c_{ij} = 1\} = \frac{C}{\sqrt{2\pi\sigma^2}} e^{\frac{-(i-j)^2}{2\sigma^2}} + \mathrm{Baseline}. \tag{3.26}$$

The baseline is considered for $\sigma \propto N$. In this network it can be shown that there exists a critical σ at which a second order phase transition occurs, to the appearance of spatially non-uniform solutions (more precisely, the first Fourier mode). Together with this appearance of non-uniform solutions, one can observe a sort of decrease in the storage capacity. Decreasing σ further[2] results in the appearance of bumps of activity, *i.e.* fixed points of the dynamics that have large overlap with the stored pattern, and on the other hand are localized in space. An example of such bumps is shown in Fig.3. The dependence of the critical sigma and the properties of the bumps are beyond the scope of this paper and are being reported elsewhere [84], but what is important for us at this stage is the existence and stability of these spatially non-uniform retrieval states, which can be analyzed using the above formalism.

3.5. *The main points*

Let us summarize the main results of the model discussed above, which are relevant to the forthcoming sections. The first point is the way the critical storage capacity scales with the relevant parameters of the model. As we stated before, this model shows that for a diluted network, which is close to a biologically plausible

[2] It should be noted that in the case of small σ the approximations leading to disappearance of ψ and Γ from our equations are not applicable, since the loops become important again. One can use these equations in the case of small α, where ρ becomes zero and so does Γ, and the effects of the loops become unimportant. We will discuss this issue in more detail elsewhere.

structure, we obtain a relation $p_c \propto C/a$. This expresses the computational advantage of sparse coding for memory storage, in structures like the hippocampus (see sect. 4). The second point is the appearance of bumps of retrieval activity, *i.e.* sustained activity localized both in physical space and in the space of stored pattern. This phenomenon, analysed in the more complicated situation of a partially recurrent and partially feedforward network comprised of multiple layers, and operating under the influence of sustained external input, is the basis of the results of [95], reported in the next section.

4. Validation of the lamination hypothesis

Does preserving accurate coding of position, in an isocortical patch, conflict with the analysis of stimulus identity? This is obviously a quantitative question, which has to be addressed with a suitable neural network model. An appropriate model can be designed with similar features as the one considered above, but with the additional option of differentiating multiple layers. In particular, the model of a cortical patch, receiving inputs from a thalamic array of units, can be investigated in its ability to generate localized retrieval states, that correspond to the stored patterns modulated by bumps, studied analytically in the section above. Unlike the analytical study, which is easier to conduct in a well defined limit case, e.g. looking at the existence of asymptotic attractor states after an afferent cue that has initialized activity in the network has been removed, with simulations one can also study the dynamics of localization and retrieval in time, with a cue that follows its own time course. Contrasting a model network with differentiated layers with one that has the same number of units and connections, but statistically equivalent layers, allows to approach the question of the role of lamination. The presence of several layers would anyway make an analytical treatment, while not impossible, very cumbersome, and computer simulations appear to be the method of choice. This is the approach taken in Ref. [95], the results of which are briefly summarized here.

A patch of cortex was modeled as a wafer of 3 arrays, each with $N \times N$ units. Each unit receives C_{ff} feedforward connections from a further array of $N \times N$ "thalamic" units, and C_{rc} recurrent connections from other units in the patch. Both sets of connections are assigned to each receiving unit at random, with a Gaussian probability in register with the unit itself, and of width S_{ff} and S_{rc}, respectively[3]. To model, initially, a *uniform*, non-laminated patch, the 3 arrays are identical in properties and connectivity, so the C_{rc} recurrent connections each unit receives are drawn at random from all arrays. To model a laminated patch,

[3] Periodic boundary conditions are used, to limit finit size effects, so the patch is in fact a torus.

later, different properties and connectivity are introduced among the arrays, but keeping the same number of units and connections, to provide for a correct comparison of performance. The 3 arrays will then model supragranular, granular and infragranular layers of the isocortex [95]. A local pattern of activation is applied to the thalamic units, fed forward to the cortical patch and circulated for N_{iter} time steps along the recurrent connections, and then the activity of some of the units in the patch is read out. To separate out "what" and "where" information, the input activation is generated as the product of one of a set of M predetermined global patterns, covering the entire $N \times N$ input array, by a local focus of activation, defined as a Gaussian tuning function of width R, centered at any one of the N^2 units. The network operates in successive *training* and *testing* phases. In a training phase, each of the possible $M \times N \times N$ activations is applied, in random sequence, to the input array; activity is circulated in the output arrays, and the resulting activation values are used to modify connections weights according to a model associative rule. In a testing phase, input activations are the product of a focus, as for training, by a *partial cue*, obtained by setting a fraction of the thalamic units at their activation in a pattern, and the rest at a random value, drawn from the same general distribution used to generate the patterns. The activity of a population of output units is then fed into a decoding algorithm - external to the cortical network - that attempts to predict the actual focus (its center, p) and, independently, the pattern i used to derive the partial cue. I_i is extracted from the frequency table $P(i, i_d)$ reporting how many times the cue belonged to pattern $i = 1, \ldots, M$ but was *d*ecoded as pattern i_d:

$$I_i = \sum_{i,i_d} P(i, i_d) \log_2 \frac{P(i, i_d)}{P(i)P(i_d)} \tag{4.1}$$

and a similar formula is used for I_p. The learning rule used to modify connection weights was

$$\Delta w_{ij} \propto r_j^{post} \cdot (r_i^{pre} - < r^{pre} >) \tag{4.2}$$

applied, at each presentation of each training phase, to weight w_{ij}. Weights are originally set at a constant value (normalized so that the total strength of afferents equals that of recurrent collaterals), to which is added a random component of similar but asymmetrical mean square amplitude, to generate an approximately exponential distribution of initial weights onto each unit. r denotes the firing rates of the pre- and postsynaptic units, and $< \ldots >$ an average over the corresponding array.

Among the several parameters that determine the performance of the network, $R << S_{rc}$ was fixed, while S_{ff} was varied from $S_{ff} \simeq R$ up to $S_{ff} \simeq S_{rc}$. It is intuitive that if the feedforward connections are focused, $S_{ff} \simeq R$, “where” information can be substantially preserved, but the cortical patch is activated over a limited, almost point-like extent, and it may fail to use efficiently its recurrent collaterals to retrieve "what" information. If the other hand $S_{ff} \simeq S_{rc}$, the recurrent collaterals can better use their attractor dynamics, leading to higher I_i values, but the spread of activity from thalamus to cortex means degrading I_p. This conflict between I_p and I_i is depicted in Fig. 4, which reports their joint values extracted from simulations, as they vary as a function of the spread of the afferents, at the end of the training phase (full curve). What is decoded is the activity of all units in the upper array of the patch. Since the patch is not differentiated, however, the other two arrays provide statistically identical information. Further, since information of both the what and where kinds is extracted from a number of units already well in the saturation regime [94], even decoding all units in all 3 arrays at the same time, or only, say, half of the units in any single array, does not alter the numbers significantly. I_i is monotonically increasing with S_{ff}. I_p, instead, decreases with S_{ff}, and as a result one can vary S_{ff} to select a compromise between what and where information, but *not optimise both* simultaneously. This conflict between what and where persists whatever the choice of all the other parameters of the network, although of course the exact position of the $I_p - I_i$ limiting boundary varies accordingly. Is it possible to go beyond such boundary?

4.1. Differentiation among isocortical layers

Several modifications of the "null hypothesis" uniform model were explored, as reported in [95]. Figure 4 illustrates, along with the results of the uniform model, results pertaining to slightly different versions of a 3-layer laminated model. Basically, the granular layer is differentiated by (i) focusing the thalamic afferents to the granular layer, while those to the two pyramidal layers are diffuse; (ii) restricting the recurrent collateral system of the granular units, by focusing the connections departing from granular units and decreasing the number of connections arriving at layer IV from the pyramidal layers; finally (iii) layer IV units follow a non-adaptive dynamics, and they do not operate during training, but only during testing. The non-adapting dynamics is effected, in the simulations by making their effect on postsynaptic units, whatever their layer, scale up linearly with iteration cycle. Thus, compared to the model pyramidal units, whose firing rate would adapt over the first few interspike intervals, in reality (but is kept in constant ratio to the input activation, in the simulations), the firing rate of granule units, to model lack of adaptation, is taken to actually increase in time for a given input activation.

Differentiating infra- from supra-granular connections is effected by simply replacing the connections from layer IV to the infragranular pyramidal units with connections to the same units from supragranular units. In the real cortex, the supragranular layers project mainly onward, to the next stage of processing. The infragranular layers project mainly backward [14], or subcortically. Among their chief target structures are the very thalamic nuclei from which projections arise to layer IV. It is clear that having different preferential targets would in principle favour different mixes of what and where information. In particular, cortical units that project back to the thalamus would not need to repeat to the thalamus “where” a stimulus is, since this information is already coded, and more accurately, in the activity of thalamic units. They would rather report in its full glory the genuine contribution of cortical processing, that is, the retrieval of identity information. Units that project to further stages of cortical processing, on the other hand, should balance the "what" added value with the preservation of positional information. With this combination of modifications, layer III becomes the main source of recurrent collaterals [73, 104], which are spread out and synapse onto both supra- and infra-granular units and also, to a lesser degree, layer IV units.

The effect of the overall model of the differentiation can be appreciated by decoding the activity in the three layers, separately, as shown in Fig. 4 by the isolated symbols. From layer IV one can extract a large I_p but limited I_i; from layer III one obtains a balanced mix. From layer V, on the other hand, one can extract predominantly "what" information, I_i, at the price of a rather reduced I_p content. Thus, the last connectivity change, by effectively reducing the coupling between granular and infragranular layers, has made the latter optimize “what” information, while neglecting “where” information, of limited interest to their target structures. Can we understand the advantage brought about by lamination? The modifications required in the connectivity of layer IV are intuitive: they make granule units more focused in their activation, in register with the thalamic focus, while allowing the pyramidal units, that receive diffuse feedforward connections, to make full use of the recurrent collaterals. What is less intuitive is the requirement for non-adapting dynamics in the granule layer. It turns out that without this modification in the dynamics, the laminated network essentially averages linearly between the performances of uniform networks with focused and with diffuse connectivity, without improving at all on a case with, say, intermediate spread parameters for the connections. This is because the focusing of the activation and the retrieval of the correct identity interfere with each other, if carried out simultaneously, even if the main responsibility for each task is assigned to a different layer. Modifying the dynamics of the model granules, instead, enables the recurrent collaterals of the pyramidal layers to first better identify the attractor, i.e. the stored global pattern, to which the partial cue “belongs”, and to start the dynamical convergence towards the bottom of the corresponding basin of attraction [7]. Only *later on*, once this process is – in most cases – safely underway, the granules make their focusing effect felt by the pyramidal units. The focusing action, by being effectively delayed after the critical choice of the attractor, interferes with it less – hence, the non-linear advantage of the laminated model.

5. What do we need DG and CA1 for?

If the synapses on the recurrent collaterals among pyramidal cells of the primitive cortex were endowed, as it is likely, with associative, "Hebbian", plasticity, such as that based on NMDA receptors [28], that cortex could have operated as an associative memory [22] – provided it had an effective way of distinguishing its operating modes. A generic problem with associative memories based on recurrent collaterals is to distinguish a storage mode from a retrieval mode. To be effective, recurrent collaterals should dominate the dynamics of the system when it is operating in retrieval mode; whereas while storing new information the dynamics should be primarily determined by afferent inputs, with limited interference from the memories already stored in the recurrent collaterals. The recurrent collaterals, instead, should modify their weights to store the new information [90]. In the model considered analytically in section 3, the learning phase is not explicitly considered. In the simulations of the laminated model, in section 4, the distinction is partially inserted by hand, by forcing the layer IV units to be silent during training.

5.1. Distinguishing storage from retrieval

The most phylogenetically primitive solution to achieve a similar effect is to use a modulator which acts differentially on the afferent inputs (originally, those arriving at the apical dendrites) and on the recurrent connections (predominantly lower on the dendritic tree). Acetylcholine (ACh) can achieve this effect, exploiting the orderly arrangement of pyramidal cells dendrites [47]. Acetylcholine is one of several very ancient neuromodulating systems, well conserved across vertebrates, and it is likely that it operated in this way already in the early reptilian cortex, throughout its subdivisions. In recent years, Mike Hasselmo has been emphasizing this role of ACh in memory, with a combination of slice work and neural network modeling [48, 49]. This work has been focused on the hippocampus – originally, the medial wall – and on piriform cortex – originally, the lateral wall. In the hippocampus, however, it appears that mammals have devised a more refined trick to separate storage from retrieval, and perform both efficiently: operating the dentate gyrus preprocessor. It is illuminating, in fact, to contrast the avian and mammalian hippocampi. They are structurally very different, with birds having stayed close to their reptilian progenitors, and mammals having detached the dentate gyrus from Ammon's Horn, as mentioned above. Yet, at the behavioural level, the hippocampus of birds has been implicated in spatial memory in a role qualitatively similar to the prevailing description for the rodent hippocampus. Evidence comes from pigeons [18] and other species, and there is extensive literature to document it [26, 27].

Initially, the neural network approach, aiming at explaining structure from function, seemed to apply indiscriminately to hippocampal function in both birds and mammals, and therefore to be unable to say anything about the structural differences between the two. In his early paper, David Marr guessed the importance of recurrent collaterals, a prominent feature of the CA3 subfield [8], even though his own model was not really affected by the presence of such collaterals, as shown later [102]. Although the paper by Marr was nearly simultaneous with two of the most exciting experimental discoveries related to the hippocampus, that of place cells [75] and that of long term synaptic potentiation [19] for a long time it did not seem to inspire further theoretical analyses – with the exception of an interesting discussion of the collateral effect in a neural network model [44]. Marr himself become disillusioned with his youthful enthusiasm for unraveling brain circuits, and in his mature years took a much more sedate – and less neural – interest in vision. From 1987, however, McNaughton and Morris (1987) and then an increasing number of other investigators rediscovered the young Marr, and tried to elaborate those ideas in order to pin down the contribution of specific elements of the hippocampal circuitry. Edmund Rolls (1989) and several others have emphasized the crucial role probably played by the CA3 recurrent collaterals, that may form an autoassociator, a well studied network model of a content addressable memory. An autoassociator may subserve both the storage of episodic memories, e.g. in humans, and the storage of memory for space, e.g. in rats [15]). The emphasis on the essential role of the CA3 recurrent collaterals opened the way for attempting to understand the specialization of the dentate gyrus, in mammals [90]. A quantitative analysis of different network architectures (essentially, an autoassociator, CA3, operating with and without dentate gyrus to aid it in storing new memories) indicated an information theoretic advantage of one over the other in forming new representations. The models used were very abstract, and thus amenable to theoretical analysis [92] instead of just simulation, yet broadly consistent with generic cortical circuitry at all levels of details below the one being investigated. Conceptually, the function ascribed to the dentate is equivalent to the function ascribed to acetylcholine – to enhance the relation between hippocampal activity and afferent inputs during memory storage. The quantitative argument, however, allows a functional prediction at the neural level, which can be tested with suitable experiments. The prediction is that if mossy fibers are inactivated, the system is not able to acquire new hippocampal memories; or, more precisely, new memories rich in their information content. It should be able, nevertheless, to retrieve the memories already in store (and perhaps to form very impoverished representations of new memories). Somewhat surprisingly, the prediction has already been borne out of a purely behavioral experiment, in which mice were tested in a Morris water maze while transmission of dentate granule cell action potentials was reversibly blocked [61]. Another behavioural experiment has provided converging evidence [63]. Physiological experiments using similar techniques, in conjunction with measures of the information content of neural patterns of activity, will allow for more stringent tests of the argument. The mammalian 'invention' of the dentate gyrus, an ingenuity which likely took a long time to evolve from the simpler early reptilian organization, may thus represent a quantitative not qualitative improvement: qualitatively, we had acetylcholine already; but we managed to further improve on that.

5.2. CA1 in search of a role

If DG can be understood as a CA3 preprocessor, perhaps CA1 should be understood as a CA3 postprocessor. Yet studies based solely on the notion of the usefulness of a further associative memory and recoding stage after CA3 (Treves, 1995) failed to illustrate impressive advantages to adding such a stage. More interesting hints come from neuropsychological studies in rats [58], that indicate a more salient role for CA1 along the temporal dimension. CA3 may specialize in associating information that was experienced strictly at the same time, whereas CA1 may link together, more than CA3, information across adjacent times. A way to formulate a qualitative implication of such a functional differentiation is to state that CA1 is important for *prediction*, i.e. for producing an output representation of what happened just after, at the time of storage, whatever is represented by the pattern of activity retrieved at the CA3 stage. Note, however, that reading the Kesner review in full indicates that the table at the end is a well-meaning simplification. Their Fig.31.2 suggests that CA3 may be involved in temporal pattern separation just as much as CA1. Moreover, the role of either DG or CA3 in temporal pattern association has never really been assessed. Further, available studies on the role of CA1 fail to make a clear distinction between tasks in which massive hippocampal outputs to the cortex are crucial, and tasks in which a more limited hippocampal influence on the cortex may be sufficient. In the first case, lesioning CA1 should have an effect independently of what CA1 specifically contributes to information processing, simply because one is severing the main hippocampo-cortical output pathway. In the second, CA3 outputs through the fimbria/fornix could enable hippocampal-mediated influences to be felt, deprived, though, of the specific CA1 contribution.

Structurally, CA3 and CA1 are contiguous portions of the dorsomedial cortex. When this reorganizes into the mammalian hippocampus, CA3 and CA1 differentiate in two important ways. First, only CA3 receives the projections from the dentate gyrus, the mossy fibers. Second, only CA3 is dominated by recurrent collaterals, while most of the inputs to CA1 cells are the projections from CA3, the Schaffer collaterals (Amaral *et al*, 1990). In [96] the hypothesis was explored that the differentiation between CA3 and CA1 may help solve precisely the computational conflict between pattern completion, or integrating current sensory information on the basis of memory, and prediction, or moving from one pattern to the next in a stored sequence. Neural network simulations, based on the same sort of model as those analyzed in section 3 and reviewed in section 4, were used to assess to what extent CA3 would take care of the former, while CA1 would concentrate on the latter. With the simulations, at the price of some necessary simplification, one can compare the performance of the differentiated circuit with a non-differentiated circuit of equal number and type of components (one in which CA3 and CA1 have identical properties, e.g. both receive mossy fibers and are interconnected with recurrent collaterals). Lesion studies, instead, can only compare the normal circuit with others with missing components, and it is thus difficult for them to say the last word on the meaning of a differentiation. The hypothesis was not really supported by neural network simulations. The conflict indeed exists, but the crucial parameter that regulates it appears to be simply the degree of firing frequency adaptation in pyramidal cells. The differentiation between the architectures of CA3 and CA1 has a minor effect on temporal prediction, while it does significantly increase the information content of hippocampal outputs.

After those simulations were completed, new experimental results from the labs of Edvard Moser [66] and James Knierim [62] have shed a completely new light on the significance of the CA3-CA1 differentiation. As explained in forthcoming papers [64, 67], activity in CA3 and CA1 differs remarkably when rats are asked to navigate in environments that some cues suggest are the same, and others indicate they are different. CA3 appears to take an all-or-none decision, usually allocating nearly orthogonal neural representations to even very similar environments, and switching to essentially identical representations only above a high threshold of physical similarity. Activity in CA1, instead, varies smoothly to reflect the degree of similarity. This functional differentiation, and the finding that new representations in CA3 emerge slowly, presumably through iterative processing, are entirely consistent with the recurrent character of the CA3 network, and the prevailing feedforward character of the CA1 network. Thanks to these experimental findings, therefore, we are beginning to finally ‘understand’ CA1, and to make complete sense of the events that drastically altered the structure of our medial pallium nearly 200 million years ago.

6. Infinite recursion and the origin of cognition

A cornerstone of the search for cortical network mechanisms of cognition is the observation, old but often neglected, that each small patch of neocortex is internally wired up in the same basic way. This holds across areas as well as across mammalian species, with relatively minor differentiations and specializations that do not alter the neocortical microcircuitry scheme [30, 35, 81] nor the basic overall cortical plan [59, 105]. It holds also in areas implicated in language functions in humans. This suggests that the local network mechanisms subserving the rich variety of cognitive functions are always essentially the same, and functional differentiation corresponds solely, to a first approximation, to differences in the long-range connections of different cortical areas. The local ŞtransactionŤ, or elementary cortical network operation, is likely to be roughly the same everywhere [68], in sensory cortex as in association cortex.

Further, the long-range connections, denoted as the A system by Braitenberg [21] in contrast with the B system of local connections that do not leave the gray matter, follow indeed a specific wiring plan, which - when compared to simple mammalian species - is similar (although more complex) in elaborated species such as ours. However these connections do not seem to differ in other ways than in their overall wiring diagram: their layers of origin and termination, their synaptic mechanisms, their plasticity, their modulation by different neurotransmitters, all follow the same set of basic rules across cortical areas and across mammalian species.

One is led therefore to speculate that an understanding of the cortical operations underlying cognitive functions requires two main steps. First, the local network transaction has to be captured by a functional description, abstract enough to apply independently of areas and modalities yet accurate enough at the network level to be useful as a building block for system-level analyses. Second, global network operations have to be reduced to the combination of multiple instances of the universal local transaction, implemented along the wiring diagram relevant to each cognitive function.

6.1. Infinite recursion and its ambiguities

Are there clues about the nature of such global network operations that come from purely cognitive analyses? In a recent review, Marc Hauser, Noam Chomsky and Tecumseh Fitch [50] re-evaluate the requirements for the faculty of language. They state that language in the broad sense requires an adequate sensory-motor system and an adequate conceptual-intentional system, which however are both unlikely to be uniquely human attributes. They further propose that what may be uniquely human is a third necessary component of the faculty of language, that is a computational mechanism for recursion. Such a mechanism would provide the capacity to generate an infinite range of expressions from a finite set of elements. They also speculate that a capacity for infinite recursion may have evolved for reasons other than language, such as number processing, navigation, or social relations. In a related analysis, Daniele Amati (personal communication) wonders what could be a component that distinguishes uniquely human cognitive abilities, which he calls *H-abilities*, from the *1-H* abilities shared with other species; and he identifies this component with a capacity for producing arbitrarily long sequences that, at an abstract level, obey certain rules. This implies an ability to process cognitive states remote from those directly elicited by sensory inputs, and to generate such states recursively, *i.e.* a notion very close to the Chomskian one of infinite recursion in language, as manifested in a generative grammar. Thus a computational mechanism for infinite recursion may be utilized in other *H*-abilities, for example (as proposed by Amati) in the production of music (see [51], and other articles in the same issue).

Recursion, referred to the generation of infinite sequences of elements drawn from a finite alphabet, is an abstract and very loose notion. Computationally, at the most pedestrian level, it might simply mean that the transitions from one element to the next follow certain rules instead of being effectively random. A mathematical formulation of a grammar can be reduced in fact to the study of a system with certain forbidden transitions among its elements (see e.g. [72] and references therein). What are the elements, how they may be represented in the brain, and how restrictive are the rules they have to adhere to, remains to be clarified. Linguistically, and in other cognitive domains, recursion is often implied to mean something less pedestrian, like an embedding of clauses one inside the other in syntax, or the nesting of do-loops in Fortran codes. Recursion in this more sophisticated sense tends to be domain-specific, however, and is hardly ever infinite. An approach to explain infinite recursion mechanistically, in general, while taking into account domain-specific connotations is therefore likely to be ill-directed. More promising appears an almost opposite, nearly bottom-up approach, that considers the generic, pedestrian meaning of recursion, assumes the quality of its being infinite as critical, and focuses on the universal cortical transaction at the local network level.

6.2. Memory – statics and dynamics

In one of his ambitious and difficult papers discussing the organization of the brain, David Marr [69] proposed to regard the cerebral cortex in terms of its ability to decode the outside world using memory of previous experiences. In their 1991 book, Braitenberg and Schüz [22] summarized a series of insightful observations on the quantitative anatomy of the cortex, concluding that in general terms it operates as an associative memory machine. Over the last 15 years the interpretation of local cortical networks as attractor networks performing memory computations [9], which has informed our sect. 3, has diffused across the neuroscience community, leading to increased attention to the role of recurrent collateral processing even in early vision [80] and in slices [29, 86]. Memory computations take different forms [82], including self-organized recoding useful for categorization, pattern association (or directed memory in the terminology of Marr, [70]), and autoassociation or free memory. Common to all is the use of neurons as simple devices summating multiple inputs, of representations distributed over the activity of many neurons, and of associative plasticity mechanisms at their synaptic connections, as used in the earlier sections of this chapter.

With such minimal and neurally plausible ingredients, memory operations at the single neuron level can be depicted as simple analog operations on vectors of synaptic weights and on vectors of firing rates. These analog computations, widely accepted as the neural basis of memory in cortical networks, are seemingly far removed from the symbolic computations often subsumed as the logical basis of language and other higher cognitive faculties. Yet, apparent differences notwithstanding, analog computations at the single neuron level can implement symbolic computations at the local network level. The crucial element for this to occur is the discrete nature of local network attractors. The discreteness of local attractors can provide the error-correction capability and the robustness to noise that are often associated with the processing of discrete symbols.

In most simple models, local attractors are viewed as final states reached by a relaxation type of dynamics, that is, they coincide with the distribution of firing rates across neurons, that the local network would tend to reach in the absence of new perturbing inputs. This is however not necessarily the case. In his proposal of the notion of *synfire chains* [1], Moshe Abeles has envisioned inherently dynamical attractors, in which the identity of neurons firing in each attractor changes rapidly with time, along chains of links, each comprised of simultaneously firing neurons. The attractive nature of a chain expresses itself both in the convergence towards one of the sequences of links that are stored in memory, and in the progressive synchronization of the units comprising each link [17, 52]. Distinct sequences and distinct links within a sequence can share a number of participating units, and if this number does not exceed a value that can in principle be calculated, each chain continues to operate as a dynamical attractor. Further, distinct sequences can share the very same link or set of links, provided the activation of a link depends on previous links extending sufficiently into the past as to disambiguate each sequence. Although the original synfire chain model may be oversimplified, theoretically this notion has the merit of unleashing the computational capabilities of attractors, with their analog-to-symbolic transformation, into the temporal dimension. If provisions are made for the composition of individual chains, and for ŞswitchŤ links with multiple possible outcomes, synfire chains can implement the structure of transition probabilities of a grammar. It has indeed been noted how synfire chains, or objects of a similar nature, could be at the basis of language [77]; interestingly, the concept derives from the experimental observation of neural activity recorded in the frontal cortex of monkeys [3]. In contrast, the *Şstatic* notion of fixed-point attractors finds its most salient experimental inspiration in data recorded in the temporal lobe [71].

In the temporal cortex, if a local static attractor may be taken to correspond to a feature represented over a limited patch of cortex, a global attractor extending over many patches may be taken to correspond to an item from semantic memory [31, 39]. In the past we have analyzed quantitatively simple multi-modular associative memory networks, to check whether they could serve as models of semantic memory. In line with the distinction between the A and B systems of connections among pyramidal cells [21], we considered models in which each module, including N units, is densely connected through associatively modifiable weights (in fact each pair of units in the same module are pre- and post-synaptic to each other, so the number of local connections per unit C_B equals $N - 1$) while different modules are sparsely connected (each unit receives C_A connections, coming from other units widely distributed over M modules). Anatomical evidence suggests that C_B and C_A are large numbers of similar magnitude, *e.g.* of order 10^4 in primates [2]. N determines the number of local attractor states, denoted here as S, which analytical studies show scales up with the number of local connections per units, *i.e.* is proportional to C_B.

In a first study of a multi-modular network, we concluded that the number p of global attractor states cannot be much larger than S for the system to retrieve each memory item correctly. Analytical results show that if p is much larger than S, random combinations of local attractors, which do not correspond to any stored global pattern of activity, prevail as fixed points over the meaningful, stored combinations, which the C_A long-range connections per unit try to enforce [76]. Thus a simple-minded multimodular network could not serve as an effective semantic memory, since it would be limited to storing a very low number of items, of the same order as that of local attractor states. In a subsequent study, we identified two modifications to the first model we had considered, which increase its storage capacity beyond such a limited value [42]. The first modification is a long-range connectivity that is not uniformly sparse across modules, but is concentrated between a module and a subset of other modules that strongly interact with it. The second modification is to consider global activity patterns, or semantic memory items, that are not defined across all modules, but only over a subset, different for each pattern, which tends to include strongly interacting modules. With these combined modifications, the storage capacity, as measured by p, can increase well beyond the local capacity S, although its exact value is difficult to calculate, and depends on the details of the model. The tentative conclusion of the second study, therefore, was that a viable model of semantic memory based on a collection of interacting local associative networks should include (a) non-uniformly distributed long-range connections and (b) activity patterns distributed over a sparse fraction of the modules [42].

6.3. Memory latching as a model of recursive dynamics

The analyses above refer to the operation of semantic memory retrieval, which has to be initiated by an input that conveys a partial cue. In the temporal cortex, the so-called stimulus specific delay activity, which is observed for up to a few seconds following the offset of the stimulus, is typically weak and disrupted by successively intervening stimuli [23]. In the frontal cortex, similar delay activity can instead be quite strong and persist in the face of intervening stimuli [41], reflecting the overall weaker influence that sensory inputs to the cortex have on frontal networks, compared to that on networks of the temporal lobe (as modeled in [79]). It becomes pertinent to ask, then, especially in the case of frontal cortex, what type of dynamics may follow semantic memory retrieval: what happens to a network comprised of multiple associative modules, once it has been activated by a cue and it has retrieved a given semantic memory. In the following, it is proposed that what can happen depends critically on the number of semantic memories stored, that is, on the number of global attractor states. While allowing for a special contribution of the frontal cortex to temporal integration, due to its position in the overall cortical plan [43], and while broadly compatible with the declarative/procedural model of Ullman [100], the proposal focuses on a network mechanism that is not restricted to frontal cortex, but that in human frontal cortex may have found a novel expression because of a purely quantitative feature: the abundance of its connections.

The hypothesis requires one additional ingredient, which however in the cortex comes for free, so to speak. This is a passive mechanism for moving a local network out of an attracting state, after some time. A combination of firing rate adaptation in pyramidal cells, short-term depression at excitatory synapses and slow rebound inhibition would produce such an effect, and in different proportions would tend exclusively to inactivate the local network or also to favour its transition to a different attractor state, or even to enable flip-flop switching between pairs of states, as in binocular rivalry [60]. Globally, under certain conditions the collection of modules will move continuously from global attractor to global attractor or, more precisely, it will hop from state to state, given the discrete nature of the attractor states. It may rapidly pass through intermediate states, but in a well behaving semantic system mixture states are unstable (see [83] for a simplified model) and the trajectory, in the absence of new inputs, will essentially include periods close to attracting states, which would be fixed points except for the adaptation/inhibition mechanism, and rapid transitions between them. The system *latches* between attractors.

We now focus on whether such transitions will continue to occur, one after the other in the absence of inputs, and, if they occur, on the degree to which they follow rules, or are effectively random. When relatively few global attractors exist, in the high-dimensional space in which they live, the attractors will tend to be orthogonal, or approximately equally distant from each other. This is a statistical tendency that follows simply from the high dimensionality of the space, without special assumptions. In such a regime, transitions will be nearly random, if they occur at all. This is because as the system moves out of the previous global attractor none of the other attractors will be strongly engaged to take over; small fluctuations in the instantaneous condition of the system may favour a particular hopping among many essentially equiprobable ones, or else selective activity in the system may simply die out. When more global patterns exist, they populate more densely their high dimensional space, and at some point each pattern will have a subset of other patterns that are closer to it, or more similar, than the rest. In such a regime transitions between states will tend to be structured, and the dynamics will appear to follow certain rules, *i.e.* a grammar. The critical density of global attractor states at which structured transitions begin to prevail depends markedly on how patterns are generated, and one has to make more concrete assumptions in order to proceed with more quantitative arguments. It is not fully clear at this stage whether the transition between the two regimes takes the sudden character of a phase transition, akin perhaps to a percolation transition [46]. In general, however, it should remain valid that such critical density does not depend on the long-range connectivity. The storage capacity for semantic memory, instead, does depend on the connectivity. The hypothesis, then, is that a connectivity increase may increase the storage capacity of a frontal semantic multi-modular network, until it can store enough patterns that, when left without inputs, it can follow structured dynamics, which express a sort of transition rules. This hypothesis can be formulated in more detail by considering a concrete model, amenable to computer simulations.

Before discussing the toy model, it is tempting to freely speculate on the relation, within this framework, between the universal grammar, posited to underlie all human languages, and the grammar constraining each particular language, characterized by its choice of parameters (see e.g. [13]). The universal grammar should reflect the associative nature of the semantic network, largely embodied in a time-independent matrix of similarities between global attractors, but also endowed with the restricted extent of time arrows characteristic of any action semantics [106]. Such time arrows, or directed associations in Marr's terms, can be realized by simple and biologically plausible mechanisms, e.g. by spike-timing dependent synaptic plasticity. The same mechanisms can operate, when learning a specific language, to resolve the residual temporal order ambiguities left by the fact that action semantics does not specify all the temporal relations necessary to produce (one-dimensional) speech. Thus, in this interpretation, language parameters are set (arbitrarily, from a formal point of view, that is according to one's mother tongue) when funneling the more loosely time-constrained action semantics into the strict order of sequential discourse. Also parameters that seemingly do not reflect simple temporal order, like the polysynthesis parameter, might be indirect by-products of such a funneling effort.

7. Reducing local networks to Potts units

Consider again a network comprised of M modules each of which functions as an autoassociative network. Assume that each module stores S patterns and that, together with the intra-modular connections, there are also connections running between units in different modules. The full analysis of such a system, when including in addition non-uniform connectivity like the one discussed in sect. 2, would be very hard; in order to proceed one should thus consider some simplified model. The first natural choice is to consider a network with all to all connectivity inside modules and dilute connectivity between any two of them. This was the model investigated by O'Kane and Treves [76]. The critical factor in the revised model considered by Fulvi Mari and Treves [42] is the existence of what a *null state* as a new attractor that a module can reach in addition to all the stored patterns in it. This null state differs form the normal attractors in the sense that if a module goes to its null-state, it would have no effect on the other modules. Basically this null-state is something like the zero activity state for a single neuron, generalized to the network level. The technical problem associated with this model is that even though it appears to have a larger storage capacity compared to the network without null state, a full analysis of storage capacity cannot be done analytically. To circumvent this problem one can first make a further drastic simplification, and consider a new reduced model based on Potts neural networks [20, 57]. Then one essentially neglects the internal structure inside each module and represents the state of each module with by its correlation with the '0' (null) state or with one of the S attractor states. At its simplest, this can be just one discrete variable, taking one of $S+1$ values. Such a discrete variable simply indicates the closest stored pattern to the current state of the module. Then we model the interactions between two different modules, which in reality is the set of all weights associated to connections between them, with a $S(S+1)/2$-dimensional weight vector.

7.1. A discrete-valued model

At any time we associate a Potts variable s_i which takes one of the values $0 \dots S$ to the i^{th} module in the following way: s_i takes the value $q, q \neq 0$ if and only if pattern q is the closest pattern to the current activity of the network; and $s_i = 0$ if its closest pattern is the null state. Obviously being the closest one is nothing but having the largest overlap. Note that to facilitate comparisons with refs. [20, 57], one should convert to the notation Q (=$S+1$), for the number of Potts states. The interaction between modules i and j would be modeled through a set of weights $w_{ij}^{kl}, i, j = 1 \dots M; k, l = 0 \dots S$ symmetric in both $\{ij\}$ and $\{kl\}$. Now suppose that at time t the configuration of the system is $\{s_i\}$. Then at time $t+1$ we randomly choose one of the modules, say module i, and calculate a set of local fields $\{h_i^s\}, s = 0 \dots S$ defined as:

$$h_i^s = \sum_{j=1, j\neq i}^{M} \sum_{k,l=0}^{S} w_{ij}^{kl} u_{s_i,k} u_{s_j,l} \tag{7.1}$$

where $u_{k,l} = (S+1)\delta_{k,l} - 1$.

At time step $t + \Delta t$ the state variable s_i is set equal to the value s which maximizes h_i^s. The effect of Hebbian plasticity on the weights, which results in the formation of network attractors coinciding with, or near to the specified global patterns, can be described, for example, by the learning rule:

$$w_{ij}^{kl} = \frac{1}{(S+1)^2 M} \sum_{\mu=1}^{p} u_{\xi_i^\mu,k} u_{\xi_j^\mu,l} (1 - \delta_{k0})(1 - \delta_{l0}) \tag{7.2}$$

in which ξ_i^μ is the local attractor in module i which participates in the global pattern μ. It is drawn from a uniform probability distribution, *i.e.* all local attractors are assumed equally likely to participate in a global pattern. With this weight matrix, global patterns defined by $\{\xi_i^\mu\}$ (or network states very close to them) become the global attractors of the network, provided their number does not exceed a critical value (when M is large; in a small network the critical value is not well defined, as evident in the simulations below). Notice that we have considered the peculiar role of the null state in the dynamics of the network through the delta functions above. Also it should be noted that we have not yet considered whether the fraction of modules in the null state in each global memory pattern is the same or different as the fraction of modules in any other of the S 'genuine' local attractors.

7.2. Storage Capacity

In order to find the storage capacity of this network, we start by writing the Hamiltonian of the system. This is where we need the symmetry property of the weight matrix. If the weights are symmetric, as in 7.2, the dynamics of the network can be described by the following Hamiltonian:

$$H = -\frac{1}{2M} \sum_{i,j,j \neq i} \sum_{k,l} w_{ij}^{kl} u_{s_i,k} u_{s_j,l}. \tag{7.3}$$

One can then apply the classical methods of spin glasses to obtain the mean field equations of the system. The above formulation is basically nothing but a variation of the Potts-neural network first investigated by Kanter [57]. Kanter's model does not include the notion of the null state, and it treats all $S+1$ local states in the same way. It also assumes full connectivity, so that the number of units providing input to any given unit, C, equals $M-1$. For such a network Kanter found that the storage capacity for small values of S scales like $MS(S+1)$. As noted by Kanter, this critical storage load scales up with the number S of Potts states *squared* because, effectively, a connection weight between a pair of Potts units is comprised of $S(S+1)/2$ independently tunable synaptic variables. When the network is loaded close to its memory capacity each such variable ends up storing up to a fraction of a bit, as in the Hopfield model [57]. This result, it turns out, is valid only when S is small, and cannot be generalized to large values of S, which is the case of interest for us. In the large S limit we found that the critical load scales like $MS(S+1)/\log(kS)$. The numerical factor k is in practice quite large (of order 10^6), and the correction term $\log S$ becomes important only for S very large.

To apply a Potts model to our multimodular semantic network one needs to consider a number of extensions of the Kanter model. The first is incomplete connectivity between the Potts units. As for the analog extensions of the Hopfield model [82] the formula for the storage capacity is modified in that the number C of connections each unit receives replaces M, the number of modules, and the numerical prefactor becomes larger (due to less reverberation of the noise along closed loops).

7.3. Sparse coding

In the above formulation, all local patterns have the same probability of appearing in a given global pattern. We are particularly interested, instead, in the case where this probability is much higher for the null state than for the others. In other words the fraction a of modules in genuine local attractors (those different from the null state) should be small. This is equivalent to the notion of *sparse coding* in autoassociative memories. Adding the additional '0' state is in fact analog to considering 0-1 spin extensions of the Hopfield model with sparse coding [24, 98]. As in the associative networks with sparsely coded patterns, one expects that using sparse coding, the modular network will have a larger storage capacity. For sparsely coded global patterns we can rewrite the definition of the weights as:

$$w_{ij}^{kl} = \frac{1}{(S+1)^2 M} \sum_{\mu=1}^{p} (u_{\xi_i^\mu,k} - B_k)(u_{\xi_j^\mu,l} - B_l)(1-\delta_{k0})(1-\delta_{l0}) \tag{7.4}$$

where the $\{B_k\}$'s, following [20], are defined through the equality:

$$Pr\{\xi_i^\mu = k\} = \frac{1+B_k}{S+1}. \tag{7.5}$$

Bolle *et al* [20], while not aiming to consider a null state, studied a generic Potts neural network with biased patterns, *i.e.* with non-zero $\{B_k\}$, although without considering optimal threshold setting, a bit like in [10]. Their formalism can be slightly modified and utilized to study a sparsely coded Potts neural network, with a null state. Optimal threshold-setting amounts, as in the transition from [10] to [98], to removing the constant coupling among non-null states, i.e. adding a term

$$\Delta w_{ij}^{kl} = \frac{-p}{(S+1)^2 M}(1+B_k)(1+B_l)(1-\delta_{k0})(1-\delta_{l0}) \tag{7.6}$$

This is the form of the couplings used in the simulations reported below. A full analytical treatment is still to be carried out, but based on signal to noise analyses and computer simulations we expect a scaling behavior like $p_c \simeq CS^2/a \log(S/a)$ for large S and small a. That is, the storage capacity benefits from sparser codings, unlike what happens without optimal threshold-setting.

7.4. A Potts model with graded response

In more realistic models of semantic storage, the stabilization into local attractors cannot be assumed to be an all-or-none phenomenon, and global attractor states cannot be assumed to be independent of one another and spatially uncorrelated. To deal with the first aspect, in the simulations we abandon the discrete Potts units used in the original storage capacity calculations, in favour of graded, analog variables representing the degree of overlap of local activity with a local attractor, and summing up to one:

$$\sum_{k=0}^{S} s_{i,k} = 1 \tag{7.7}$$

which reflect input variables $\sigma_{i,k}$ according to a standard sigmoidal activation transform:

$$s_{i,k} = \frac{\exp \beta\sigma_{i,k}}{\sum_{l=0}^{S} \beta\sigma_{i,l}} \tag{7.8}$$

where β has the role of an inverse temperature, and the $\sigma_{i,k}$'s could simply be taken to reflect the weighted summation of inputs from other modules. To model somewhat more accurately the dynamics of entering and leaving a local attractor, however, it is convenient to assume the $\sigma_{i,k}$'s to integrate another set of variables, which themselves reflect summed inputs:

$$\tau_1 \dot{\sigma}_{i,k} = -\sigma_{i,k} + h_{i,k} - h_{i,k}^T \tag{7.9}$$

for $k \geq 1$, with the local fields $h_{i,k} = \sum_{j,l} w_{ij}^{kl} s_{j,l}$. Note the difference with the discrete-valued model in Eq. 7.1. The attractor-specific thresholds $h_{i,k}^T$'s evolve with a slower time constant to track recent correlation with the corresponding local attractor:

$$\tau_2 \dot{h}_{i,k}^T = s_{i,k} - h_{i,k}^T. \tag{7.10}$$

For $k = 0$, the 'activation' variable $\sigma_{i,0}$ acts as a general threshold for all local attractors, modulated on an even slower time scale, τ_3, by the extent to which activity in the network is correlated to local attractors, as opposed to being in the null state:

$$\begin{aligned} \sigma_{i,0} &= r_0^T - h_{i,0} \\ \tau_3 \dot{h}_{i,0} &= \sum_{k=1}^{S} s_{i,k} - h_{i,0}. \end{aligned} \tag{7.11}$$

In the simulations below, the inverse temperature β and the fixed threshold baseline r_0^T were given values estimated to favour near optimal retrieval behaviour, while the time constants τ_1, τ_2 and τ_3 were given values of *e.g.* 10, 33 and 100 basic integration time steps (a time step was indicatively taken to correspond to 1 msec of real neuronal dynamics. With such differential equations the graded variables describing local network behaviour evolve in time similarly to the collective variables describing an autoassociator network of integrate-and-fire units with adaptation [15].

7.5. Correlated patterns

Correlations among patterns can drastically reduce the storage capacity of an autoassociative network. However we hypothesize that in some models with correlations, one of which is adopted in the simulations sketched below, storage capacity is indeed reduced, but essentially by a prefactor dependent on the correlations, preserving the general dependence of p_c on the connectivity per unit, C –Ű a linear dependence; and on the number S of local attractors – roughly, a quadratic dependence.

Memory retrieval was simulated in a network of Potts units, in which global activity patterns to be stored as memory items were generated by a two-step algorithm, that could be parametrically varied from producing independent to highly correlated patterns. In the first step, a number of underlying *factors* were generated, defined simply as distinct random subsets of the entire set of Potts units. In the simulations, each subset included 50 units out of the total 300 units, and a total of 200 such factors were generated. The overlaps in the spatial distribution of different factors therefore are purely random, and clustered around their mean value $50/300 = 8.33$.

In the second step, global patterns were generated from the factors, which had been indexed by r in order of decreasing mean importance. For each global pattern, the specific importance of each factor was given by a coefficient γ_r^μ obtained by multiplying the overall factor $\exp(\zeta r)$ by a random number, taken to be 0 with probability $1 - a$, and otherwise drawn with a flat distribution between 0 and 1, specifically for pattern μ. A value taken by factor r, s_r, was randomly drawn among the S 'genuine' attractors, and a contribution γ_r^μ was added to the field onto each Potts unit over which factor r was defined, in the direction s_r. After accumulating contributions from all factors, the direction in which each unit received the largest field was computed, and the aM units receiving the largest maximal fields were assigned the corresponding direction s_r in pattern μ, while the remaining $(1 - a)M$ units were assigned the null state in pattern μ.

With this procedure, pairs of Potts units have uncorrelated activity when averaged across patterns (because the different patterns that both engage the pair will span nearly evenly the different local states). Pairs of patterns, instead, can by highly correlated once averaged across units, particularly if they share one or a few most important factors; and positively correlated if these factors have been assigned the same direction in Potts space. Thus correlations among patterns will be higher if the importance of different factors decreases rapidly (e.g., in the simulations the value $\zeta = 0.02$ was used, equivalent to assuming of order 50 'important' factors); and they will tend to vanish if all factors are equally important, in general ($\zeta = 0$). When correlations are very high each pattern tends to be significantly correlated with a specific subset of the others, those sharing the main factor that influences them, and positively correlated with a fraction $1/S$ of this subset. In this scheme, the number of memory items significantly overlapping with one recently retrieved, and which can be the target of a non-random transition, scales up as p/S, and does not depend on C. By contrast, the storage capacity for retrieval, although severely reduced by correlations, should still scale as $p_c \simeq CS^2/a$. This leads to the two diagrams in Fig. 6, which indicate that conjoint semantic retrieval and structured transitions should be possible only above critical values for C and S. Translated into the language of an underlying multimodular network, the expectation is that there should be critical values for both the short and the long-range connectivity, C_A and C_B, beyond which a model which follows this factorial scheme would be able of both semantic retrieval and infinite recursion.

Before discussing the simulations of the Potts model, it is useful to clarify how its connectivity parameters could be mapped onto those of an underlying multimodular network model. In the *reduced* Potts model, each unit receives C connections from other units, for a total of $CS(S-1)/2$ independently variable weights per unit. A storage capacity of $p_c \simeq CS^2/[a\log(kS)]$ patterns, each of which contains about $Ma\log_2(S)$ bits of information, implies that the total information that can be stored in the reduced network is of order $I_{tot} \simeq MCS^2$, that is, of order one bit per synaptic variable. In the full multimodular network, including N units per module, each unit would receive C_A single-variable weights from units in other modules. Note that one can further take S, the number of local attractors to be of order the number C_B of short-range (local) connections per unit in the underlying model, that is of order N. If also the full network, like the reduced network, can store of order one bit per synaptic variable, in this case it would amount, even counting only long range connections, to $I_{tot} \simeq MNC_A \simeq MC_BC_A$. This implies that the bound on the number of global patterns, or semantic items, should scale up as $p_c \simeq C_AC_B/[a\log(kS)]$, that is not only it should increase with sparser modular coding (the a factor), but it should also scale up with the *product* of the number of long- and short-range connections per unit in the underlying model, not with their *sum*. This is a possibility left open in the Fulvi Mari & Treves [42] calculation, which should be verified by further analysis and simulation. From a quantitative point of view, it would resurrect the idea [21, 22] that multimodular cortical networks can serve as efficient semantic memory storage devices, raising their capacity from several thousands to several millions of items.

7.6. Scheme of the simulations

Whereas simulating the full multimodular network is a long-term project, the reduced Potts model requires only manageable CPU times and memory loads and can easily be simulated on a standard PC. Figure 7 shows a sample of the types of network dynamics which emerge in the simulation of the reduced Potts model.

When adaptation is turned off, typically the network remains in the retrieved attractor indefinitely. When adaptation is on, it gradually decreases the overlap between current network activity and the retrieved attractor. During this decay phase, other attractors see their overlaps increase. If one of them becomes sufficiently strong to pass an effective threshold (around 0.5 in the simulations), it manages to attract the entire network, and rapidly it reaches values close to 1, before decaying away in its turn. This transition can be repeated several times (bottom panel), reminiscent of the series of transitions seen in monkey frontal cortex [4]. The crucial ingredient for an indefinite repetition, and thus for infinite recursion to occur, is that any activated global pattern must have at least one *neighbour* that can reach an overlap above threshold before its predecessor has decayed away. Although this is a dynamical phenomenon, it is closely related to the (static) matrix of similarities among stored patterns. The more significantly correlated global patterns exist to the one currently activated, the more likely is *latching* to proceed. For it to proceed indefinitely, each of the patterns activated in sequence must be able to activate the next, and this is more likely to occur when the density of patterns is higher, as posited in Fig. 6. To check more quantitatively the expectation expressed in those diagrams, we have run extensive simulations in which I have varied systematically C, S and the storage load p, and kept other parameters constant.

Fig. 8 summarizes how these 3 parameters determine the network ability to combine the retrieval of the first, cued pattern with successive latching to different patterns. The light areas correspond to regions where both retrieval and latching occur frequently (averaging across thousands of independent runs). In the dark areas either retrieval tends to fail (towards the top of both plots) or latching tends to die out (towards the bottom of both plots). The simulations clearly demonstrate the existence of a limit p_c on the storage load, beyond which retrieval of the pattern that best matches a partial cue is not possible (the striped regions of Fig. 6, and the top portions of Fig. 8). Below this limit (e.g. at the marked points on the $p-C$ and $p-S$ planes in Fig. 6) cued retrieval does occur, and latching can occur as well if p is high enough, but still below p_c, hence inside the right *Śwedges* appearing in both plots of Fig. 6 and Fig. 8. Note that Fig. 8, represents the results of simulations limited in the size of the system but also in the time of each run (30 time steps), and this contributes to its smoother, graded appearance than Fig. 6. It is expected, though, that comparing over longer runs behaviour corresponding to the marked points in Fig. 6, long series of structured transitions will prevail only in the higher p regime (the upper marked point in each panel of Fig. 6), possibly above a *percolation* critical point. This could be assessed quantitatively even looking only at a limited time window, by measuring the entropy of states that follow the activation of each attractor. For structured transitions, this entropy is smaller than for random transitions, thereby quantifying the metric content [93] of the underlying grammar. We are currently working at the fully analytic approach sketched above, will should allow clarification of these issues, beyond the limits of computer simulations and their dependence on specific choices of parameters.

7.7. Conclusions

The proposal [97] is that a generic capacity for infinite recursion (intended in its basic meaning) may have evolved as a consequence of the refinement of the semantic system. Such a refinement may have been triggered by the increase in connectivity among pyramidal cells in the cortex, particularly for some mammalian lineages including primates, and particularly in the temporal and frontal lobes [36, 37]. Such a development may have then been accelerated in the frontal cortex, relative to the temporal lobe and its sensory semantics, because action semantics invoked more structure along the time dimensions. This may have led to a capacity for syntax in communication in humans, favored by the further connectivity increase in their frontal cortex.

This proposal is still vague in several details, and it requires an analytical approach to be validated at least at the level of the self consistency of the mathematical model, even before implications for the evolution of cognition are explored in full. Its relation to a number of related approaches are discussed in [97], while here we note the distinction from the concept of phase transitions explored, in relation to language dynamics, in [74], and the potential relation to studies of the chaotic behaviour of analog systems which are close to neural networks [5]. Further work on Potts neural networks and their analog versions may pave the way to a better understanding of the rich dynamics of multi-modular neuronal networks, and indirectly contribute to illuminate the mysteries surrounding the sudden appearance, perhaps 40,000 years ago, of qualitatively new cognitive capabilities in our species.

Acknowledgments

This chapter is based on three lectures given by AT at the Les Houches 2003 summer school, integrated by reports on analytical work in progress, written by YR. We are most grateful to Edvard and May-Britt Moser, Stefan Leutgeb, Jim Knierim and coworkers for making us appreciate the hippocampal experimental findings from their labs before publication. We are thankful to Baktash Babadi and Pouya Pakarian for the nice discussions and the hospitality. AT enjoyed the opportunity of giving lectures on these topics at the IPM, School of Cognitive

Sciences in Tehran and also at Bar Ilan, Brain Research Center in Tel Aviv-Ramat Gan.

References

[1] Abeles M (1982) Local Cortical Circuits. (Springer, Newyork)

[2] Abeles M (1991) Corticonics: Neural Circuits of the Cerebral Cortex (Cambridge Univ. Press, Cambridge)

[3] Abeles M et al (1993) Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. *J Neurophysiol* 70:1629-1638

[4] Abeles M et al (1995) Cortical activity flips among quasi-stationary states. *Proc Natl Aca Sci* 92:8616-8620

[5] Afraimovich VS et al (2004) Heteroclinic contours in neural ensembles and the winnerless competition principle. *Int J Bifurc Chaos* 14:in press.

[6] Allman J (1990) Evolution of neocortex. In *Cerebral Cortex*, vol 8A *Comparative Structure and Evolution of Cerebral Cortex* Jones, E.G. & Peters, A., eds. (Plenum Press, New York), 269-283

[7] Amit DJ (1989) Modelling Brain Function (Cambridge Univ. Press, New York).

[8] Amaral DG et al (1990) Neurons, numbers and the hippocampal network. *Progress in Brain Research* 83:1-11

[9] Amit DJ (1995) The Hebbian paadigm reintegrated: local reverberations as internal representations. *Behavioral and Brain Sciences* 18:617-657

[10] Amit DJ et al (1987) Information storage in neural networks at low levels of activity. *Phys Rev* A 35:2293

[11] Amit DJ, Tsodyks MV (1991) Quantitative study of attractor neural network retrieveing at low spike rates: I. Substrates – spikes, rates and neuronal gain. *Network: Comput Neural Syst* 2:259

[12] Amit DJ & Brunel N (1997) Dynamics of a recurrent network of spiking neurons before and following learning. *Network: Comput Neural Syst* 1:381

[13] Baker MC (2002) The Atoms of Language. (Oxford University Press, New York)

[14] Batardiere A et al (1998) Area-specific laminar distribution of cortical feedback neurons projecting to cat area 17: Quantitative analysis in the adult and during ontogeny. *J Comp Neurol* 396:493-510

[15] Battaglia FP & Treves A (1998a) Stable and rapid recurrent processing in realistic autoassociative memories. *Neural Computation* 10: 431-450

[16] Battaglia FP & Treves A (1998b) Attractor neural networks storing multiple space representations: a model for hippocampal place fields. *Physical Review E* 58:7738-7753

[17] Bienenstock E (1995) A model of neocortex. *Network: Comput. Neural Syst.* 6:179-224.

[18] Bingman VP & Jones T-J (1994) Sun-compass based spatial learning impaired in homing pigeons with hippocampal lesions *Journal of Neuroscience* 14:6687-6694

[19] Bliss TV & Lomo T (1973) Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology* 232:331-356

[20] Bolle D et al (1993) Mean-field theory for the Q-state Potts-glass neural network with biased patterns. *J. Phys. A: Math. Gen.* 26:549

[21] Braitenberg V (1978) Cortical architectonics: general and areal. In: Brazier MAB & Petsche H (eds) Architectonics of the cerebral cortex. (Raven, New York)

[22] Braitenberg V & Schuz A (1991) Anatomy of the Cortex (Springer-Verlag, Berlin).

[23] Brunel N (2003) Dynamics and plasticity of stimulus-selective persistent activity in cortical network models. *Cereb. Cortex* 13:1151-1161

[24] Buhmann J et al (1989) Associative memory with high information content.*Phys. Rev. A* 39:2689-2692.

[25] Carroll RL (1988) Vertebrate Paleontology and Evolution (W H Freeman & Co., New York).

[26] Clayton N & Krebs JR (1995) Memory in food-storing birds: from behaviour to brain. *Current Opinion in Neurobiology* 5:149-154

[27] Clayton NS, Griffiths DP, Emery NJ & Dickinson A (2001) Elements of episodic-like memory in animals. *Philosophical Transactions of the Royal Society of London* B 356:1483-1491

[28] Collingridge GL & Bliss TV (1995) Memories of NMDA receptors and LTP. *Trends in Neuroscience* 18:54-56

[29] Cossart R et al (2003) Attractor dynamics of network UP sates in the neocortex.*Nature* 423:283-288.

[30] DeFelipe J et al (2002) Microstructure of the neocortex: comparative aspects. *J Neurocytol* 31:299-316

[31] Devlin J et al (1998) Category specific semantic deficits in focal and widespread brain damage. A computational account. *J Cogn Neurosci* 10:77-94

[32] Diamond IT & Hall WC (1969) Evolution of neocortex. *Science* 164:251-262

[33] Diamond IT et al (1985) Laminar organization of geniculocortical projections in Galago senegalensis and Aotus trivirgatus. *J Comp Neurol* 242:610

[34] Donoghue JP et al (1979) Evidence for two organizational plans in the somatic sensory-motor cortex in the rat. *J Comp Neurol* 183:647-666

[35] Douglas RJ & Martin KAC (1991) A functional microcircuit for cat visual cortex. *J. Physiol.* 440:735-769

[36] Elston GN (2000) Pyramidal cells of the frontal lobe: all the more spinous to think with. *J Neurosci* 20:RC95(1-4)

[37] Elston GN et al (2001) The pyramidal cell in cognition: a comparative study in human and monkey. *J Neurosci* 21:RC163(1-5)

[38] Erickson RP et al (1967) Organization of the posterior dorsal thalamus of the hedgehog.*J Comp Neurol* 131:103-130

[39] Farah M & McClelland J (1991) A computational model of semantic memory impairment: modality specificity and emergent category specificity. *J Exp Psychol: Gen* 120:339-357

[40] Finlay BL & Darlington RB (1995) Linked regularities in the development and evolution of mammalian brains. *Science* 268:1578-1584

[41] Freedman DJ et al (2003) A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J Neurosci* 23:5235-5246

[42] Fulvi Mari C & Treves A (1998) modeling neocortical areas with a modular neural network. *Biosystems* 48:47-55

[43] Fuster JM (2002) Frontal lobe and cognitive development. *J Neurocytol* 31:373-385

[44] Gardner-Medwin AR (1976) The recall of events through the learning of associations between their parts. *Proceedings of the Royal Society of London* B 194:375-402

[45] Haberly LB (1990) Comparative aspects of olfactory cortex. In Cerebral Cortex, vol. 8B: Comparative Structure and Evolution of Cerebral Cortex. (Jones EG, Peters A, eds.), pp.137-166. New York: Plenum Press

[46] Hammersley JM (1983) Origins of percolation theory. *Ann Israel Phys Soc* 5:47-57

[47] Hasselmo ME & Schnell E (1994) Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: Computational modeling and brain slice physiology. *Journal of Neuroscience* 14:3898-3914

[48] Hasselmo M et al (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* 15:5249-5262

[49] Hasselmo M et al (1996) Encoding and retrieval of episodic memories: role of cholinergic and GABAergic modulation in hippocampus. *Hippocampus* 6:693-708

[50] Hauser MD et al (2002) The faculty of language: what is it, who has it, and how did it evolve? *Science* 298:1569-1579

[51] Hauser MD & McDermott J (2003) The evolution of the music faculty: a comparative perspective. *Nature Neurosci* 6:663-668

[52] Hertz J & Prugel-Bennet A (1996) Learning synfire chains: turning noise into signal. *Int J Neural Syst* 7:445-450

[53] Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Aca Sci USA* 79:2554-2558

[54] Jerison HJ (1990) In Cerebral Cortex, vol. 8A: Comparative Structure and Evolution of Cerebral Cortex, eds.Jones, EG & Peters, A (Plenum Press, New York), pp.285-309

[55] Jones EG (1998) Viewpoint: the core and matrix of thalamic organization. *Neuroscience* 85:331-45

[56] Kaas JH (1982) In: Contributions to sensory physiology, vol. 7 (Academic Press, New York) pp 201-240

[57] Kanter I (1988) Potts-glass models of neural networks, Phys. Rev. A. 37:2739

[58] Kesner RP et al (2002) Subregional analysis of hippocampal function in the rat. In *Neuropsychology of Memory*, LR Squire & DL Schacter (Eds.), 3rd Ed. (Guilford Press)

[59] Krubitzer L (1995) The organization of neocortex in mammals: are species differences really so different? *Trends Neurosci.* 18:408-417

[60] Laing CR & Chow, CC (2002) A spiking neuron model for binocular rivalry. *J. Comput. Neurosci.* 12:39-53

[61] Lassalle JM et al (2000) Reversible inactivation of the hippocampal mossy fiber synapses in mice impairs spatial learning, but neither consolidation nor memory retrieval, in the Morris navigation task *Neurobiol. Lear. Mem.* 73:243-257

[62] Lee I et al (2003) Differential coherence of CA1 vs CA3 place field ensembles in cue-conflict environments. *Soc Neurosci abs* 29:91.11

[63] Lee I & Kesner RP (2004) Encoding versus retrieval of spatial memory: double dissociation between the dentate gyrus and the perforant path inputs into CA3 in the dorsal hippocampus. *Hippocampus* 14:66-76.

[64] Lee I, Yoganarasimha D, Rao G & Knierim JJ (2004) Autoassociative network properties of the ensemble representation of environments in the CA3 field of the hippocampus. Submitted

[65] Lende RA (1963) Cerebral cortex: a sensorimotor amalgam in the Marsupialia. *Science* 141:730-732

[66] Leutgeb S et al (2003) Differential representation of context in hippocampal areas CA3 and CA1. *Soc Neurosci abs* 29:91.5

[67] Leutgeb S, Leutgeb JK, Treves A, Moser M-B & Moser EI (2004) Distinct ensemble codes in hippocampal areas CA3 and CA1. Submitted

[68] Lorente de Nó R (1938) Architectonics and structure of the cerebral cortex. In Physiology of the Nervous System (Fulton JF, ed) pp. 291-330. (Oxford University Press, New York)

[69] Marr D (1970) A theory for cerebral neocortex. *Proc Roy Soc Lond* B 176:161-234

[70] Marr D (1971) Simple memory: a theory for archicortex. *Phil Trans Roy Soc (London)* B 262:23-81

[71] Miyashita Y & Chang HS (1988) Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* 331:68-70

[72] Namikawa J & Hashimoto T (2003) Dynamics and computation in functional shifts. Los Alamos arXiv:nlin.CD/0302048

[73] Nicoll A & Blakemore C (1993) Patterns of local connectivity in the neocortex. *Neural Comput* 5:665-68

[74] Nowak MA et al (2002) Computational and evolutionary aspects of language. *Nature* 417:611-617

[75] O'Keefe J & Dostrovsky J (1971) The hippocampus as a spatial map: preliminary evidence from unit activity in the freely moving rat. *Brain Research* 34:171-175

[76] O'Kane D & Treves A (1992) Why the simplest notion of neocortex as an autoassociative memory would not work. *Network: Comp Neural Syst* 3:379-384

[77] Pulvermuller F (2002) A brain perspective on language mechanisms: from discrete neuronal ensembles to serial order. *Progr Neurobiol* 67:85-111

[78] Rauschecker JP et al (1995) Processing of complex sounds in the macaque nonprimary auditory cortex. *Science* 268:111-114

[79] Renart A et al (1999) Associative memory properties of multiple cortical modules. *Network: Comp Neural Syst* 10:237-255

[80] Ringach DL et al (2003) Dynamics of orientation tuning in macaque V1: the role of global and tuned suppression. *J. Neurophysiol.* 90:342-352

[81] Rockel AJ et al (1980) The basic uniformity in structure of the neocortex. *Brain* 103:221-24

[82] Rolls ET & Treves A (1998) Neural Networks and Brain, (Oxford University Press: Oxford)

[83] Roudi Y & Treves A (2003) Disappearance of spurious states in analog associative memories. *Phys Rev E* 67:041906

[84] Roudi Y & Treves A (2004) An associative network with spatially organized connectivity. Submitted.

[85] Shiino M & Fukai T (1993) Self-consistent signal-to-noise analysis of the statistical behavior of analog neural networks and enhancement of the stoarge capacity. *Phys. Rev. E* 48:867

[86] Shu Y et al (2003) Turning on and off recurrent balanced cortical activity. *Nature* 423:288-293

[87] Treves A (1990) Graded-response neurons and information encoding. *Phys Rev A* 42:2418

[88] Treves A & Rolls ET (1991) *Network: Comp. Neural. Syst.* 2:371

[89] Treves A (1991) Dilution and sparse coding in theshold-linear nets. *J Phys A: Math Gen* 24:327

[90] Treves A & Rolls ET (1992) Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus* 2:189-199

[91] Treves A (1995) Quantitative estimate of the information relayed by the Schaffer collaterals. *J Comput Neurosci* 2:259-272

[92] Treves A et al (1996) How much of the hippocampus can be explained by functional constraints? *Hippocampus* 6:666-674

[93] Treves A (1997) On the perceptual structure of face space. *Biosystems* 40:189-196

[94] Treves A (2001) In Handbook of Biological Physics, vol. 4: Neuro-Informatics and Neural Modelling, eds Moss F & Gielen S (Elsevier, Amsterdam) pp. 825-852

[95] Treves A (2003) Computational constraints that may have favoured the lamination of sensory cortex, *J Comput Neurosci* 14:271-282

[96] Treves A (2004a) Computational constraints between retrieving the past and predicting the future, and the CA3-CA1 differentiation. *Hippocampus* 14:on-line early view.

[97] Treves A (2004b) Frontal latching networks: a possible neural basis for infinite recursion. *Cogn Neuropsy*: in press

[98] Tsodyks MV & Feigel'man MV (1988) The enhanced storage capacity in neural networks with low activity level. *Europhysics Lett* 6:101-105

[99] Ulinski PS (1990) The cerebral cortex of reptiles, In Cerebral Cortex, vol. 8A: Comparative Structure and Evolution of Cerebral Cortex, eds EG Jones & A Peters (Plenum Press, New York) pp 139-215

[100] Ullman MT (2001) A neurocognitive perspective on language: the declarative/procedural model. *Nat Rev Neurosci* 2:717-726

[101] Whitfield IC (1979) The object of the sensory cortex. *Brain Behav Evol* 16:129-154

[102] Willshaw D & Buckingham J (1990) An assessment of Marr's theory of the hippocampus as a temporary memory store. *Philosophical Transaction of the Royal Society of London* B 329:205-215

[103] Wilson EO (1975) Sociobiology. The New Synthesis (Harvard Univ. Press, Cambridge, MA)

[104] Yoshioka T et al (1992) Intrinsic lattice connections of macaque monkey visual cortical area V4. *J. Neurosci.* 12:2785-2802

[105] Young MP et al(1994) Analysis of connectivity: neural systems in the cerebral cortex. *Rev Neurosci* 5:227-250

[106] Zanini S et al (2002) Action sequencing deficit following frontal lobe lesions. *Neurocase* 8:88-99

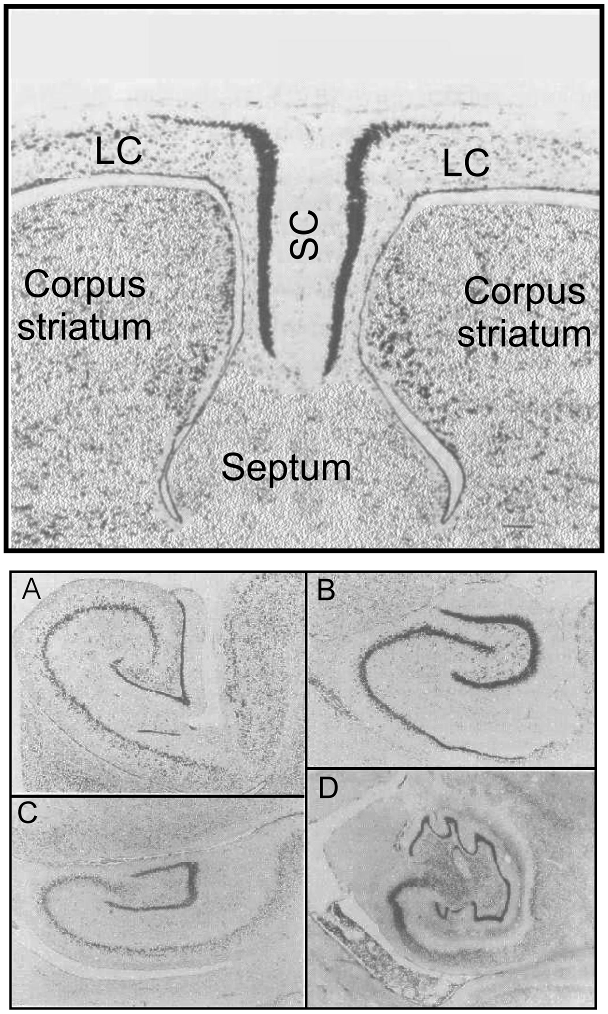

Fig. 1. The structural phase transition in the hippocampus. The medial pallium of a reptile (a lizard), top, with indicated the Large Cell (LC) and Small Cell (SC) subdivisions. Examples of the reorganized medial pallium in 4 highly divergent mammalian species, bottom: A – opossum; B – rat; C – cat; D – human. The homolog of the SC subdivision has become the detached dentate gyrus, which sends connections to the CA3 portion of the homolog of the LC subdivision, that has remained continuous with the rest of the cortex.

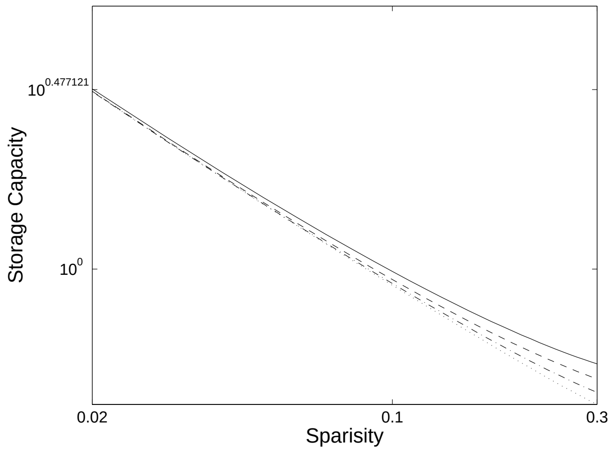

Fig. 2. Storage Capacity vs. a for $C/N = 0$ (full curve), $C/N = 0.05$ (dashed line) and $C/N = 1$ (dotted line).

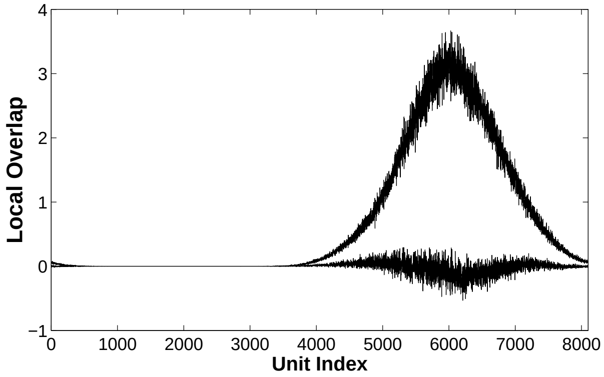

Fig. 3. The result of simulating a network of N=8100 units on a 1D ring, with C=405, p=5 and σ=300. The big bump is the local overlap with the retrieved pattern, and the small fluctuating curve is the overlap with one of the non retrieved patterns. Periodic boundary conditions were used.

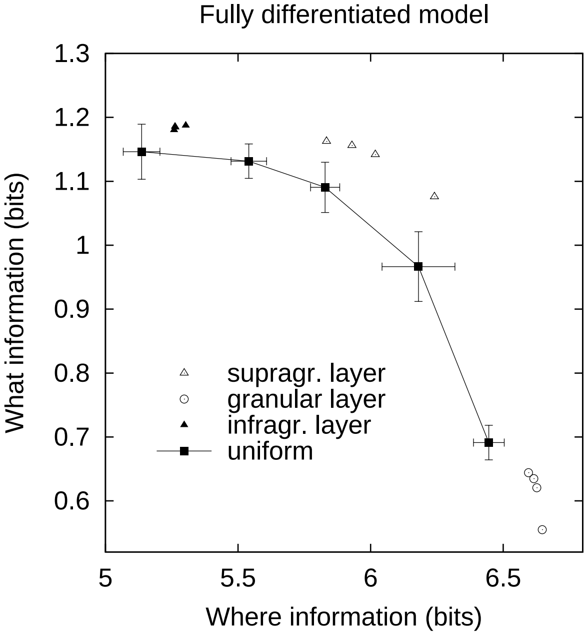

Fig. 4. I_i and I_p values obtained, after 3 training epochs, with the uniform model and, for 4 different parameter choices, for the fully differentiated model. 3 of the data points for the infragranular layer (black triangles) are nearly superimposed.

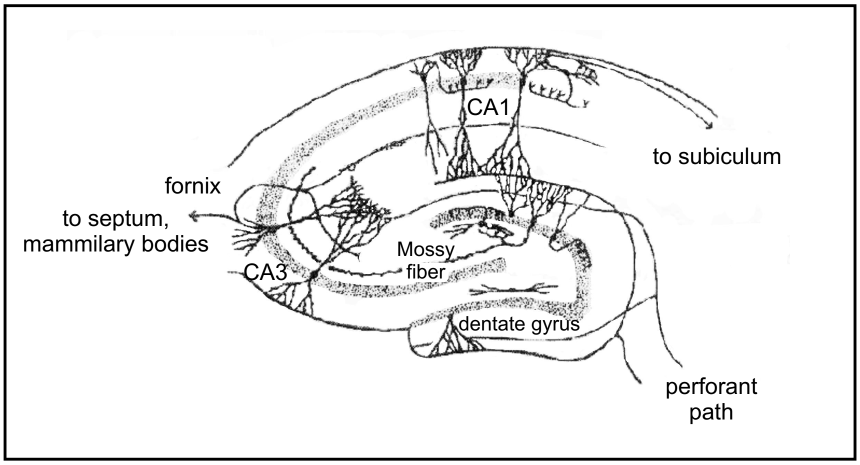

Fig. 5. Scheme of some of the main subfields and synaptic systems of the hippocampus proper. Redrawn from [82]

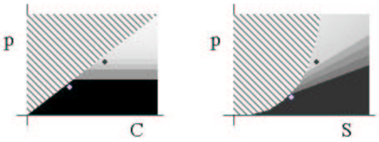

Fig. 6. Useful ranges for the number of global attractors. In the striped area above the critical line, which is linear in the C (left) and almost quadratic in S (right) semantic retrieval is not possible, because p is above the maximum storage load. Below the critical line, there is expected to be a (dark) region of low p values where long sequences of structured transitions are not possible. This region extends up to p values that are independent of C and proportional (in the multifactor model) to S. The allowed region for both semantic retrieval and infinite recursion, therefore, is close to the upper right corner of both the $p - C$ and the $p - S$ plane (uniform light area). The transition from dark to light should be sudden in a system with large S and C (akin to a percolation phase transition).

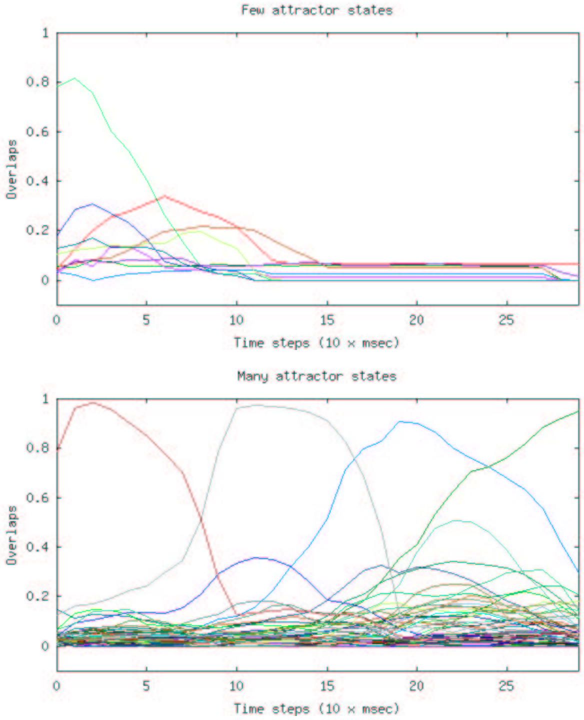

Fig. 7. Examples of global attractor retrieval with and without ensuing structured transitions among attractor states. Both examples were produced by simulating a Potts model with 300 units, $S = 10$, global patterns generated by a multi-factor model with $\zeta = 0.02$, and by applying a cue to 50% of the Potts units. Top panel, $C = 5$ and $p = 10$, and selective activity decays away after retrieval, as a second attractor is almost recruited, but it does not have a sufficient overlap with the first to emerge above an effective threshold. Bottom panel, $C = 25$ and $p = 50$, and a sequence of attractors dynamically replace each other, with the next one being recruited by its strong association with the previous one, thus generating structured transitions.

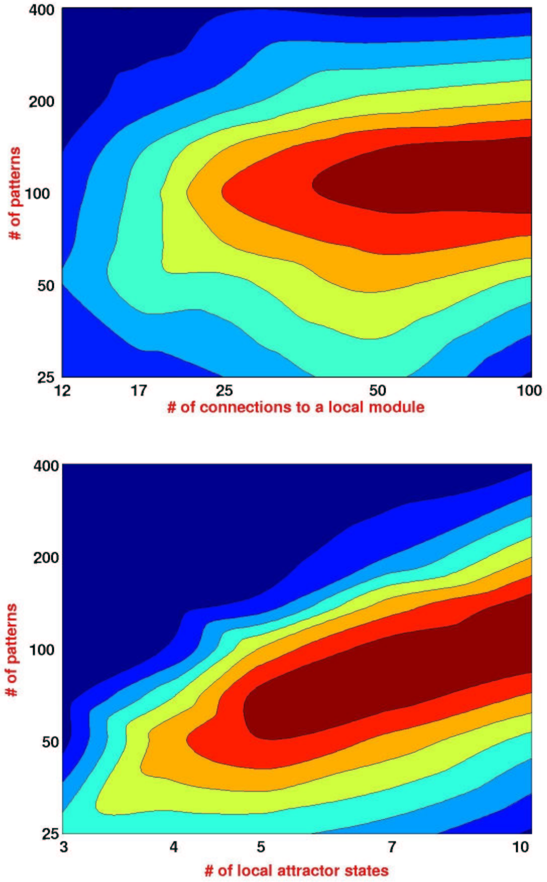

Fig. 8. Simulation results expressed as phase diagrams similar to Figure 6, but plotted on bi-logarithmic scales. In both the $p - C$ (top) and the $p - S$ plane (bottom), what is plotted in shades of gray is the product of a measure of retrieval ability (the degree to which activity is still best correlated with the cued pattern after 7 time steps) with a measure of latching ability (the degree with which after 30 time steps activity is still specifically correlated with one pattern, but not with the one cued). Both measures run from 0 to 1, and white corresponds to their product being higher than 0.3. Each diagram was obtained with 5x7 simulation datapoints, interpolated by Matlab. A datapoint reports the average of thousands of simulations with identical parameters.