Biologically Inspired Cognitive Architectures (2012) 1, 44–54

Available at www.sciencedirect.com

SciVerse ScienceDirect

journal homepage: www.elsevier.com/locate/bica

RESEARCH ARTICLE

Advantages of dilution in the connectivity of attractor networks in the brain

Edmund T. Rolls

Oxford Centre for Computational Neuroscience, Oxford, UK

Received 5 March 2012; received in revised form 28 March 2012; accepted 28 March 2012

KEYWORDS
Diluted connectivity;
Cortical network;
Memory capacity;
Attractor network;
Hippocampus;
Competitive network

Abstract
A fundamental question about brain function is why the connectivity in the cortex is diluted, in that neurons in a local region of the neocortex and in the CA3 part of the hippocampal cortex typically have a probability of having a synaptic connection between them that is less than 0.1. In both these types of cortex, there is evidence that the excitatory interconnections between neurons are associatively modifiable, and that the system supports attractor dynamics that enable memories to be stored, which are used in for example short-term memory and in episodic memory. The hypothesis proposed is that the diluted connectivity allows biological processes that set up synaptic connections between neurons to arrange for there to be only very rarely more than one synaptic connection between any pair of neurons. If probabilistically there were more than one connection between any two neurons, it is shown by simulation of an autoassociation attractor network that such connections would dominate the attractor states into which the network could enter and be stable, thus strongly reducing the memory capacity of the network (the number of memories that can be stored and correctly retrieved), below the normal large capacity for diluted connectivity. Diluted connectivity between neurons in the cortex thus has an important role in allowing high capacity of memory networks in the cortex, and helping to ensuring that the critical capacity is not reached at which overloading occurs leading to an impairment in the ability to retrieve any memories from the network. This intra-area diluted connectivity complements the diluted connectivity in the feedforward connections between cortical areas that helps the representations built by competitive learning to be stable.

Introduction

A key feature of the architecture of the neocortex and the CA3 region of the hippocampus is that there are many excitatory interconnections between the pyramidal cells that are

E-mail address: Edmund.Rolls@oxcns.org
URL: http://www.oxcns.org

2212-683X/$ - see front matter
http://dx.doi.org/10.1016/j.bica.2012.03.003

associatively modifiable. This prototypical architecture provides the basis for cortical attractor networks which enable memories to be stored and recalled from a fragment (Rolls, 2008, 2010); for continuing neuronal firing to implement short-term memory (Rolls, 2008) and thereby to provide for top-down attention (Rolls, 2008) by providing the source of the biased competition (Desimone & Duncan, 1995; Deco & Rolls, 2005) or biased activation (Grabenhorst & Rolls, 2010); and for decision-making when there is competition between two inputs to produce a high firing rate attractor state which represents the decision (Deco, Rolls, Albantakis, & Romo, 2012; Rolls, 2008; Rolls & Deco, 2010; Wang, 2008). However, a key property of this connectivity is that it is diluted, that is the probability that any one hippocampal CA3 neuron will contact another is approximately 0.04 (Rolls, 1989, Chap. 13; Treves & Rolls, 1992), and the probability that a neocortical pyramidal cell will contact another nearby neocortical pyramidal cell is in the order of 0.1 (see below and Rolls (2008)). In this paper, the fundamental issue in brain design of why cortical connectivity is diluted is addressed.

In both the hippocampal cortex and neocortex, there is evidence that the excitatory interconnections between neurons are associatively modifiable, and that the system supports attractor dynamics that enable memories to be stored (Rolls, 2008). In the hippocampal cortex, the memory stored may be episodic, about a particular recent event or episode (Rolls, 2008, 2010; Rolls & Kesner, 2006) (such as where one parked one's bicycle today, or where one was for dinner yesterday, with whom, and what ideas were discussed). In the neocortex, the memory might be a long-term semantic memory (McClelland, McNaughton, & O'Reilly, 1995; Rolls, 2008) (for example the classification of plant and animal species). The prototypical cortical neuronal network for the storage of these memories is an autoassociation or attractor network in which the recurrent collateral connections make associatively modifiable synapses onto neurons in the same network (Amit, 1989; Hertz, Krogh, & Palmer, 1991; Hopfield, 1982; Kohonen, Oja, & Lehtio, 1981, Cha 4; Rolls, 2008; Rolls & Treves, 1998). A memory in such a system is implemented by the set of the neuronal firing rates across the whole population when the memory is stored. The particular firing rate vector comprised of the firing chosen at random of all the neurons is one of the memory patterns that is stored and must later be retrieved. During retrieval, presentation of even a fragment of the memory can produce recall of the whole memory (Rolls, 2008). If a sparse distributed representation is used (in which a small proportion of the neurons chosen at random fires for each memory pattern), then the number of different memories (or memory patterns) that can be stored and correctly retrieved is in the order of the number of recurrent collateral synapses onto each neuron, that is in the order of 10,000 (Rolls, 2008; Rolls & Treves, 1998; Treves, 1991; Treves & Rolls, 1991). The accuracy of the retrieval of each memory pattern can be measured by the correlation between the retrieved firing rate pattern (i.e. vector of firing rates) and the stored memory pattern.

A difference between the hippocampal cortex and the neocortex is that the CA3 network is a single network allowing any representation to be associated with any other representation, providing an implementation of episodic memory (Rolls, 2008, 2010). In contrast, the neocortex has local connectivity with a radius of approximately 2 mm, and this enables the whole of the cerebral cortex to have many separate attractor networks, each storing a large number of memories (O'Kane & Treves, 1992; Rolls, 2008).

The hypothesis proposed here is that this diluted connectivity in the cortex allows biological processes that set up synaptic connections between neurons to arrange for there to be only very rarely more than one synaptic connection between any two neurons. If probabilistically there was more than one connection between any two neurons, it is shown here that such multiple connections between a proportion of the neurons would dominate the attractor states into which the network could enter and be stable, thus severely reducing the memory capacity of the network below the normal limit for diluted connectivity, which is approximately that of a fully connected network with the same number of recurrent collateral connections onto any neuron (Bovier & Gayrard, 1992; Rolls & Treves, 1998; Rolls, Treves, Foster, & Perez-Vicente, 1997; Treves, 1991; Treves & Rolls, 1991). The implication is that a major advantage of the diluted connectivity in the cortex is that it makes it likely that there is rarely more than one synapse between any pair of neurons, and the computational advantage of this is that the memory capacity is high, that is, depends to the first order on the number of recurrent collateral connections C per neuron, and is not greatly reduced below that by the presence of many instances of multiple connections between pairs of neurons. This provides a theory for this fundamental aspect of the design of the neocortex and the hippocampal cortex. Other advantages of the diluted connectivity are presented in the Discussion.

Methods

The autoassociative or attractor network architecture being studied

The architecture and functional properties of autoassociative or attractor networks are described in detail by Hertz et al. (1991), by Amit (1989), and by Rolls (2008). Here it is assumed that the memory patterns are stored in the autoassociative network by an associative (or Hebbian) learning process in an architecture of the type shown in Fig. 1 as follows. The firing of every output neuron i is forced to a value y_i determined by the external input e_i. Then a Hebb-like associative local learning rule is applied to the recurrent synapses in the network:

$$\delta w_{ij} = \alpha y_i y_j. \quad (1)$$

where α is a learning rate constant, and y_j is the presynaptic firing rate.

During recall, the external input e_i is applied, and produces output firing, operating through the non-linear activation function described below. The firing is fed back by the recurrent collateral axons shown in Fig. 1 to produce activation of each output neuron through the modified synapses on each output neuron. The activation h_i produced by the recurrent collateral effect on the ith neuron is the sum of the activations produced in proportion to the firing rate of each axon y_j operating through each modified synapse w_{ij}, that is,

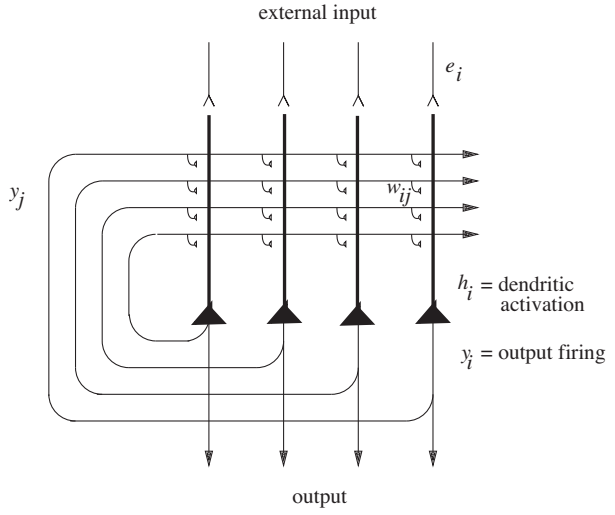

Fig. 1 The architecture of an autoassociative neural network. The external input e_i is applied to each neuron i by unmodifiable synapses. This produces firing y_i of each neuron, or a vector of firing on the output neurons **y**. Each output neuron i is connected by a recurrent collateral connection to the other neurons in the network, via modifiable connection weights w_{ij}. This architecture effectively enables the output firing vector **y** to be associated during learning with itself. Later on, during recall, presentation of part of the external input will force some of the output neurons to fire, but through the recurrent collateral axons and the modified synapses, other neurons in **y** can be brought into activity. This process can be repeated a number of times, and recall of a complete pattern may be perfect. Effectively, a pattern can be recalled or recognized because of associations formed between its parts. This requires distributed representations.

$$h_i = \sum_j^C y_j w_{ij} \tag{2}$$

where $\sum_j^C$ indicates that the sum is over the C input axons to each neuron, indexed by j.

The output firing y_i is a function of the activation produced by the recurrent collateral effect (internal recall) and by the external input (e_i):

$$y_i = f(h_i + e_i) \tag{3}$$

The activation function should be non-linear, and may be for example binary threshold, linear threshold, sigmoid, etc. (Hertz et al., 1991; Hopfield, 1982; Rolls, 2008).

The storage capacity of attractor networks with diluted connectivity

With non-linear neurons used in the network, the capacity can be measured in terms of the number of input memory patterns **y** (each a firing rate vector comprised by the firing rate of each neuron forming a vector of firing rates across the population of neurons) produced by the external input **e**, see Fig. 1, that can be stored in the network and recalled later, even from a fragment of the stored memory pattern, whenever the network settles within each stored pattern's basin of attraction. The accuracy of the recall of each memory pattern can be measured by the correlation between the recalled firing rate vector and the stored firing rate vector. The first quantitative analysis of storage capacity, measured by the number of memory patterns that can be stored and later recalled correctly, considered a fully connected (Hopfield (1982)) autoassociator model, in which neurons are binary elements with an equal probability of being 'on' or 'off' in each pattern, and the number C of inputs per neuron is the same as the number N of output units (Amit, Gutfreund, & Sompolinsky, 1987). (Actually it is equal to $N - 1$, since a neuron is taken not to connect to itself.) Learning is taken to occur by clamping the desired patterns on the network and using a modified Hebb rule, in which the mean of the presynaptic and postsynaptic firings is subtracted from the firing on any one learning trial (this amounts to a covariance learning rule, and is described more fully in Appendix A4 of Rolls & Treves (1998)). With such fully distributed binary random patterns, the number of patterns that can be learned is (for C large) $p \approx 0.14C = 0.14N$, hence well below what could be achieved with orthogonal patterns or with an 'orthogonalizing' synaptic matrix (Amit et al., 1987; Hopfield, 1982). Many variations of this 'standard' autoassociator model have been analyzed subsequently (Amit, 1989; Hertz et al., 1991; Rolls & Treves, 1998).

This analysis has been extended to autoassociation networks that are much more biologically relevant in the following ways (Rolls, 2008; Rolls & Treves,1998; Rolls et al., 1997; Treves,1990, 1991; Treves & Rolls, 1991). First, some or many connections between the recurrent collaterals and the dendrites are missing (this is referred to as diluted connectivity, and results in a non-symmetric synaptic connection matrix in which w_{ij} does not equal w_{ji}, one of the original assumptions made in order to introduce the energy formalism in the Hopfield (1982) model). Second, the neurons need not be restricted to binary threshold neurons, but can have a threshold linear activation function (see Fig. 1.3 of Rolls (2008)). This enables the neurons to assume real continuously graded firing rates to different stimuli (Treves, 1990; Treves & Rolls, 1991), which are what is found in the brain (Rolls, 2008; Rolls & Tovee, 1995; Rolls & Treves, 2011; Treves, Panzeri, Rolls, Booth, & Wakeman, 1999). Third, the representation need not be fully distributed (with half the neurons 'on', and half 'off'), but instead can have a small proportion of the neurons firing above the spontaneous rate (Treves & Rolls, 1991), which is what is found in parts of the brain such as the hippocampus that are involved in memory (Rolls, 2008). Such a representation is defined as being sparse, and the sparseness a of the representation can be measured, by extending the binary notion of the proportion of neurons that are firing, as

$$a = \frac{\left(\sum_{i=1}^N y_i/N\right)^2}{\sum_{i=1}^N y_i^2/N} \tag{4}$$

where y_i is the firing rate of the ith neuron in the set of N neurons. Treves and Rolls (1991) have shown that such a network does operate efficiently as an autoassociative network, and can store (and recall correctly) a number of different patterns P as follows

$$P \approx \frac{C^{RC}}{a \ln\left(\frac{1}{a}\right)} k \tag{5}$$

where C^{RC} is the number of synapses on the dendrites of each neuron devoted to the recurrent collaterals from other neurons in the network, and k is a factor that depends weakly on the detailed structure of the rate distribution, on the connectivity pattern, etc., but is roughly in the order of 0.2–0.3.

The main factors that determine the maximum number of memories that can be stored in an autoassociative network are thus the number of connections on each neuron devoted to the recurrent collaterals, and the sparseness of the representation (Rolls, 2008; Rolls & Treves, 1998; Treves, 1991; Treves & Rolls, 1991). For example, for $C^{RC} = 12{,}000$ and $a = 0.02$, P is calculated to be approximately 36,000. This storage capacity can be realized, with little interference between patterns, if the learning rule includes some form of heterosynaptic long-term depression that counterbalances the effects of associative long-term potentiation (Treves & Rolls (1991); see Appendix A4 of Rolls & Treves (1998)). It should be noted that the number of neurons N (which is greater than C^{RC}, the number of recurrent collateral inputs received by any neuron in the network from the other neurons in the network) is not a parameter that influences the number of different memories that can be stored in the network. The implication of this is that increasing the number of neurons (without increasing the number of connections per neuron) does not increase the number of different patterns that can be stored (see Rolls & Treves (1998) Appendix A4), although it may enable simpler encoding of the firing patterns, for example more orthogonal encoding, to be used. This latter point may account in part for why there are generally in the brain more neurons in a recurrent network than there are connections per neuron (Rolls, 2008). In addition, the random stochastic fluctuations (or 'noise' (Rolls & Deco, 2010)) related to the finite number of spiking neurons is smaller with diluted compared to fully connected networks when the number of connections C per neuron and hence the storage capacity is equated (Rolls & Webb, 2012).

The network simulated

The network can be described under four headings which correspond to the four stages in which the simulation of the network operates. The formal specification of the operation of the network is the same as that of the network analysed by Treves (1990) (see also Treves (1991)) and simulated by Rolls et al. (1997), except where indicated. First, the patterns that the net is to be trained on are binary, that is, a fraction of the neurons, which defines the sparseness a, are 1, and the remainder are 0. Second, the weights are set according to a Hebbian covariance rule. Third, the weight matrix is ablated, that is a proportion of its elements are probabilistically set to zero, to achieve an effective dilution of recurrent connectivity. In other conditions, in addition some of the weights are multiplied by 2, 3, etc. as defined by a Poisson distribution to investigate the effects of multiple synaptic connections between some of the neurons. Fourth, the net undergoes testing with incomplete persistent external cues until the state has settled into retrieval or otherwise.

Pattern generation

Random binary patterns with a sparseness a of 0.1 were used.

The sparseness of the retrieved patterns was measured, to ensure that the network was operating in such a way that the sparseness of the retrieved patterns was close to that of the stored patterns. In these simulations, in contrast to earlier simulations (Rolls et al., 1997), the sparseness was set by altering the threshold T_{thr} for the activation of a neuron to produce firing in such a way that the sparseness reached a value of $a = 0.1$. This has the advantage that it can be implemented by an automatic algorithm, and ensures that the sparseness of the retrieved pattern is close to the value desired of a.

Learning

The learning mechanism is a form of Hebbian covariance synaptic modification, a one-step application of a simple rule which takes account of simple pairwise covariance relationships within each pattern. The exact rule is as follows. Note that the form of the covariance rule is commutative with respect to units i and j, therefore forcing a fully connected net with such a rule to have symmetric weights.

$$w_{ij} = \frac{1}{Na^2}\sum_{\mu=1}^{p}\left(y_i^{\mu} - a\right)\left(y_j^{\mu} - a\right) \tag{6}$$

where w_{ij} is the weight between units i and j. y_i^{μ} represents the firing rate of unit i within pattern μ. This is a simple covariance rule, and a represents the mean activation of the net.

Connectivity

Networks with asymmetrical dilution of the connectivity were investigated. The dilution applied is described in the Results where the effects of different types of dilution, and of some probability of multiple synapses between pairs of neurons, are analyzed. The total number of neurons is N, each neuron receives exactly C inputs, the dilution is C/N, and $P = \alpha C$ is the number of patterns stored in the net. The critical loading of the net, when it fails to operate as a memory, is denoted as α_c.

Testing recall by the net

During recall, the activity of each neuron in the network was asynchronously updated according to a rule which, by analogy with the theoretical analysis, considered a local field h_i at each unit i consisting of an internal field and external field, as follows:

$$h_i = \sum_{j(\neq 1)} w_{ij} y_j + \sum_{\mu} s^{\mu}\frac{e_i^{\mu}}{a} \tag{7}$$

where y_j represents the output of neuron j, and s^{μ} represents the relative strength of pattern μ, see below.

The external field (the last term in the above equation with e_i^{μ} the firing rate of the external input to neuron i produced by the pattern μ) is equivalent to the clamping, persistent external cue, which is believed to be provided by for example the direct perforant path afferents into CA3 from entorhinal cortex (Treves & Rolls, 1992). The ratio between

the average number of perforant path synapses per CA3 cell and that of the recurrent collaterals is in this model allowed to determine their relative influence on the firing of CA3 cells. Anatomical evidence available from the rat suggests that the ratio of the external input (the retrieval cue) to the internal recall provided by the recurrent collaterals should be in the order of 0.25 (see Treves & Rolls, 1992), and s^{μ} was set to produce this ratio (for example when the retrieval cue had a correlation of 0.5 with the originally learned pattern).

The internal field (the first term in the above equation) is equivalent to the recurrent activation provided by the recurrent collaterals in CA3. This is implemented through a standard autoassociation update rule involving weighted inputs from each of the other units. As explained above, this is qualified by the connectivity enforced through zero weights.

The activation function of the neurons is a threshold linear function of the local field h_i, with a gain factor g described by Treves (1990) set to 0.5, and with the threshold T_{thr} adjusted after each iteration in the recall to produce a retrieval sparseness that was close to 0.1 (which was the sparseness of the stored patterns), as described above.

$$y = \begin{cases} g(h - T_{\text{thr}}) & , \ h > T_{\text{thr}} \\ 0 & , \ h < T_{\text{thr}} \end{cases}$$

The recall of the net was measured by the correlation of the retrieved pattern with that stored, when incomplete retrieval cues were used. The performance of the network was also measured by the information retrievable from the network in bits per synapse about the set of stored patterns, as follows

$$I = \frac{\alpha}{\log_2} \sum_{k=0}^{m} \sum_{l=0}^{n} c_k^l \log \frac{c_k^l}{c_k c^l} \tag{8}$$

with Treves (1990) and Rolls (2008) providing further details. Briefly, for each element of the retrieved network state and the corresponding stored pattern, the firing was first discretised into bins, and then the expression above was evaluated. In the above, c_k is the probability that the pattern element is in the kth bin of m bins, c^l is the probability that the retrieved element is in the lth bin of n bins, and c_k^l is the probability that the retrieved element is in the lth bin, and the pattern element is in the kth bin. In the implementation of this calculation, due to practical limitations, both the patterns and network states were binned into 15 bins. Note that the factor α means that the result is in bits per synapse, which is proportional to the total information stored in, and retrievable from, the whole network.

Parameters

The network functioned with a set of parameters chosen to be biologically relevant. Where the parameters are not in correspondence with measurable quantities, they were optimized to the values required for the theory to apply (Treves, 1990). This subsection details some of these parameters, and the reason for their choice.

The total number of neurons in the net was set to 1000 for the results reported.

The gain parameter, equivalent to the gradient of the linear threshold function producing the output of each unit from its incoming local field, was tested at a number of different values, and found to be optimal in the region of $g = 0.5$ for binary neurons. The actual value of g used was 0.5 for the binary patterns.

The loading α was expressed as the ratio P/C, where P is the number of patterns stored, and C is the number of connections per neuron. The loading was varied between 0.1 and 1.2 to investigate the maximum value of the storage capacity α_c, in terms of patterns, as well as to investigate the effect of over-loading. The number of patterns was varied from 40 to 480. The net was allowed to iterate for a maximum of 30 epochs.

Results

First a model of how the connections may be set up between neurons is formulated. The initial value of the synaptic weights is 0. This is only the prescription for whether a neuron i will receive a synaptic connection from neuron j. Let us assume that some genetic factors set the number of connections to be received by a neuron of class A from class B to be a constant C (Rolls & Stringer, 2000), the number of neurons in the network to be N, and the average connection probability to be p, so that $C = pN$. Previous analyses (Rolls & Treves, 1998; Treves, 1991; Treves & Rolls,1991) have assumed that when diluted connectivity is present there are exactly C connections onto each neuron i, that $C < N$, that the connections between pairs of neurons need not be reciprocal and symmetrical, and that there is at most 1 connection from neuron i to neuron j and vice versa. It is this last condition that is relaxed in this investigation, to examine the consequence of different scenarios that might arise according to different prescriptions for determining whether there is a synaptic connection between any two neurons in the network. Three prescriptions are considered.

Full connectivity

With full connectivity, there is one and only one synapse in a given direction between neurons i and j, and all neurons are reciprocally connected with equal weights in the two directions. This situation arises in a fully connected autoassociation network trained with an associative (Hebbian) synaptic modification rule, which produces a symmetric synaptic weight matrix (Rolls, 2008).

Although this is the system favoured for formal analysis (Amit, 1989; Hopfield,1982), I suggest that this would be extremely difficult for real biological systems to set up. There would have to be a biological system that would have to detect whether among something in the order of say 10,000 synapses being received by neuron i (Rolls, 2008), there was more than one from neuron j. This is rejected for the purposes of this paper as being implausible. There would also have to be a mechanism for ensuring that every neuron in the set of N neurons had at least 1 connection to every other neuron in the set of N neurons. This is rejected for the purposes of this paper as being implausible. There would also need to be a way for genes to specify that a particular set of N neurons specified a particular network. This is rejected for the purposes of this paper as being implausible. Thus it is argued that fully connected networks are unlikely to be found in the brain if they are set up by a simple

mechanism such as a neuron making synapses at random with nearby neurons. Thus this genetically simple mechanism to prescribe connections for neurons to make would not lead to a fully connected network.

Connectivity with on average one synapse in a given direction between neurons *j* and *i*

Let us assume that among a population of *N* neurons, a biological process forms connections at random to any one neuron *i* from the other neurons *j* until the average number of connections to any neuron *i* from any one neuron *j* is 1. In this situation, some neurons *i* will receive more than one connection from neuron *j*, and some will receive 0 connections from neuron *j*. In fact, the number of connections received by neuron *i* from neuron *j* will follow approximately a Poisson distribution with mean (λ) = 1. Given that all neurons are trained using an associative synaptic modification rule of the form shown in Eq. (1), this will mean that some neuron pairs will have a synaptic strength that is twice, three times, etc. the strength of the connections between the neurons with one synapse between them. It is suggested, and tested next, that the consequences of this are that the energy landscape is considerably deformed. Hopfield (1982) was able to show that in a fully connected network trained associatively the recall state can be thought of as the local minimum in an energy landscape, where the energy would be defined as

$$E = -\frac{1}{2}\sum_{i,j} w_{ij}(y_i - \langle y \rangle)(y_j - \langle y \rangle). \quad (9)$$

This equation can be understood in the following way. If two neurons are both firing above their mean rate (denoted by <*y*>), and are connected by a weight with a positive value, then the firing of these two neurons is consistent with each other, and they mutually support each other, so that they contribute to the system's tendency to remain stable. If across the whole network such mutual support is generally provided, then no further change will take place, and the system will indeed remain stable. If, on the other hand, either of our pair of neurons was not firing, or if the connecting weight had a negative value, the neurons would not support each other, and indeed the tendency would be for the neurons to try to alter ('flip' in the case of binary units) the state of the other. This would be repeated across the whole network until a situation in which most mutual support, and least 'frustration', was reached. What makes it possible to define an energy function and for these points to hold is that the matrix is symmetric (see Amit, 1989; Hopfield, 1982; Hertz et al., 1991).

It is shown next by simulations that if connectivity of the type defined in this section is present, with some pairs of neurons having several connections between them, that the energy landscape is distorted in such a way that there are deep energy minima in some parts of the energy landscape, and this reduces the capacity of the network, that is, its ability to store 0.14 *N* binary patterns with sparseness *a* = 0.5 (Hopfield, 1982), or a larger number of patterns with more sparse representations as shown in Eq. (5) (Treves & Rolls, 1991).

Simulations were run with the attractor network described in Section ''The network simulated'' and by Rolls et al. (1997) with 1000 neurons. The sparseness of the randomly chosen binary patterns was 0.1. The connectivity was set to have a mean number of connections between any pair of neurons, λ, equal to 1.0 with a Poisson distribution, resulting in the number of connections between pairs of neurons shown in Table 1. The algorithm for setting the number of connections was to select an output neuron *i*, and then set it to have exactly the number of inputs from each of the other neurons in the network (chosen in a random sequence to prevent connections from a neuron already chosen) shown in Table 1. This resulted in an asymmetric connection matrix with each neuron *i* having exactly the number of 0, single, double, etc. connections to other neurons in the network shown in Table 1. In practice, the algorithm was applied after training by multiplying the synaptic weights by the values shown in Table 1. The resulting synaptic connectivity matrix thus reflected what would be produced by any one neuron having a mean number of connections with other neurons set according to a Poisson distribution with the mean number of connections between any pair of neurons, λ, equal to 1. In some cases a pair of neurons would have no synaptic connections, and in other cases 1, 2, 3, etc. synaptic connections as shown by the probabilities and numbers for a network of 1000 neurons shown in Table 1.

Fig. 2a and b shows the results of applying just the degree of dilution of the synaptic weights indicated in Table 1 as the baseline control condition, without any double, triple, etc. synapses. In this case, there were 368 weights of zero out of the starting number of *C* = 1000, and the remainder were multiplied by 1. The correlation of the final state of the network after recall with the stored pattern, as a function of loading $\alpha = P/C$, is shown in Fig. 2a. The two lines in each graph correspond to two retrieval cue levels: the lower line is for a retrieval cue correlated 0.5—0.55 with the original stored pattern, and the upper line is for a retrieval cue correlation of 0.9—0.95. The information retrieved in bits per synapse after recall with the stored pattern, as a function of loading $\alpha = P/C$ is shown in Fig. 2b. (In this Figure, the loading value α shown on the abscissa refers to the fully connected case with *C* = 1000 connections per neuron, and should be multiplied by 1.58 given that with the dilution there are in fact *C* = 638 connections per neuron, as shown in Table 1.) The results in Fig. 2a and b show good performance up to high loading levels when only dilution of the connectivity is present, and there are no multiple synapses between neuron pairs present.

These results emphasize that with diluted connectivity, and asymmetric connections between pairs of neurons, the attractor network still displays its predicted memory capacity, and ability to complete memories from incomplete patterns (Bovier & Gayrard, 1992; Perez Castillo & Skantzos, 2004; Rolls & Treves, 1998; Rolls et al., 1997; Treves, 1991; Treves & Rolls, 1991).

Fig. 2c and d shows the results of applying the identical degree of dilution of the synaptic weights to that used in Fig. 2a and b, but now with the number of double, triple, etc. weight synapses indicated in Table 1 column 3. At low levels of loading $\alpha \leqslant 0.4$ the performance is as good as or better than the control baseline dilution-only simulation shown in Fig. 2a and b, but at higher loading levels the performance drops off very greatly (Fig. 2c and d).

Table 1 The probability of different numbers of connections X between neurons for different values of λ, the average number of connections between neurons in the network, based on a Poisson distribution. The column labelled $C(N = 1000)$ shows the numbers of synaptic connections to a neuron i multiplied by X in prescribing a network with $N = 1000$ neurons with $\lambda = 1$.

X	$\lambda = 1$	$C(N = 1000)$	$\lambda = 0.1$	$\lambda = 0.04$
0	0.3679	368	0.9048	0.9608
1	0.3679	368	0.0905	0.0384
2	0.1839	184	0.0045	0.0008
3	0.0613	61	0.0002	0.0000
4	0.0153	15	0.0000	0.0000
5	0.0031	3	0.0000	0.0000
6	0.0005	0.5	0.0000	0.0000

Fig. 2a and b shows that α_c, the critical loading capacity of the network beyond which the memory fails and little information can be retrieved from it, is approximately 0.4. With just diluted connectivity, without multiple synaptic connections, the critical capacity α_c was higher than 1.2, as shown in Fig. 2a and b.

It is therefore concluded that prescription 2 (Section ''Connectivity with on average one synapse in a given direction between neurons j and i'') for setting the connectivity is seriously flawed, as it reduces the capacity of the attractor network, i.e. the number of patterns that it can store per synapse onto each neuron (Rolls, 2008; Treves & Rolls, 1991), because of the multiple synapses found between a proportion of the neurons. The interpretation of the loss of capacity with some multiple synapses present is that this distorts the energy landscape by producing irregular deep and broad areas which attract patterns that are some distance away, so that it is not possible to have a large number of approximately equal-size basins of attraction in the energy landscape. The interpretation of the somewhat better performance at low loading with some multiple synapses present is that if there are few basins present, but they are deep and large due to the multiple synapses, then patterns some distance away can be drawn into a nearby attractor, which, given large spacing between the attractor basins, is likely to be the correct attractor basin.

I further hypothesize that double synapses would distort the patterns being stored if they are graded (for example have an exponential distribution of firing rates across the population of neurons for any one stimulus), which is typical of neural representations (Franco, Rolls, Aggelopoulos, & Jerez, 2007; Rolls, 2008; Rolls & Treves, 2011; Rolls et al., 1997; Treves, 1990). It is possible that the graded firing rate representation could no longer be accurately stored: the graded representation would be distorted by the extra firing of neurons with double connections. Simulations of the type described by Rolls et al. (1997) could be performed to check whether this becomes an issue, depending on the number of double, triple, etc. synapses between some neuron pairs.

Diluted connectivity with $C < N$

Let us assume that among a population of N neurons, a biological process forms connections at random between any pair of neurons, so that in the resulting network there are C incoming connections to any of N neurons, with $C < N$. Assuming symmetry, this implies that the number L_{ij} of the connections from any neuron j to any neuron i follows a binomial distribution $\mathrm{Bin}\left(C, \frac{1}{N-1}\right)$:

$$\begin{aligned}\mathrm{Prob}\{L_{ij} = k\} &= \frac{C!}{k!(C-k)!}\left(\frac{1}{N-1}\right)^k\left(1-\frac{1}{N-1}\right)^{C-k} \\ &= \frac{C(C-1)\cdots(C-k+1)}{1\cdot 2\cdots k}(N-1)^{C-k} \end{aligned} \quad (10)$$

where k is between 0 and C. When $C = \lambda N$ and $N \to \infty$, this distribution is well approximated by the Poisson distribution with mean λ.

For example, let us assume a diluted connectivity with the average number of connections per neuron $\lambda = 0.1$. We set up connections with the same general prescription as in prescription 2 (Section ''Connectivity with on average one synapse in a given direction between neurons j and i''). However, if $\lambda = 0.1$, the probability that any two neurons will have two connections between them is 0.0045, as shown in Table 1. (In a network with $N = 1000$ neurons, there would be 91 single connections onto each neuron, and only 5 double connections onto each neuron, with no triple connections.) This is a much smaller probability. Additional simulations of the type illustrated in Fig. 2 were performed with $\lambda = 0.1$. The results showed that having this small proportion of double strength synapses in the network produced very little reduction in the correlation of the retrieved patterns with those stored, or in the capacity of the attractor network to store many patterns using the type of analysis shown in Fig. 2c and d. The number of patterns that can be stored is still of order C (Treves & Rolls, 1991), with the constant k in Eq. (5) little affected by this small number of multiple connections between pairs of neurons.

This scenario applies in the real brain. For example, an estimate for the dilution C/N in the neocortex might be 0.1. (For the neocortex, assuming 10,000 recurrent collaterals per pyramidal cell, that the density of pyramidal cells is 30,000 mm^3 (Rolls, 2008), that the radius of the recurrent collaterals is 1 mm, and that we are dealing with the superficial (or deep) layers of the cortex with a depth of approximately 1 mm, the dilution between the superficial (or deep) pyramidal cells would be approximately 0.1.)

A similar but somewhat more diluted scenario applies in the hippocampus. In the rat hippocampus, there are $N = 300{,}000$ CA3 neurons, and each neuron receives $C = 12{,}000$ synapses (see Fig.3) (Rolls, 1989; Treves & Rolls,

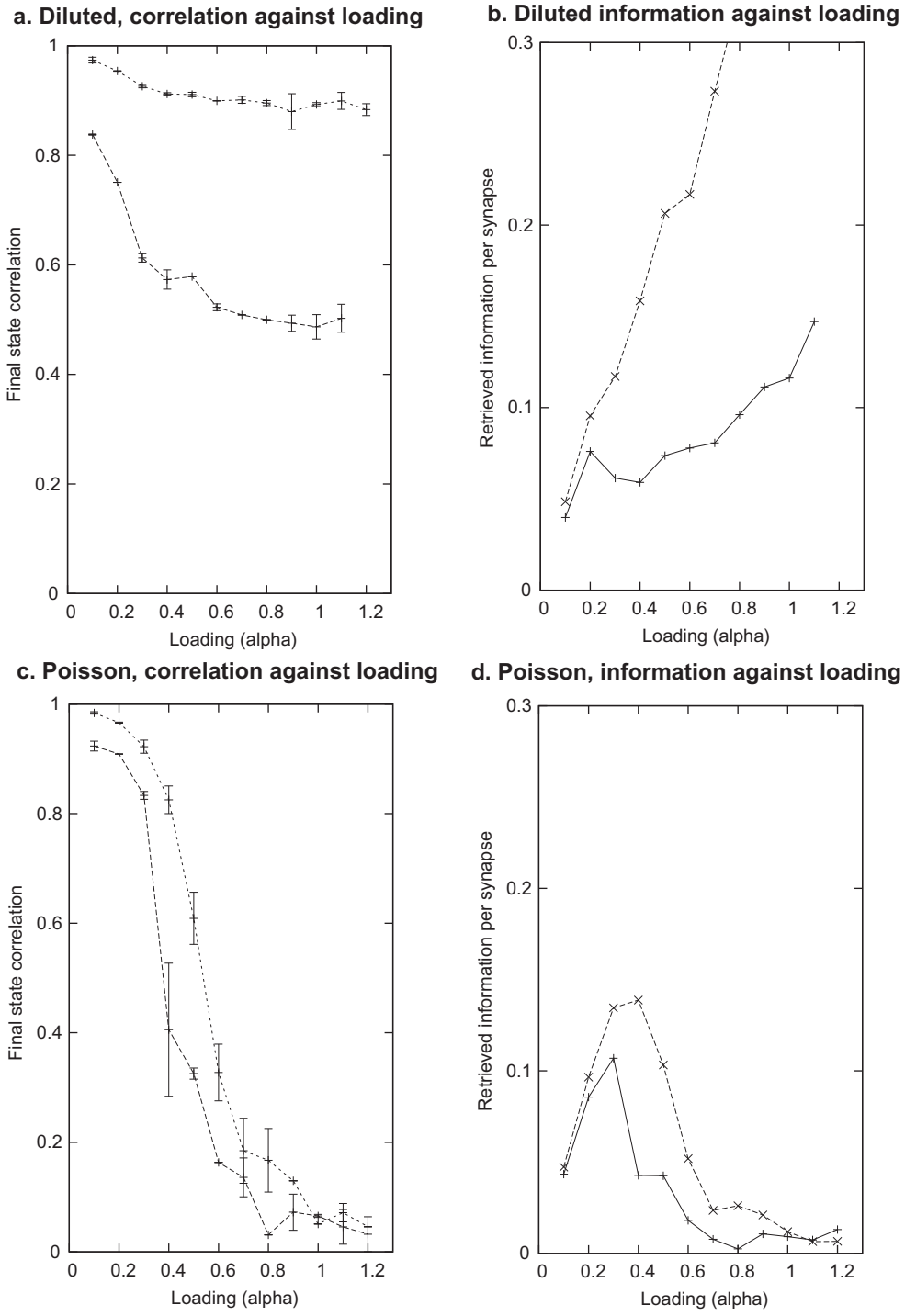

Fig. 2 (a and b) The performance of the attractor network with diluted connectivity only with $\lambda = 1$ for the average number of connections received by each output neuron. In this case, there were 368 weights of zero, and the remainder were multiplied by 1. (c and d) The performance of the attractor network with the numbers of connections between neurons specified by the full Poisson distribution shown in Table 1 diluted connectivity only with $\lambda = 1$ for the average number of connections received by each output neuron. In this case, there were 368 weights of zero, and 368 were multiplied by 1, 184 by 2, 61 by 3, etc. (a and c) The correlation of the final state of the network after recall with the stored pattern, as a function of loading $\alpha = P/C$. The two lines in each graph correspond to two retrieval cue levels: the lower line is for a cue correlated 0.5—0.55 with the original stored pattern, and the upper line is for a cue correlation of 0.9—0.95. The error bars represent the standard deviations. (b and d) The information retrieved in bits per synapse after recall with the stored pattern, as a function of loading $\alpha = P/C$.

1992). (It is assumed that each CA3 neuron correspondingly makes 12,000 recurrent collateral synapses.) The dilution of this network C/N thus equals 0.04. In a network with $N = 1000$ neurons and $\lambda = 0.04$, there would be 38 single

synapses onto each neuron, and only 1 double synapse, as is evident from Table 1. This low number of double synapses, with no triple, etc. synapses, has little effect on the operation of the network, as shown by further simulations of the type illustrated in Fig. 2. It is proposed that this network operates as an autoassociation or attractor network, and is key to how the hippocampus operates to store episodic memories (Rolls, 1996, 2008, 2010) (see Chap. 2 of Rolls (2008)), and here the number of different memories that can be stored in the network is at a premium, and is provided for by this level of dilution of the connectivity which ensures few multiple synapses between any pair of neurons.

Discussion

The results of the simulations therefore support the proposal made in this paper that the reason why networks in the neocortex and hippocampal cortex have diluted connectivity (Rolls, 2008) is that the diluted connectivity ensures that the energy landscape and thereby the memory capacity is not disturbed by multiple connections between pairs of neurons, with the network still operating correctly in the diluted regime where C/N is in the range 0.1—0.01 (Rolls, 2008; Rolls & Treves, 1998; Rolls et al., 1997; Treves, 1991). The somewhat greater dilution of connectivity (with fewer multiple synapses) in the hippocampal CA3 network (0.04) than the estimate of its value in the neocortex (0.1) may be related to the great importance of achieving a high memory capacity with good retrieval of all stored patterns in the hippocampus where this may be useful for episodic memory, for which memory capacity in a single network is important (Rolls, 2008, 2010).

In both these types of cortex, there is evidence that the excitatory interconnections between neurons are associatively modifiable, and that the system supports attractor dynamics that enable memories to be stored (Rolls, 2008).

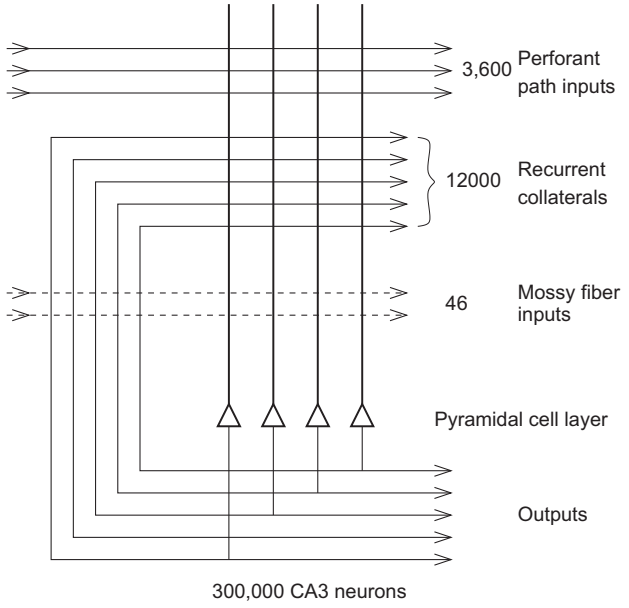

Fig. 3 The numbers of connections onto each CA3 cell from three different sources in the rat. (After Treves and Rolls (1992) and Rolls and Treves (1998).)

One of the points made in this paper is that with a local associative rule, the presynaptic and the postsynaptic activity at a given synapse to determine the strength of that connection between a pair of neurons, which is a widely held assumption (Rolls, 2008; Rolls & Treves, 1998), then consequences of the type described here on memory capacity would ensue. A difference between these cortical structures is that the CA3 network is a single network allowing any representation to be associated with any other representation, providing an implementation of episodic memory (Rolls, 2008; Rolls, 2010). In contrast, the neocortex has local connectivity with a radius of approximately 2 mm, and this enables the whole of the cerebral cortex to have many separate attractor networks, each storing a large number of memories (O'Kane & Treves, 1992; Rolls, 2008).

Another potential advantage of diluted connectivity in which the number of neurons N is greater than C^{RC}, the number of recurrent collateral inputs received by any neuron in the network from the other neurons in the network, is that this may enable simpler encoding of the firing patterns, for example more orthogonal encoding, to be used. For example, much of the information available from the firing rates of a population of neurons about which stimulus was presented can be read by a decoding procedure as simple as a dot product, which is what neurons compute using their synaptic weight vector, and which is very biologically plausible (Rolls, 2008; Rolls & Treves, 2011).

Another advantage of diluted recurrent collateral connectivity is that it can increase the storage capacity α_c of autoassociation (attractor) networks, that is the number of patterns that can be stored per recurrent collateral synapse (Treves & Rolls, 1991). (This useful effect applies provided that representations are not too sparse, because when the sparseness is very low, $a << 0.01$, the performance becomes similar to that of a fully connected network (see Treves & Rolls (1991) Fig. 5a; and Rolls & Treves (1998) Fig. A4.2).)

Another advantage of diluted connectivity (for the same number of connections per neuron) is that this increases the stability and accuracy of the network as there is less spiking-related noise producing stochastic fluctuations in the diluted network (which has more neurons in it), with little cost in increased decision times (Rolls & Webb, 2012).

Another advantage of diluted connectivity is its role in stabilizing competitive networks in the brain (Rolls, 2008). Competitive self-organizing feed-forward unsupervised networks learn to categorize inputs based on the similarity of vectors in the input space (Carpenter, 1997; Hertz et al., 1991; Rolls, 2008). These networks are generally stable if the input statistics are stable. If the input statistics keep varying, then the competitive network will keep following the input statistics. Diluted connectivity can help stability, by making neurons tend to find inputs to categorize in only certain parts of the input space, and then making it difficult for the neuron to wander randomly throughout the space later (Rolls, 2008).

Very diluted connectivity in feedforward networks can also play a role in pattern separation, by encouraging different output neurons to respond to different, random, combinations of the inputs because each neuron is connected to a different combination of inputs. This finds application in the brain in for example the dentate granule cell mossy fibre to CA3 connections which are very dilute but strong. The

architecture helps grid cell representations to be transformed into place or spatial view representations (Rolls, 2008, 2010; Rolls & Kesner, 2006; Rolls, Stringer, & Elliot, 2006; Treves & Rolls, 1992).

In conclusion, it is proposed that the reason why networks in the neocortex and hippocampal cortex have diluted connectivity in the recurrent collateral connections (Rolls, 2008) is that the diluted connectivity ensures that the energy landscape and thereby the memory capacity is not disturbed by multiple connections between pairs of neurons, with the network still operating correctly in the diluted regime where C/N is in the range 0.1–0.01 (Rolls, 2008; Rolls & Treves, 1998; Rolls et al., 1997; Treves, 1991). It is shown here that having a proportion of multiple connections between neurons in an attractor network trained by an associative rule produce a major reduction in $\alpha_c = P/C$, the memory capacity of the network beyond which adding further memories drastically impairs the ability of the network to retrieve any memories. That important reason for dilution in the connectivity of cortical networks, which helps them to be specified by relatively few and simple genetic rules consistent with a limited number of genes in the whole genome in the order of 30,000 compared to the number of synapses which is in the order of 10^{15} (Rolls & Stringer, 2000), is accompanied by other advantages of the dilution of cortical connectivity elucidated in the Discussion.

Acknowledgements

The research was supported by the Oxford Centre for Computational Neuroscience. Professor Alessandro Treves, Cognitive Neuroscience, SISSA, Trieste, Italy, and Professor Tatyana Turova, Department of Mathematics, University of Lund, Sweden, are warmly thanked for interesting discussions and advice.

References

Amit, D. J. (1989). *Modeling brain function. The world of attractor neural networks*. Cambridge: Cambridge University Press.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1987). Statistical mechanics of neural networks near saturation. *Annals of Physics (New York), 173*, 30–67.

Bovier, A., & Gayrard, V. (1992). Rigorous bounds on the storage capacity of the dilute Hopfield model. *Journal of Statistical Physics, 69*, 597–627.

Carpenter, G. A. (1997). Distributed learning, recognition and prediction by ART and ARTMAP neural networks. *Neural Networks, 10*(8), 1473–1494.

Deco, G., & Rolls, E. T. (2005). Neurodynamics of biased competition and cooperation for attention: A model with spiking neurons. *Journal of Neurophysiology, 94*, 295–313.

Deco, G., Rolls, E.T., Albantakis, L. & Romo, R. (2012). Brain mechanisms for perceptual and reward-related decision-making. *Progress in Neurobiology* [Epub 2 February].

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193–222.

Franco, L., Rolls, E. T., Aggelopoulos, N. C., & Jerez, J. M. (2007). Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex. *Biological Cybernetics, 96*, 547–560.

Grabenhorst, F., & Rolls, E. T. (2010). Attentional modulation of affective vs sensory processing: Functional connectivity and a top down biased activation theory of selective attention. *Journal of Neurophysiology, 104*, 1649–1660.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Wokingham, UK: Addison Wesley.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA, 79*, 2554–2558.

Kohonen, T., Oja, E., & Lehtio, P. (1981). Storage and processing of information in distributed memory systems. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (pp. 105–113). Hillsdale, NJ: Erlbaum.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419–457.

O'Kane, D., & Treves, A. (1992). Why the simplest notion of neocortex as an autoassociative memory would not work. *Network, 3*, 379–384.

Perez Castillo, I., & Skantzos, N. S. (2004). The Little-Hopfield model on a sparse random graph. *Journal of Physics A: Mathematical and General, 37*, 9087–9099.

Rolls, E. T. (1989). Functions of neuronal networks in the hippocampus and neocortex in memory. In J. H. Byrne & W. O. Berry (Eds.), *Neural models of plasticity: Experimental and theoretical approaches* (pp. 240–265). San Diego, CA: Academic Press.

Rolls, E. T. (1996). A theory of hippocampal function in memory. *Hippocampus, 6*, 601–620.

Rolls, E. T. (2008). *Memory, attention, and decision-making. A Unifying Computational Neuroscience Approach*. Oxford: Oxford University Press.

Rolls, E. T. (2010). A computational theory of episodic memory formation in the hippocampus. *Behavioural Brain Research, 215*, 180–196.

Rolls, E. T., & Deco, G. (2010). *The noisy brain: Stochastic dynamics as a principle of brain function*. Oxford.: Oxford University Press.

Rolls, E. T., & Kesner, R. P. (2006). A theory of hippocampal function, and tests of the theory. *Progress in Neurobiology, 79*, 1–48.

Rolls, E. T., & Stringer, S. M. (2000). On the design of neural networks in the brain by genetic evolution. *Progress in Neurobiology, 61*, 557–579.

Rolls, E. T., & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology, 73*, 713–726.

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

Rolls, E. T., & Treves, A. (2011). The neuronal encoding of information in the brain. *Progress in Neurobiology, 95*, 448–490.

Rolls, E. T., & Webb, T. J. (2012). Cortical attractor network dynamics with diluted connectivity. *Brain Research, 1434*, 212–225.

Rolls, E. T., Treves, A., Foster, D., & Perez-Vicente, C. (1997). Simulation studies of the CA3 hippocampal subfield modelled as an attractor neural network. *Neural Networks, 10*, 1559–1569.

Rolls, E. T., Stringer, S. M., & Elliot, T. (2006). Entorhinal cortex grid cells can map to hippocampal place cells by competitive learning. *Network: Computation in Neural Systems, 17*, 447–465.

Treves, A. (1990). Graded-response neurons and information encodings in autoassociative memories. *Physical Review, A, 42*, 2418–2430.

Treves, A. (1991). Dilution and sparse coding in threshold-linear nets. *Journal of Physics A: Mathematical and General, 24*(1), 327–335.

Treves, A., & Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain? *Network, 2*, 371–397.

Treves, A., & Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus, 2*, 189—199.

Treves, A., Panzeri, S., Rolls, E. T., Booth, M., & Wakeman, E. A. (1999). Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli. *Neural Computation, 11*, 601—631.

Wang, X.-J. (2008). Decision making in recurrent neuronal circuits. *Neuron, 60*, 215—234.