How Much of the Hippocampus can be Explained by Functional Constraints?

Alessandro Treves[1*], William E Skaggs[2] and Carol A Barnes[2]
[1]*SISSA, Cognitive Neuroscience, Trieste, Italy*
[2]*Neural Systems, Memory and Aging, Arizona Res. Labs, Tucson, AZ, USA*

June 5, 1996

Running title: hippocampal structure explained by function

Keywords: associative memory, storage capacity, redundancy, forgetting, sparse coding

* To whom reprint requests should be addressed at: SISSA, Cognitive Neuroscience, via Beirut 2-4, 34013 Trieste, Italy.

In the spirit of Marr, we discuss an information-theoretic approach that derives from the role of the hippocampus in memory constraints on its anatomical and physiological structure. The observed structure is consistent with such constraints, and, further, we relate the quantitative arguments developed in earlier analytical papers to experimental measures extracted from neuronal recordings in the behaving rat.

1 THIS IS NOT A COMPUTATIONAL MODEL, REALLY

Across scientific disciplines, *computational* usually qualifies approaches based on the use of the *computer*, as opposed to theoretical analysis or physical experiment. In the study of the brain, the term vaguely suggests, in addition, that an approach is aimed at the *computations* performed by a given ensemble of neurons or a given structure. What is reported in this chapter is related to the latter but not much to the former meaning of the word, since it is almost entirely based on formal analytical derivations, not on computer simulations. Moreover, the term *model* usually implies a definite system, as specified by a collection of formulas or by a set of computer instructions or even by an organic or living preparation, chosen to study in simplified form phenomena pertaining to the "original". Our approach is not based on the study of a definite model, but rather on the use of different formal models, each calibrated according to the specific questions asked, and on the model-independent analysis of neuronal activity.

The hippocampus is both a structure emerging from mammalian evolution and a system dedicated to its own particular operations on the information it processes. While some aspects of its organization may be the semi-accidental result of its evolutionary history, for others, in particular for the quantitative values of biologically tunable parameters, it is legitimate to argue that they must be geared to *optimize information processing.* This approach is aimed, then, at understanding which aspects of the organization of the hippocampus – anatomical or physiological – stem from this higher level requirement of optimising the function it performs. At the most abstract level, this function is equivalent to manipulating information in certain ways; accordingly, information theory is the basis of our approach.

Knowledge about hippocampal anatomy and physiology are to be regarded, obviously, as an input to this approach. It is perhaps less obvious, but equally true, that ideas about the role which the hippocampus plays in managing information within the brain are also an input, and not an outcome, of this approach. In short, the goal of the approach is neither to discover structure nor to expound function, but solely to hypothesise or establish *explicit* relations, whenever possible, between structure and function. The success of the approach must be evaluated on the basis of the number of predictive (ideally quantitative) relationships it allows to establish, and the fraction of these that are validated by direct experiment. It should not be evaluated on its inability to explain those aspects of the structure which it does not link to function; nor on the inaccuracy, omissions or fallacy of the structural and functional descriptions it builds upon.

2 THE HIPPOCAMPUS AS A MEMORY DEVICE

Marr's system level theory of the hippocampus (1971) was, in broad terms, the same description of its functional role in memory currently shared by several investigators and taken as an input for the present analysis. His perception of the role of formal models as providing explicit links between structure and function, leading to verifiable predictions (ranked with his curious star system) set the paradigm for others, including us, to follow. What was lacking in his time was a) detailed knowledge about both structure and function, but also, b) the mathematics adequate to analyse formal models refined enough for his purposes. Therefore a) the discussion of how the theory would be implemented within the hippocampus remained rather vague, although it inspired subsequent work in which the correspondence was made more precise (McNaughton and Morris, 1987; Rolls, 1989); b) the quantitative results of his analysis were not applicable to the real system. Nevertheless, Marr's attempt to *explain* the hippocampus remains the most important reference point for later analyses.

Following Marr, the theory considered here for the function of the hippocampus is that it serves as an intermediate-term memory store, in which neural representations of certain events are stored on-line as the events are experienced, and from which they can be retrieved, off-line, by a so-called cue. This is a widespread conceptualization of the role played by the hippocampus, originally based on evidence from human patients (Scoville and Milner, 1957) and later discussed also in the context of experimental findings with other primates and rodents (see the debates following Rawlins, 1985; Eichenbaum *et al.* , 1994). The findings largely agree that memory retention by the hippocampus is limited in time (Squire, 1992); what is more controversial is whether hippocampal forgetting follows the transfer - mediated by cued retrieval - of the same episodic information to neocortical permanent storage sites, possibly after reorganization into a semantic system (*cf.* Gaffan, 1993). In any case, typical forgetting or transfer times would for humans be of the order of years, whereas Marr (1971), who based a similar "transfer" notion on the idea that it would occur during sleep, when neocortex is shut off from sensory inputs, assumed these times to be of the order of days.

The main alternative theory, that the hippocampus operates as a spatial computer (*e.g.,* O'Keefe, 1990), will not be considered here, not as a tacit denial that space has intimate connections with hippocampal function, but rather because we argue that the fact that the information being processed is wholly or partially spatial in nature does not necessarily constrain hippocampal structure (*cf.* Treves *et al.* , 1992). Even within the "memory" camp, a substantial discussion has been devoted to the characterization of (i) the type of information which reaches the hippocampus, and (ii) the transformations it goes through within the hippocampus (*e.g.,* Nadel, 1991, and the "forum" that follows). These are both important aspects, neither of which is attended to in this chapter. We focus only on simpler and more abstract quantitative aspects, such as the amount of information in a representation. For example, whether the spatial information in a representation is in terms of an egocentric or allocentric frame, or refers to the animal's location in the rat or to external space in the monkey, are possibilities which will not be discriminated here, as long as they correspond to xx bits of space information. If it were shown, to continue the example, that knowledge about where the rat is in the arena comes to the hippocampus in polar coordinates and is within the hippocampus transformed to Cartesian coordinates, this would again be irrelevant to the present discussion, except for the possible loss in resolution and information resulting from the transformation.

If we are not going to argue about what the hippocampus feeds itself with, nor about how it digests it, what is it that will matter to us? From our information-theoretical viewpoint, the hippocampus, in order to carry out the memory function it is specialized for, must be able to:

1. generate, on-line, appropriate neuronal representations of the events it has to store in memory;
2. store these representations on-line, and thus in a single shot;
3. hold multiple representations simultaneously in storage within the same system;
4. retrieve each representation from partial cues;
5. send back the retrieved information in a readable and robust format.

These simple requirements in fact significantly constrain the structure of the biological device that must fulfill them, especially if they are taken quantitatively, that is if they have to be met with optimal or near optimal solutions. This is what is discussed next, with a subsection devoted to each of the requirements, which, so as to follow the logic of the argument, are considered in the scrambled order: 4, 2, 3, 5, 1: *a content-addressable memory implemented as a cascade of Hebbian associative networks, with a free autoassociator at its core, a post-processor at the end, and a pre-processor at the front.*

2.1 A content-addressable memory....

Requirement 4, the ability to retrieve information from partial cues, that is from arbitrary subsets of the information to be retrieved, is equivalent to requiring that the device in question be a content-addressable memory. The qualitative, explicit nature of the cue could be further defined as being a sensory component of a multimodal episodic memory, or part of the context, or in many other ways. At a quantitative and abstract level, the utility of such a device arises from the difference between the amount of information that has to be supplied with the cue, and the amount of information that can be retrieved from the device. If this difference is zero or the former is more than the latter, the device does not operate as a memory but merely as a converter (though the activity of individual units within the device may still show memory effects, such as place cells maintaining their specificity in the dark). Therefore, quantifying the *information gain* provided by the system is crucial for establishing whether its role in memory is substantial or purely coincidental. The information content required for the cue to be effective in eliciting retrieval depends on the type of memory and on its load, in the same sense that an e-mail address has a different size, on average, from a regular mail address. If the memory is taken to hold p item (relevant ranges for p are discussed below), the minimum information necessary in the cue is

$$I_{cue} = \log_2 p. \tag{1}$$

The information content of each memory item is also dependent on the type of memory, but in general for a system based on the parallel operation of N processors, if a memory item is represented by the values taken at a point in time by N variables associated with each element, it may be expected to be of order

$$I_{item} \sim N i_{elem}, \tag{2}$$

where i_{elem} is the average information provided by individual elements taken separately[1]. Eq. 2, far from being a a simple change of notation from I_{item} to i_{elem}, would express, if verified, a deep property of the type of representation used by hippocampal cells: its additivity in terms of information, or in other words that different cells convey different information. In contrast, if *redundant* representations were used, I_{item} would be much less than the sum of individual i_{elem}'s; whereas if the representations were *synergistic*, it would be more than the sum.

The most important condition, for our device to be effective as a memory, is then that I_{item} be much larger than I_{cue}, and this will be the case if Eq. 2 is approximately valid, that is if I_{item} grows linearly or almost linearly with N, and of course if N is large. Translated into plain hippocampal terms, the hippocampus can function as a memory if its many cells operate independently or near independently (a few global or quasi-global constraints are acceptable, whereas many detailed mutual constraints on their activity would be functionally destructive) [2].

A very important experimental result, then, is that the hippocampus indeed appears to display such functional diversity, in that the evidence available is consistent with Eq. 2. The collection of such evidence is not simple, because (i) it is only possible to extract the portion of the information contained in a pattern of cellular activations which is about a limited accessible set of correlates, such as position in space for hippocampal cells in the rat; (ii) the bias in information estimates due to limited sampling (Treves and Panzeri, 1995) requires the collection of large amounts of recorded activity; and (iii) recording from multiple cells simultaneously is a major task, which can now be handled but only up to about a hundred cells (Wilson and McNaughton, 1993), still few compared with the hundreds of thousands present in the system. The first aspect, in particular, implies that regardless of whether the information provided by different cells is actually independent, the total measured information will never exceed a ceiling set by the log of the limited number of variables (here, spatial bins) considered.

Figure 1 shows that individual cells contribute independent information, at least until information values begin to saturate as they approach the ceiling of the maximum information possibly associated with spatial position, as defined in the experiment. In other words, different cells are *redundant* only inasmuch as they cannot answer more (about a definite question: here, spatial position) than a full answer (*cf.* Rolls *et al.* , 1996). To a theoretical, infinitely complex question it is possible, and

[1]Essentially, the average difference between the total entropy of the variable associated with the element, and its entropy conditional to a given memory item.

[2]To focus on the common cohesive behaviour of entire populations of cells, as often done in neurophysiology and in brain imaging (talking e.g. about global neuromodulators, or epileptogenesis, or activated areas) is to negate the functionally useful aspects of parallel processing systems, that arise from the diversity and incoherence of individual cell functions.

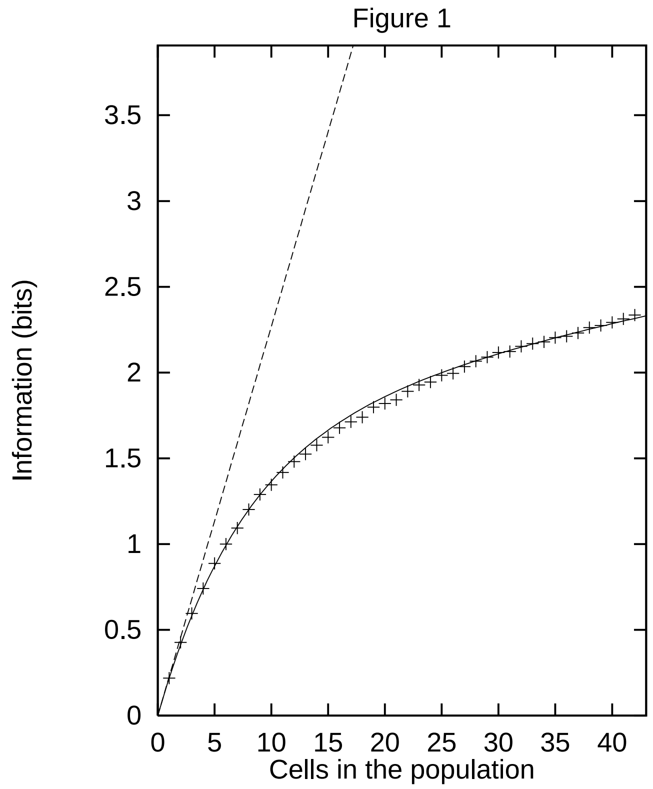

Fig. 1 Average information extracted from subsets of hippocampal cells from a sample of 42, about the position of the rat in a 3-arm maze, *vs.* the size of the subset. The data points (+) are fitted with a simple model (solid line) that explains the deviation from a linear increase solely in terms of ceiling effects: a separate ceiling $I_{max}^{arm} = \log_2 3$ for providing information about which arm the rat is occupying, and one $I_{max}^{segm} = \log_2 5$ for providing information about which of the 5 segments in which each arm has been discretized is currently occupied by the rat. The model is expressed by the equation

$$I(n) = I_{max}^{arm}[1 - \phi_{arm}^{n}] + I_{max}^{segm}[1 - \phi_{segm}^{n}], \tag{3}$$

with the ϕ's fit parameters. The total ceiling $I_{max} = \log_2 15$ (15 being the number of spatial bins) is at the top of the ordinate axis. The linear increase extrapolated from the full information i_{elem} provided by the responses of single cells (not a fit parameter) is also indicated (dashed line). For details about this type of analysis see Rolls *et al.* , (1996).

consistent with current evidence, that a population of N cells would provide an answer with an information content scaling with N, hence much larger than the minimum information required to be supplied with the cue (I_{cue}).

A separate issue to be examined is whether the theoretical minimum for I_{cue}, $\log_2 p$, is enough for a specific implementation of a content addressable memory to effectively retrieve a memory item. This has been shown to be the case for the associative networks discussed below.

2.2 ..implemented as a cascade of Hebbian associative networks...

Requirement 2, the ability to store neuronal representations in one shot, is satisfied by associative neuronal networks operating with *Hebbian* types of synaptic plasticity, as shown by a variety of formal models, including those considered by Willshaw et al (1969), Kohonen (1974), and Hopfield (1982). Among all content addressable memories, many others could be conceived, of course, that could store representations in one shot. We focus, however, on systems made up of neurons, characterized by a weighed summation of inputs from other units (as opposed to a completely arbitrary operation) and connected by synapses whose efficacy can be modulated by activity, but only locally in space and time. This restricts the class of systems with the required ability pretty much to variations of associative networks (Treves and Rolls, 1991). The essential reason for this is not the sophistication of associative networks, but precisely their simplicity. More complex networks (for example backpropagation networks, in which however synaptic plasticity is not a local effect) tend to prefer iterative learning because their complexity yields to instabilities when faced with rapid, one-shot learning (McClelland *et al.* , 1995). A basic associative memory function may nevertheless coexist, as a sort of minimum common denominator, with other functionalities within the same networks.

The crucial ingredient that endows networks of real neurons, operating with distributed representations, with an associative memory ability is a Hebbian type of synaptic plasticity, such as the type known as LTP (long-term potentiation), which is induced, *e.g.*, through the action of NMDA receptors. Associative LTP is present at several synaptic systems in the hippocampus, including the most intensely studied perforant path to granule cells and Schaffer collateral to CA1 systems (but possibly not the mossy fiber to CA3 system). All such systems (not the mossy fibers) operating in cascade, along the direction of preferred activation flow, can contribute to an associative memory function[3]. The involvement of LTP in associative memory has never been conclusively demonstrated, nor it is likely that it will be in the near future (Barnes, 1995); nevertheless, its very existence provides what would otherwise be a missing link in our logic, and a serious one at that.

Turning to a quantitative analysis, the number p of representations that can be stored with any

[3] Plasticity in the mossy fiber system would be useless for associative memory purposes due to the small number of synapses per receiving cell (Treves and Rolls, 1992); but may be useful in satisfying requirement 1 (see below), although interestingly Kandel and coworkers (Huang *et al.* , 1995) failed to find a learning deficit in rats following the selective blockade of such plasticity.

reasonably efficient type of Hebbian plasticity is broadly determined by the relation

$$p < p_c \sim 0.2 \frac{C}{a \log(1/a)} \tag{4}$$

where C is the average number of associative inputs per cell, and a is the average sparseness (Treves, 1990) of the representations. This relation is valid for a large class of networks that store distributed representations (firing patterns) with Hebbian "learning rules" that model associative types of synaptic plasticity. It has been established originally for the sparsely coded version of the Hopfield autoassociator with binary units (Tsodyks and Feigel'man, 1988; Buhmann *et al.* , 1989) and later found to hold also for a variety of specific models with graded response units, that differ in the type of connectivity (the architecture), the statistics of the firing patterns, or the exact learning rule used (Treves and Rolls, 1991). It holds also for more realistic models that incorporate a description of the dynamics of real neurons (see below) and is thus expected to apply to real networks in the brain, even allowing for some semantic structure in the encoding of what the simplest models treat as independent episodic representations[4]

The *sparseness* of the firing patterns, or intuitively speaking the proportion of units active in a representation, is the most important quantity that balances storage capacity with representational capacity. Sparse coding (a small) allows more memory items to be stored, but the information content of each item (hence the representational capacity of the network) decreases. Experimentally, relatively sparse coding is found to prevail in those areas such as the hippocampus (in rats, Barnes *et al.* , 1990; and monkeys, Rolls *et al.* , 1989) which are closely associated with a simple role in associative memory; whereas in areas whose role in memory is thought to be different, and in any case minor with respect to sensory encoding, such as the monkey temporal cortex, firing patterns are found not to be sparse at all (Rolls and Tovee, 1995).

If the quantities C and a vary across the cascade of networks, the relevant ones are of course those that produce the minimum p (that is, the memory bottleneck). Since C, whatever the type of connectivity, is of the order of N or less, p also cannot exceed N by much, and its logarithm, that is I_{cue}, is a small number with respect to I_{item} even if the mutual information per cell, i_{elem} is a small fraction of a bit, as found to be the case with the sparse firing of hippocampal cells (see Figure 1).

If p is limited and the memory device has to function over a lifetime, a need obviously arises for a mechanism that erases old memories and hence allows for the storage of new ones. Among the simplest of these *forgetting* mechanism (forgetting intended at the hippocampal level, not necessarily at the behavioural level) are constraints on the range of variability of synaptic efficacies - which are certainly present at least in the sense that synaptic conductances cannot become negative - and the gradual, passive decay of synaptic enhancement[5]. Denoting with τ_{item} the mean permanence time of a representation in the hippocampal system, and with dp/dt the acquisition rate, we have

[4]Recent evidence (Skaggs and McNaughton, 1996) points at one basic "semantic" aspect that seems to be preserved in hippocampal memories: the temporal order in which different representations were activated. Besides, rat spatial maps can be regarded as a further semantic structure linking the memories of nearby positions in the environment.

[5]In the Marr (1991) model with binary synaptic elements, the limit p_c on p was set by the requirement that enough synapses remain unsaturated, that is at the lower of the two efficacy values. Thus saturation determines

$p \simeq \tau_{item} dp/dt$. If the hippocampus operates at close to its memory capacity, as an efficient system should, then

$$dp/dt \sim 0.2 \frac{C}{\tau_{item} a \log(1/a)} \tag{5}$$

which implies that if the acquisition rate varies strongly across time, some of the other three parameters should be tunable to maintain optimal performance. This could be verified experimentally by manipulating the acquisition rate and monitoring both sparseness and hippocampal forgetting through multiple single unit recording and behavioural procedures, as suggested before (Treves and Rolls, 1994; Treves *et al.* , 1996). Similarly, independent changes in C, for example a reduction with aging, might be partially compensated by tuning a or τ_{item} (Barnes *et al.* , 1994). If, further, τ_{item} is related to the exponential time constant measurable in LTP decay, as their similar reduction with aging suggests (Barnes and McNaughton, 1985), this might allow to follow changes in the time parameter with purely neurophysiological means. These aspects might also be probed, in the future, by direct measures of the average plasticity *in vivo*, as explained in detail elsewhere (Treves *et al.* , 1996).

2.3 ..with a free autoassociator at its core...

We now examine requirement 3, considering how to optimize storage capacity by choosing the most suitable network *architecture*, once the elements available are given (pyramidal cells, NMDA receptors, etc.) and the sparseness is set as yielding the best balance with the information content of each memory. A free autoassociator, that is an autoassociative memory concentrated in a compact network heavily interconnected with recurrent collaterals (Treves and Rolls, 1991), would be the most efficient link in the posited associative chain, on at least two accounts: the ease with which it can store new representations, and the full use it would make of memory space. The first factor arises from each output cell having access, in such an architecture, to both afferent inputs and, through at most a few synaptic steps, collateral inputs from all the other cells in the net. This will be discussed further below. The second factor arises from I_{item} being proportional to N, as discussed above, with this N being in a single-layer network both the total number of cells *and* the number of output cells. Since p is proportional to C, and the *total information* the net stores at a moment in time is just pI_{item}, this information is proportional to CN, which for a free autoassociator is also the number of synapses in the network. In fact, a reasonable estimate, established analysing a wide class of formal models, is that 0.1-0.2 bits could be stored per synapse. Directed associative memory networks (Marr, 1971), instead, are either monolayer, and then unable to produce complete retrieval (Gardner-Medwin, 1976), or multilayer, and then the output cells are just a subset of the cells in the network, as more cells are present that contribute to memory retrieval but do not access stations

---

storage capacity, and decay, which would be implemented as a "flip" back to the lower value, is the one mechanism for forgetting (McNaughton and Barnes, 1990); In the more efficient models that effectively use the formal equivalent of LTD (long-term depression) along that of LTP, storage capacity is determined by the balance between signal and noise, and saturation is an accessory ingredient which, if present, has an overwriting effect similar to (continuous) decay, thus constituting an independent potential mechanism for forgetting (Barnes *et al.* , 1994).

downstream. Although these nets can still function as autoassociators (Treves and Rolls, 1991), they use the available synaptic space less efficiently. We believe that this simple information-theoretic advantage has provided most of the evolutionary pressure for the emergence of the extensive CA3 recurrent collateral system. The value of C, about 12,000 in the rat (Amaral et al, 1990) could be intepreted as having been maximized, in order to increase storage capacity, under the constraint of maintaining cells electrically compact, to ensure that Hebbian plasticity at the dendrites reflects events occurring at the soma. Dendrites operating as effectively independent units (Softky, 1994) would not implement an associative memory.

A recurrent autoassociator can be seen as iterating in time the same operation performed just once in an equivalent but purely feedforward system. This implies *feedback*, which has several effects, and could also be taken to imply a very long time to operate, involving, as it were, repetitive cycles through the recurrent circuit. The second expectation is borne out of considering overly simplified non-dynamical models, but it is contradicted by the analysis of formal models that include the relevant dynamical biophysics of pyramidal neurons (Treves, 1993). This analysis, corroborated by computer simulations, shows that recurrent collaterals can contribute their effect over short times, of the order of the time constant for conductance inactivation at their synapses. Feedback, on the other hand, results in the following: (o) it complicates considerably the analytical methods required to understand the properties of formal models (Amit, 1989); (i) if strong, it allows for self-sustained activation, which can subserve short-term memory; (ii) it amplifies interference among different memory representations, but very little if the coding is sparse (Treves and Rolls, 1991), as it appears to be in the hippocampus; (iii) it can make it difficult for subtractive inhibition to control the activity of the pyramidal cells, while at the same time allowing them to operate efficiently as a memory - solving this conflict requires shunting inhibition (Battaglia and Treves, 1996). Effect (o) is worrisome for the modellers but not for the hippocampus; effect (i) may or may not be used by the hippocampus; effect (ii) is effectively avoided by sparse coding, which is already there for other reasons (see above); effect (iii) merely puts additional constrains, on the organization of the inhibitory component of the circuitry, and disruption of these constraints accounts for the ease with which hippocampal activity can get out of hand, e.g. in epilepsy (Traub and Miles, 1991). The existence of CA3 with its prominent recurrent collateral network suggests that these side-effects are overridden by the two advantages that a compact associative net provides, in terms of learning and in terms of effective usage of synapses.

2.4 ..a post-processor at the end...

The information retrieved within the device, in particular by the autoassociator, has to be sent back to "external users", *i.e.,* neocortical areas, with minimal waste – requirement 5. This can be accomplished by a multilayered system of Hebbian-modifiable backprojecting synapses (see Treves and Rolls, 1994). In addition, the CA1 network can be considered to be both the first step in the relay from CA3 back to neocortex, and the last stage of the hippocampal associative memory system, in other words a dedicated memory post-processor. One crucial advantage of having CA1 after the CA3 stage, is that the very compressed representation provided by CA3 pyramidal cells (in numbers, the bottleneck of the hippocampal system) can be reexpanded onto the larger number of CA1 pyramidal cells. The reexpansion appears to occur across species (Seress, 1988) and results in the same information being coded in a much more robust manner. As an artificial but intuitive example, if N CA3 cells code for $2N$ bits of information by firing at 1 of 4 equiprobable rates, $2N$ CA1 cells can code the same information one bit each, by firing at 1 of just 2 equiprobable levels, with a corresponding increase in the permissible noise level. Obviously the recoding is effective only if at least it *preserves* the overall information content of the representation. Further, CA1 can also contribute to associative retrieval itself by *increasing* this information content over and beyond that of the representation retrieved from CA3. A quantitative assessment of such information content using an analytical model (Treves, 1995) shows that both preservation and increase can occur if the CA3-to-CA1 connections, the Schaffer collaterals, are endowed with associative Hebbian modifiability. In particular, there is an optimal range of the plasticity parameter (measuring essentially the average modification per item as a fraction of the total variance) which is the one that matches the plasticity of CA3 recurrent connections.

The analysis then suggest that the two synaptic systems (within CA3 and from CA3 to CA1) may be optimally organized if they share the same type of synaptic plasticity, in particular the same molecular and biophysical mechanism based on NMDA-receptors. This analysis is now being extended, through an appropriately adapted formalism, to include the direct perforant path projections to CA1 (Panzeri, Fulvi Mari and Treves, in preparation). It is hoped that this will clarify the contribution of direct entorhinal inputs to the information content in the hippocampal output, and provide indications as to the reason for the relative abundance of perforant path and Schaffer collateral synapses onto CA1 cells.

If no substantial increase is contributed by direct perforant inputs, the information *per cell*, I_{item}/N_{CA1}, provided by a population of CA1 cells should then be lower than in CA3 in the inverse ratio of the number of cells, N_{CA3}/N_{CA1}. This is what is found in preliminary analyses of data recorded simultaneously in CA3 and CA1, and kindly provided by Matt Wilson. The ratio is about 1.4 in the rat, and the analysis must select homogeneous samples, with similar single cell firing statistics in the two areas (implying similar values for i_{elem}). In other words, the representation provided by CA1 cells appears, from preliminary evidence, to be slightly more redundant, and hence more robustly coded, than in CA3.

Note that, in addition, firing patterns appear to be sparser in CA3 than in CA1, and mean rates lower (Barnes *et al.* , 1990). The more pronounced sparseness may be related to CA3 being the recurrent autoassociator, hence more subject to interference. The lower mean rates are in principle a separate issue (exactly the same sparseness would be measured if all firing rates were scaled down by a common factor), but slow firing may be related to sparse firing, at least at a general hippocampal level, in the following indirect sense. The distribution, across *e.g.,* spatial positions, of quasi-stationary currents entering the soma has a more or less fixed shape for any cell receiving many small distributed inputs (quasi-Gaussian, in fact); these currents result in a distribution of firing rates through basically a threshold process, and one of the simplest ways to modulate the sparseness of the distribution of rates is to alter thresholds: high thresholds yield sparser distributions than low ones (see *e.g.,* Treves and Rolls, 1992). Therefore the setting of relatively high thresholds, with the side effect of low mean firing rates, may be a mechanism aimed, in the hippocampus, towards sparser codings.

2.5 ..and a pre-processor at the front.

Finally, the first requirement in our list is that adequate mechanisms exist, to generate appropriate representations for memory storage. Appropriate, in the abstract sense considered here, means rich in information about the event that is being represented – as rich as compatible with the compression inherent in using a compact associative network, and with the sparseness required to store a large number of representations. Treves and Rolls (1992) have produced a quantitative argument that indicates the advantage of delegating this task to a specialized system, the dentate gyrus, with its sparse mossy fiber projections onto CA3 cells. A different argument suggests instead that the direct perforant path to CA3 is the system apt to relay to CA3 cells the cue that initiates the retrieval of a representation. For the mossy fiber inputs to provide the required driving effect on the firing of CA3 cells, overcoming the interference resulting from the activation of recurrent collaterals, the mossy fibers need not be quite as strong and specific as *detonators* (*cf.* McNaughton and Morris, 1987): the observed physiological and anatomical properties would be adequate.

A direct experimental check of the theoretical argument entails the quantification of the information contents of CA3 representations learned before and after the inactivation or ablation of the granule cells. In intact rats, the observed values of information per cell as a function of sparseness are consistent with expectations based on formal models, as shown in Figure 2.

Following the near complete destruction of granule cells, the corresponding values should go below the lower curve of Figure 2. This would be a quantitatively sensitive way to probe an effect that may be unclear at the behavioural level. Analysis of data recorded from lesioned rats is in progress and will be reported elsewhere.

The dentate gyrus, if this hypothesis is found to be correct, might thus be regarded as a late addition to the hippocampal system (granule cells have a late ontogenetic development) that serves to greatly increase the information-theoretic efficiency of its associative networks.

The curves in Figure 2 are derived from simple formal models, in which the relevant distribution of firing rates is taken to be the asymptotic limit for long times, whereas the data points, being derived from actual experiments, must correspond to the measurement of firing rates as spike counts over a finite time interval. Experimental evidence however indicates that for a freely running rat the distribution of mean rates at each spatial position (and with it the sparseness of the distribution) does not vary much with the size of the time bin over which they accumulate; whereas the variance of such rates around their means shrinks, as of course it should, with longer bins, tending to a finite non-zero value which represents the intrinsic variability in the rates. Correspondingly, the information extracted from spike counts of increasing bin length rises until it saturates at bin sizes of order a

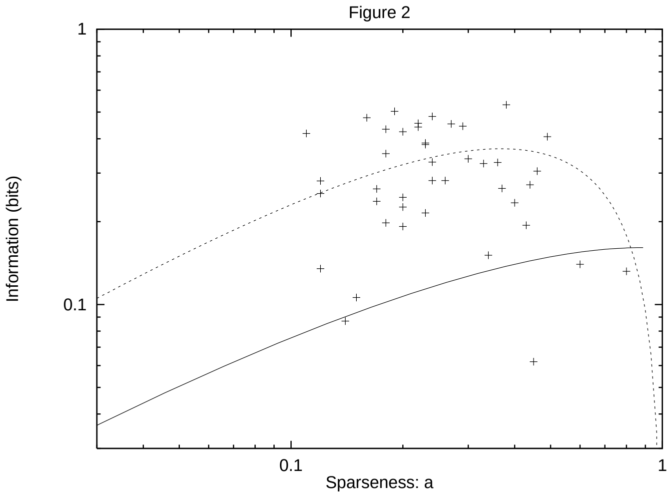

Fig. 2 The information extracted from single hippocampal cells from a sample of 42, about the position of the rat in a 3-arm maze, *vs.* the sparseness of their firing. Given the number of spatial bins, the sparseness can only range from 1/15=0.067 to 1. The upper curve is the expected average trend of $i_{elem}(a) = a\ln(1/a)$ (Treves and Rolls, 1992), while the lower curve is the expected *upper* limit of the data points if mossy fiber inputs were absent in the storage of new representations (see Treves and Rolls, 1992).

second or less, and this can be taken to be the values that should match the asymptotic values of formal models. The instantaneous information rates (Skaggs *et al.* , 1993), instead, depend only on the distribution of mean rates and are relatively constant with the size of the time bin. These facts are illustrated in Figure 3, which exemplifies single cell information kinetics in the hippocampus.

3 DISCUSSION

3.1 What is new in this approach?

The approach itself is not new, because it has been evolving for 25 years, but new powerful analytical methods have emerged for understanding in more detail both formal models of associative memories and data recorded from hippocampal cells. The latter are now obtained with massively parallel simultaneous recordings, which opens up the possibility to analyze the information conveyed by populations instead of just single cells. As a result of these developments, the correspondence between the requirements of the theory and the actual structure of the hippocampus is improving, as documented in Section 2.

3.2 What good is the math?

The arguments recapitulated above are quantitative, and as such they have to rely on the conjunct mathematical analysis of both adequate formal models and recordings of neuronal activity *in vivo.* A quantitative analysis is crucial to understand the relationship between structure and function in the hippocampus, because an entirely different structure may subserve the same qualitative function,

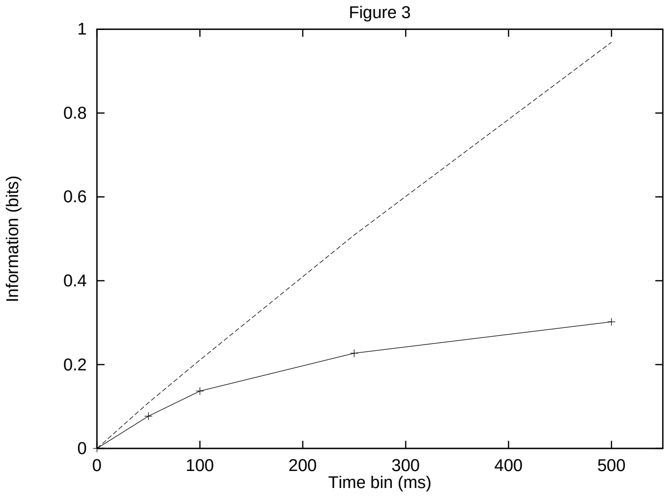

Fig. 3 Average information extracted from single hippocampal cells from a sample of 42, about the position of the rat in a 3-arm maze, *vs.* the size of the time bins used for counting spikes: 50, 100, 250 or 500 ms. For comparison, the information values, obtained by multiplying the instantaneous rates by the corresponding time bin, are also displayed by the dashed curve, which being nearly straight implies almost constant instantaneous information rates.

at the expense of information-theoretic efficiency. For example, a structure without the dentate or even without CA3 can still function as an intermediate term memory that stores and retrieves representations from partial cues. Thus it would not be surprising if behavioural studies were to find limited impairments following complete CA3 ablation. The avian structure that is supposed to be an analogue of the hippocampal formation and is implicated in spatial memory (Krebs *et al.* , 1989)) does not show, as far as the anatomy is known, anything like the same internal structure, yet it may well take a similar functional role, with different efficiency.

A point of central importance, reflected in the suggestions for experiments of Section 3.4, is that behaviour alone, as observable from outside the "black box", is certainly what ultimately matters, but is not a probe sensitive enough to really bear on the relations between structure and function within the hippocampus. Only the behaviour *of the hippocampus*, as observed by recording the activity of its units, can inform a detailed understanding of the hippocampal formation. This cellular behaviour has however to be observed *in vivo*, and of course the concurrent whole animal behaviour is a most useful auxiliary variable to keep track of. On a similar note, stressing the crucial role played by the mathematical analysis of formal models does not mean denying the utility of computer simulations, often important especially in the preliminary investigation of questions not yet reduced to a formulation accessible to analytical methods.

3.3 What is different from other approaches?

Several other researchers share the same or a similar system-level view of the role of the hippocampus in memory, but differ in the particular aspects they address or emphasize in their analysis. This should be evident also from reading other contributions to this issue. Rolls (this issue), for example, has instantiated this view in a real *computational* approach, by implementing a computer simulation of the operation of the hippocampus and entorhinal cortex, along essentially the same lines followed in our arguments. Murre (this issue), with his TraceLink computer model, does not attempt to account for the details of hippocampal circuitry, but rather for the phenomenology of amnesia in humans; but the basic ideas are, again, consistent with ours. In particular, Murre assumes that the fundamental constraint preventing on-line learning of episodic information directly in neocortex, thus generating the need for a dedicated hippocampal system, is the limited long-range cortico-cortical connectivity. In contrast, McClelland (this issue) has suggested that the fundamental constraint is that neocortical memory systems could only accomodate slow learning – and hence would require an auxiliary hippocampal fast system – if they operated like certain connectionist models.

The work of Hasselmo and colleagues (also reported in this issue), on the other hand, although consonant in perspective, focuses specifically on cholinergic modulation of the hippocampus, and proposes a mechanism which may be alternative to that proposed by Treves and Rolls (1992). Based on the observation that acetylcholine suppresses transmission by intrinsic connections and increases cellular excitability and synaptic modifiability, it is proposed that the switching "on" and "off" of the cholinergic input may be sufficient to differentiate between a storage and a retrieval phase, in particular by suppressing, when "on", transmission by CA3 recurrent collaterals, with its interference on the storage of new memories. One should note that the Treves and Rolls proposal does require a mechanism that switches off or attenuates *mossy fiber* inputs to CA3 during the *retrieval* phase; this may result from a decrease in the firing of dentate granule cells, but may also be helped by cholinergic action. However the Hasselmo proposal may be alternative in that, if cholinergic effects suffice in strongly suppressing *recurrent collateral* transmission during *storage*, no need would arise for separate input systems to CA3.

3.4 What is there to do?

There is a very large number of experiments that would greatly improve our understanding of the hippocampus from the viewpoint discussed here. Most important directions to take:

- **Parallel recording from monkeys**. While the considerations above have been mainly sculpted by data recorded in the rat, the rat or even rodent hippocampus is clearly not necessarily representative of the mammalian – and even less of the primate – one. Important new aspects have emerged from single unit recordings in monkeys (O'Mara *et al.* , 1994; Ono *et al.* , 1993) and now actively walking monkeys (Rolls *et al.* , 1995), whose implications will be even more crucial once the activity of populations will be analysed.

- **Gradient of retrograde amnesia**. This has largely been a moot issue (Gaffan, 1993) because its exploration at the behavioural level confounds hippocampal forgetting with that resulting from other factors. An analysis of forgetting at the cellular level, with chronic implants, but also at the population level, without, will clarify (a) its existence and (b) its potential modulation by acquisition rate, aging, etc. (Barnes *et al.* , 1994; Treves *et al.* , 1996).

- **Differences between processing stages**. The comparison between data recorded from different populations (Barnes *et al.* , 1990) has already been very insightful, but more detailed analysis of parallel recordings is needed to quantify the informational properties of each component of the whole system.
- **Selective blockades of plasticity**. Along with selective lesions, the disabling of plasticity at individual synaptic systems, with the fast developing genetic means (*e.g.,* Huang *et al.* , 1995), but even better with non permanent pharmacological means (Barnes, 1995), may allow to verify or disprove the detailed predictions arising from this approach (see Treves and Rolls, 1992, 1994).

Acknowledgments

Different parts of the work described here were in collaborations with Edmund Rolls, Bruce McNaughton, Stefano Panzeri and Francesco Battaglia, all of whom are to be thanked. Partial support from grants AG12609 and MH01227 (USA), CNR94.02931.CT04 (Italy) and ERB-CHRX-CT93-0245 (EC).

References

Amaral DG, Ishizuka N and Claiborne B (1990) Neurons, numbers and the hippocampal network. *Progr Brain Res* **83**, 1-11.

Amit DJ (1989) *Modelling Brain Function*, Cambridge Univ Press, New York.

Barnes CA (1995) Involvement of LTP in memory: Are we "searching under the street light"? *Neuron* **15**, 751-754.

Barnes CA and McNaughton BL (1985) An age comparison of the rates of acquisition and forgetting of spatial information in relation to long-term enhancement of hippocampal synapses. *Behav Neurosci* **99**, 1040-1048.

Barnes CA, McNaughton BL, Mizumori SJY, Leonard BW and Lin L-H (1990) Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Progr Brain Res* **83**, 287-300.

Barnes CA, Treves A, Rao G, and Shen J (1994) Electrophysiological markers of cognitive aging: Region specificity and computational consequences. *Seminars Neurosci* **6**, 359-367.

Battaglia FP and Treves A (1996) Information dynamics in associative memories with spiking neurons. *Soc Neurosci abs* **22**, in press.

Buhmann J, Divko R and Schulten K (1989) Associative memory with high information content. *Phys Rev A* **39**, 2689-2692.

Eichenbaum H, Otto T and Cohen NJ (1994) Two functional components of the hippocampal memory system. *Behav Brain Sci* **17**, 449-472.

Gaffan D (1993) Additive effects of forgetting and fornix transection in the temporal gradient of retrograde amnesia. *Neuropsychologia* **31**, 1055-1066.

Gardner-Medwin AR (1976) The recall of events through the learning of associations between their parts. *Proc Roy Soc London B* **194**, 375-402.

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

Huang Y-T, Kandel ER, Varshavsky L, Brandon EP, Qi M, Idzerda RL, McNight GS and Bourtchouladze R (1995) A genetic test of the effects of mutations in PKA on mossy fiber LTP and its relation to spatial and contextual learning. *Cell* **83**, 1211-1222.

Krebs JR, Sherry DF, Healy SD, Perry VH and Vaccarino AL (1989) Hippocampal specialization of food storing birds. *Proc Natl Acad Sci USA* **86**, 1388-1392.

Kohonen T (1977) *Associative Memory*, Springer, Berlin.

Marr D (1971) Simple memory: A theory for archicortex. *Phil Trans Roy Soc London B* **262**, 24-81.

McClelland JL, McNaughton BL and O'Reilly RC (1995) Why there are complementary learning systems in hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev* **102**, 419-457.

McNaughton BL and Barnes CA (1990) From cooperative synaptic enhancement to associative memory: Bridging the abyss. *Seminars Neurosci* **2**, 403-416.

McNaughton BL and Morris RGM (1987) Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends Neurosci* **10**, 408-415.

Nadel L (1991) The hippocampus and space revisited. *Hippocampus* **1**, 221-229.

O'Keefe J (1990) A computational theory of the hippocampal cognitive map. *Progr Brain Res* **83**, 301-312.

O'Mara SM, Rolls ET, Berthoz A and Kesner RP (1994) Neurons responding to whole-body motion in the primate hippocampus. *J Neurosci* **14**, 6511-6523.

Ono T, Nakamura K, Nishijo H and Eifuku S (1993) Monkey hippocampal neurons related to spatial and nonspatial functions *J Neurophysiol* **70**, 1516-1529.

Rawlins JNP (1985) Associations across time: The hippocampus as a temporary memory store. *Behav Brain Sci* **8**, 479-496.

Rolls ET (1989) Functions of neuronal networks in the hippocampus and neocortex in memory. In: *Neural Models of Plasticity*, JH Byrne and WO Berry, eds, Academic Press, San Diego, 240-265.

Rolls ET, Miyashita Y, Cahusac PMB, Kesner RP, Niki H, Feigenbaum J and Bach L (1989) Hippocampal neurons in the monkey with activity related to the place in which a stimulus is shown. *J Neurosci* **9**, 1835-1845.

Rolls ET, Robertson RG and Georges-Francois P (1995) The representation of space in the primate hippocampus. *Soc Neurosci abs* **21**: 586.10, 1494.

Rolls ET and Tovee MJ (1995) Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J Neurophysiol* **73**, 713-726.

Rolls ET, Treves A, and Tovée MJ (1996) The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *submitted.*

Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. *J Neurol Neurosurg Psychiat* **20**, 11-21.

Seress L (1988) Interspecies comparison of the hippocampal formation shows increased emphasis on the regiosuperior in the Ammon's horn of the human brain. *J Hirnforsch* **29**, 335-340.

Skaggs WE and McNaughton BL (1993) Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* **271**, 1870-1873.

Skaggs WE, McNaughton BL, Gothard KM and Markus EJ (1993) An information-theoretic approach to deciphering the hippocampal code. In: *Advances in Neural Information Processing Systems* **5**, SJ Hanson, JD Cowan and CL Giles, eds, Morgan Kaufmann, San Mateo, 1030-1037.

Softky W (1994) Sub-millisecond coincidence detection in active dendritic trees. *Neuroscience* **58**, 13-41.

Squire LR (1992) Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychol Rev* **99**, 195-231.

Traub RD and Miles R (1991) *Neuronal Networks of the Hippocampus*, Cambridge Univ Press, New York.

Treves A (1993) Mean-field analysis of neuronal spike dynamics. *Network* **4**, 259-284.

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

Treves A, Barnes CA and Rolls ET (1996) Quantitative analysis of network models and of hippocampal data. In T: Ono, BL McNaughton, S Molotchnikoff, ET Rolls and H Nishijo, eds, *Perception, Memory and Emotion: Frontier in Neuroscience*, Oxford, Elsevier, in press.

Treves A, Miglino O and Parisi D (1992) Rats, nets, maps and the emergence of place cells. *Psychobiol* **20**, 1-8.

Treves A and Panzeri S (1995) The upward bias in measures of information derived from limited data samples. *Neural Comp* **7**, 399-407.

Treves A and Rolls ET (1991) What determines the capacity of autoassociative memories in the brain? *Network* **2**, 371-397.

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

Treves A and Rolls ET (1994) Computational analysis of the role of the hippocampus in memory. *Hippocampus* **4**, 374-391.

Tsodyks MV and Feigel'man MV (1988) The enhanced storage capacity in neural networks with low activity level. *Europhys Lett* **6**, 101-105.

Willshaw DJ, Buneman OP and Longuet-Higgins HC (1969) Non-holographic associative memory. *Nature* **222**, 960-962.

Wilson M and McNaughton B.L (1993) Dynamics of the hippocampal ensemble code for space. *Science* **261**, 1055-1058.