Paradigmatic working memory (attractor) cell in IT cortex [1]

by
Daniel J. Amit and Stefano Fusi,
Racah Institute of Physics, Hebrew University, Jerusalem and
INFN, Istituto di Fisica, Università di Roma, "La Sapienza",
and
Volodya Yakovlev
Department of Neurobiology, Hebrew University, Jerusalem

March 25, 1996

Abstract

We discuss paradigmatic properties of activity of single cells comprising an attractor – a developed stable delay activity distribution. To demonstrate these properties and a methodology for measuring their values, we present a detailed account of the spike activity recorded from one single cell in infero-temporal (IT) cortex of a monkey performing a delayed match-to-sample (DMS) task of visual images. In particular, we discuss and exemplify: 1. the relation between spontaneous activity and activity immediately preceding the first stimulus in each trial during a series of DMS trials; 2. the effect on the visual response (i.e. activity **during** stimulation) of stimulus degradation (moving in the space of IT afferents); 3. the behaviour of the delay activity (i.e. activity **following** visual stimulation) under stimulus degradation (attractor dynamics, and the basin of attraction); and, 4. the propagation of information between trials – the vehicle for the formation of (contextual) correlations by learning a fixed stimulus sequence, as found in ref. [6]. In the process of the discussion and demonstration, we expose effective tools for the identification and characterisation of attractor dynamics.

1 Introduction

1.1 Delay activity in experiment

A number of cortical areas, such as infero-temporal cortex (IT) and pre-frontal cortex have been suggested as part of working memory system[1, 2, 3, 4]. The phenomenon is detected in primates trained to perform a delay match-to-sample (DMS) task, or a delay eye-movement (DEM) task, using, in each case, a relatively large set of stimuli. It is observed in single unit extra-cellular recordings of spikes following, rather than during, the presentation of a sensory stimulus. In these tasks, the behaving monkey must remember the identity or location of an initial eliciting stimulus in order to

[1] A colour version of the ms is found on the web, at:
http://www.fiz.huji.ac.il/staff/acc/faculty/damita

decide upon its behavioural response following a second stimulus. This latter test stimulus is, with equal likelihood, identical to, or different from, the first stimulus in the DMS task, and is simply a "go" signal in the DEM task. In these areas of cortex neurons are observed, in rather compact regions of the same area (called "modules" or "columns"), to have reproducible elevated spike rates during the delay period, after the first, eliciting stimulus has been removed.[2] Such elevated rate distributions have been observed to persist for long times (compared to neural time constants), as long as 30 seconds. The rates observed are in the range of about 10–20 spikes/sec (s^{-1}), against a background of spontaneous activity of a few s^{-1}. The subset of neurons that sustain elevated rates, in the absence of any stimulus, is selective of the preceding, first, or sample stimulus. Thus, the distribution of delay activity could act as a neural representation or memory of the identity of the eliciting stimulus, transmitted for later processing in the absence of the stimulus.

Moreover, in one experiment in which training was carried out with eliciting stimuli presented in a fixed order, it was observed that though the stimuli were originally generated to be uncorrelated, the resulting learned delay activity distributions (DADs), corresponding to the uncorrelated stimuli, were themselves correlated [6]. That is, there was an elevated probability that the same neuron would respond with an elevated delay activity for two stimuli that were close in the sequence. In fact, the magnitude of the correlations expressed the temporal separation of the stimuli in the training sequence: the closer two stimuli were in the training sequence the higher the correlation of the corresponding two DADs. (For a detailed discussion see e.g. [8, 9, 13].)

1.2 The attractor picture

These experimental findings have been interpreted as an expression of local attractor dynamics in the cortical module: A comprehensive picture has been suggested ([8, 9, 13]) which connects the DAD to the recall of a memory into a working status and views the DAD as a collective phenomenon sustained by a synaptic matrix formed in the process of learning. The collective aspect is expressed in the mechanism by which a stimulus that had been learned previously, has left a synaptic *engram* of potentiated excitatory synapses connecting the cells driven by the stimulus. When this subset of cells is re-activated, they cooperate to maintain elevated firing rates in the selective set of neurons, via the same set of potentiated synapses, after the stimulus is removed. In this way, these cells can provide each other, i.e. each one of the group, with an afferent signal that clearly differentiates the members of the group from other cells in the same cortical module. Theoretical and simulation studies have demonstrated that this signal may be clear enough to overcome the noise generated by the spontaneous activity of all other cells, provided not too many stimuli have been encoded into the synaptic system of the module. A large number of 'correct' neurons must collaborate to sustain the pattern. A rough estimate is that about

[2] Similar phenomenon has been observed in motor cortex by Georgopoulos (private communication).

1,000–2,000 cells would collaborate in a given DAD [10], out of some 100,000 cells in a column of 1mm^2. There would be considerable overlap between DADs.

The collective nature of the DAD is related to its attractor property. Since so many cells collaborate, the DAD is robust to stimulus "errors", i.e. even if a few of the cells belonging to the self-maintaining group are absent from the group initially driven by a particular test stimulus, or if some of them are driven at the "wrong" firing rates, once the stimulus is removed, the synaptic structure will reconstruct the 'nearest' distribution of elevated activities of those it had learned to sustain. The reconstruction (the attraction) will succeed provided the deviation from the learned template is not too large. All stimuli leading to the same DAD drive activities which are in the basin of attraction of that 'attractor'. Each of the stimuli presented repeatedly during training creates an attractor with its own basin of attraction (see e.g. [11]). In addition, the same module can have all neurons at spontaneous activity levels, if the stimulus has driven too few of the neurons belonging to any of the stimuli previously learned.

1.3 Attractors and learning correlations

According to the attractor picture, attractors are established through learning process. In experiments in which monkeys performed a DMS task using as many as 100 stochastically generated visual stimuli, selective delay activities were formed for all those images which were repeatedly used in the training phase[2]. This finding supports a dynamic interpretation for the delay activity and confirm the presence of a learning process, since it is rather unlikely that synaptic structures related to these stimuli had been generated prior to training. In fact, no associated delay activities were found for new images, not used during training, but introduced subsequently in the testing stage [2]. Moreover, the correlations between DADs for sequentially appearing stimuli in the fixed training order, corresponding to uncorrelated images, are even more directly related to learning, inasmuch as these correlations are dependent on the particular fixed order in which the stimuli appeared in the particular training protocol. Learning these correlations presents a puzzle which is naturally resolved in the attractor picture. The puzzle is: How does the information contained in one image (used as first, eliciting, stimulus in one trial) propagate, in the absence of the image, until the next image in the sequence is presented, several seconds later, as first stimulus in the next trial?

Attractors provide a natural vehicle for the propagation of this information (see e.g. [12, 13, 8]), as follows: During training, there are initially no delay activity attractors. If the stimuli in the training set are uncorrelated, as DADs are first formed they are *necessarily* uncorrelated, just because there is no way to communicate information between successive stimuli. Once the uncorrelated DADs have been formed, however, these DADs themselves may be used to propagate information between trials. Recall that in one half of the trials the second stimulus is identical to the first (so as to maintain an unbiased probability of "same" and "different" trials). This second stimulus may also leave behind a "delay" activity distribution, during the inter-trial-interval, identical to the DAD excited by the first stimulus, during the inter-stimulus-interval of the trial. If this delay activity is not disturbed by the motor response of the animal (and other unmonitored visual or non-visual events in the inter-trial interval), then it may persist through the inter-trial interval and will be present in the network when the next image in the sequence is presented as first stimulus in the next following trial. This delay activity will be present in the sense that there will be a time window in which the cells of the delay activity are still at elevated rates, while the cells corresponding to the new stimulus begin to be driven. Information is then available for learning the correlation of the two (See e.g. [14].)

1.4 Experimental questions

In this study, we outline an experimental program. We show recordings, and results of their analysis, from a paradigmatic cell, i.e. an IT neuron with a clear delay activity, which we were successful in holding for a long stable recording session, sufficient for a detailed series of tests. Given the very detailed and beautiful bottom-up recordings of selective delay activities reported by Miyashita, Goldman-Rakic, Kubota and Nakamura, etc., [2, 6, 15, 4, 3] etc. on the one hand, and the detailed picture of attractor networks [8, 9, 13] on the other, one naturally asks what would be expected of a cell that participates in an attractor net. We ask this question in the present paper, and describe in detail the expected answer by demonstrating the results from this particularly rich paradigmatic neuron. Whether IT or another area in fact serves this role will only be shown by a population study. However, we believe that the type of analysis outlined here will necessarily be part of that investigation. What we are showing is an example of how a cell would perform were the attractor picture valid.

Our goal in this paper is not to present data which give a definitive answer to this question of whether IT delay activity reflects an attractor network serving stimulus image memory. For that we would have to record and analyse many cells from more than one monkey. We wish, instead, to raise the issues quantitatively, and point out, by considering in detail data from a single example, what experiments need to be carried out, and what behaviour one must expect from neurons participating in an attractor neural net.

The central issue with which we deal is the neuronal behaviour following presentation of a degraded stimulus. This is because the attractor picture naturally implies basins of attraction for every attractor. That is, it predicts that though the input to IT will be different for stimuli of different degrees of degradation, the attractor, i.e. the delay activity distribution within IT, will always be the same, as long as the degraded stimulus is not too far from the original. Thus, the attractor theory predicts very different characteristics for the response of IT neurons during stimulation (reflecting the response driven by the input to IT), and the response following stimulation (reflecting the attractor state).

Testing empirically the attractor picture raises a set of issues:

1. IT observability: Are the stimuli used in these experiments identifiable at

the level of the relevant cortical region, namely IT? i.e. Do these stimuli induce neuronal responses in the recorded area which are clearly higher than spontaneous as well higher than the delay activities? and, Are these responses different for the different stimuli?

2. Identification of significant in inter-stimulus-interval (ISI): Is the activity immediately preceding the first stimulus of a trial merely spontaneous activity, or does it reflect the identity of the final stimulus in the preceding trial? In particular in the course of a series of trials.

3. Moving in the space of IT-observable stimuli: What is the effect of stimulus degradation , e.g. using degraded or noisy stimuli, what is the effect on visual responses (during the stimulus) and on attractor dynamics, as expressed in a given cell?

4. Basin of attraction: Does motion in IT-observable space bring about an abrupt change in delay activity?

5. Information transport between consecutive trials: Can delay activity act as the agent for generation of correlations between sequentially appearing stimuli, as in the scenario of [12, 13]?

2 Methods

2.1 Preparation, stimuli and task

A rhesus monkey (Macaca mulatta) was trained to perform a Delay-Match-to-Sample (DMS) task on 30 stochastically generated images, 15 'fractals' and 15 Fourier descriptors, much like in refs. [6] and [15], the two types were randomly intermixed. Three examples of the stimuli are shown in Fig. 1 from left to right: two 'fractals' and a Fourier descriptor. The left two images, the 'fractals', are the ones we used in recording most of the data reported here (see below).

The behavioural paradigm was as follows: Following the appearance of a flickering fixation point on the monitor screen in front of the monkey, the monkey was to lower a lever and keep it in the down position for the duration of the trial (see Fig. 2). After the lever was lowered, and following a 1000ms delay, the first (eliciting, or sample) stimulus was presented on the screen for 500ms. The inter-stimulus interval (ISI) was 5 seconds. Following this delay, the second (test) stimulus was shown, also for 500ms. This second test stimulus was the same as the first sample stimulus in one half of the trials, and was different in the other half of the trials (chosen randomly from the other 14 stimuli of its type, fractal or Fourier descriptors). After the second stimulus was turned off, the monkey had to shift the bar left or right depending on whether the second stimulus was the same as the first or different, respectively, and then, when the fixation point stopped flickering, to release the bar. Correct responses were rewarded by a squirt of fruit juice. See Fig. 2. The monkey

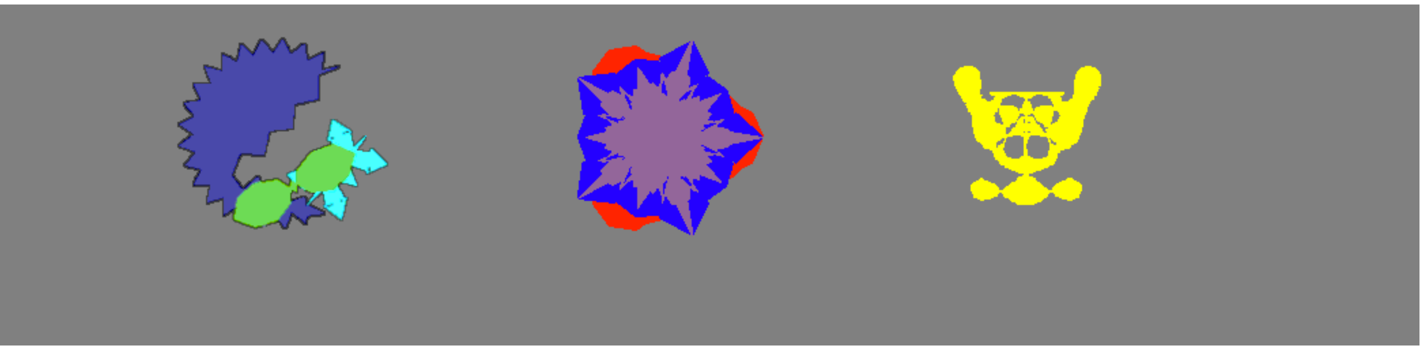


Figure 1: Three of the 30 colour images used in the delay-match-to-sample experiments. Left, 'fractal' stimulus which induced the clearest delay activity for the neuron of Fig. 3 - number 5 in the figure - and therefore called the best stimulus for this neuron; centre, 'fractal' stimulus which induced delay activity which was indistinguishable from the average pre-stimulus activity, as demonstrated in Fig. 3 - number 17 in the figure - and is therefore called the worst stimulus for this neuron; right, a Fourier descriptor type stimulus, which induced response 14 in Fig. 3.

had been trained to a performance level not less than 80% correct responses when the reported experiment took place.

2.2 Recording

We recorded extra-cellularly spike activity from single neurons of infero-temporal cortex. The experiment is still in progress so we can not give the histological verification of the recorded cell. However, the coordinates of the position of the guide tube were A14 L15 during vertical penetration into ventral cortex.

After a cell was isolated by a 6-channel spike sorter (Multi–Spike Detector; ALPHA OMEGA Engineering, Nazareth, Israel) the experiment proceeded in two stages. During the first stage, all 30 stimuli were presented a small number of times (1–3). Average PST response histograms were produced on-line, as demonstrated in Fig. 3. From these, the 'best' and 'worst' stimuli were determined by eye, using the following criteria: The 'best' stimulus was that which evoked the clearest excess of the average rate in the ISI over the average rate in the period immediately preceding the first stimulus: the pre-stimulus rate: stimulus no. 5 in Fig. 3, corresponding to Fig. 1, left. The 'worst' stimulus was selected as one that produced an overall weak response and an indistinct difference between pre-stimulus and delay period firing rate: stimulus no. 17 in Fig. 3, corresponding to Fig. 1, middle. At this stage, the pre-stimulus firing rate was interpreted naively as the average activity in a 1.0 second interval prior to the presentation of the trial's first stimulus.

For the second stage of the experiment, a new image set was created from the

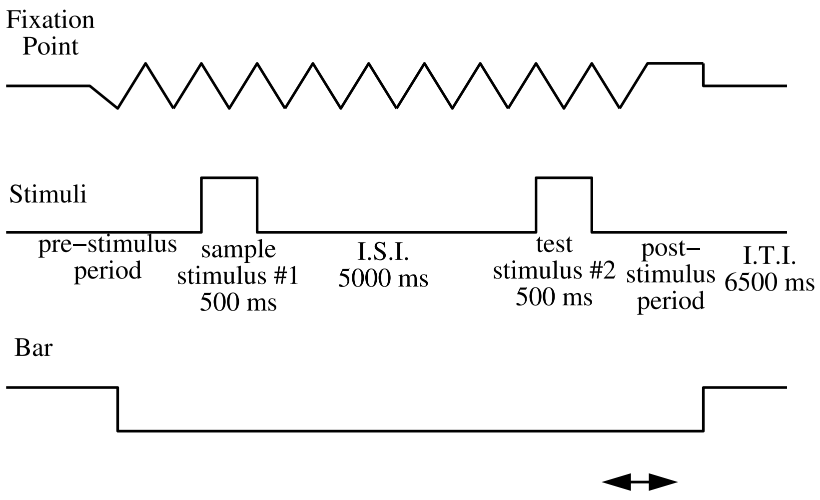


Figure 2: Schematic sequence of events in each phase of DMS task. See the text for details.

chosen 'best' and 'worst' stimuli. The RGB pixel map of each image was degraded by superimposing uniform noise at one of four levels. That is, we generated from each of these two images a set of five images: the pure, original, one (to be referred to as degradation 0) and four degraded images, each at an increasing level of degradation (1–4). The DMS task was then resumed, using as first (sample) stimulus in the DMS task a randomly selected image from the new set of 10 (original and degraded) images. The second stimulus, the test image, was randomly selected as an undegraded version of either the 'best' or the 'worst' stimulus. In Fig. 4 we reproduce the 'best' stimulus and degraded versions of this image at each of the four degradation levels used in this second stage of the experiment.

2.3 Mean rate correlation coefficient MRC

In order to test various hypotheses concerning the structure and origin of persistent activities we introduce a coefficient which measures the correlation of the trial-by-trial fluctuations of the mean firing rates – or spike counts – in two different intervals. The MRC is defined as follows: x_n and y_n ($n = 1, ..., N_{st}$ – the total number of selected trials) are, respectively, the spike counts in two intervals in the n-th trial. Then:

$$MRC = \frac{\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle}{\sqrt{(\langle x^2 \rangle - \langle x \rangle^2)(\langle y^2 \rangle - \langle y \rangle^2)}} \tag{1}$$

where $\langle ... \rangle$ denotes the estimate of the expectation of a random variable performed over all the selected trials, i.e.

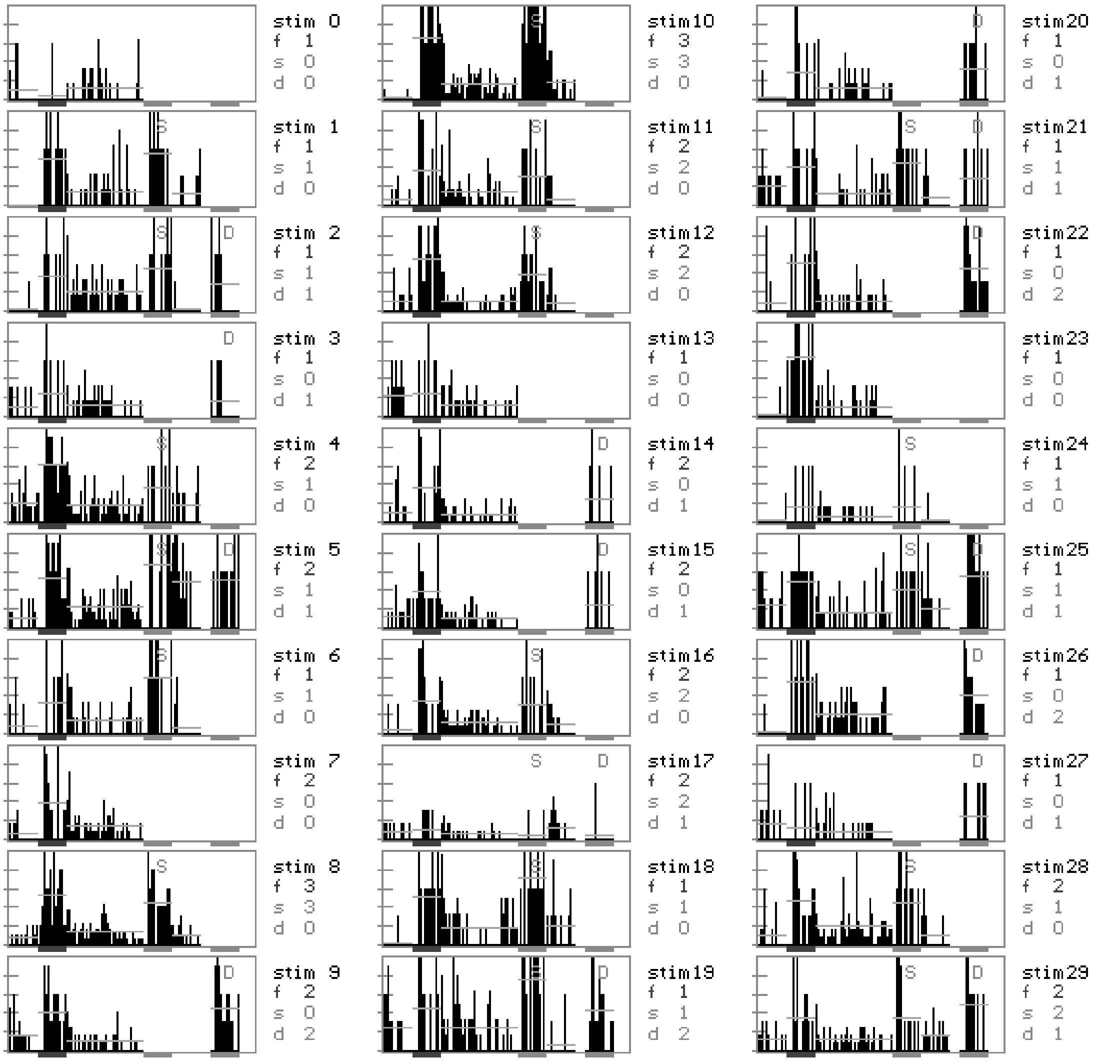


Figure 3: Spike rate histograms (full scale=50s^{-1}) for small number of trials of the entire set of 30 colour stimuli, used on-line for the selection of 'best' and 'worst' stimuli. Each window is divided into 6 temporal intervals and the neuron activity during this activity is binned and averaged for the different trials, and shown in units of spikes per second. From left: pre-stimulus interval (66ms/bin for the 1000ms just prior to the first stimulus of the trial); first, or sample, stimulus period (33ms/bin for 500ms); delay (ISI) interval (125ms/bin for 5sec); second, or test, stimulus period (showing the response for the one-half of the trials when the stimulus was the same as the first; 33ms/bin for 500ms); post-second stimulus interval (showing the activity when the second stimulus was the same as the first; 66ms/bin for 1000ms). Finally we show the response to the second stimulus of other trials, (when the first stimulus was not that of this window), but when the second stimulus was that of this window (so that the trial was a non-match trial; 33ms/bin for 500ms). Horizontal line over each interval is average rate in interval. On the right of every window: stimulus number, number of trials with this image as first stimulus (f); number of trials with the same stimulus as second stimulus in the cases of same (s) and different (d) trials. 'Best' is stimulus no. 5, the one evoking the highest excess of the average rate in the ISI over the average pre-stimulus rate; 'worst' is stimulus no. 17, the one having an overall weak response.

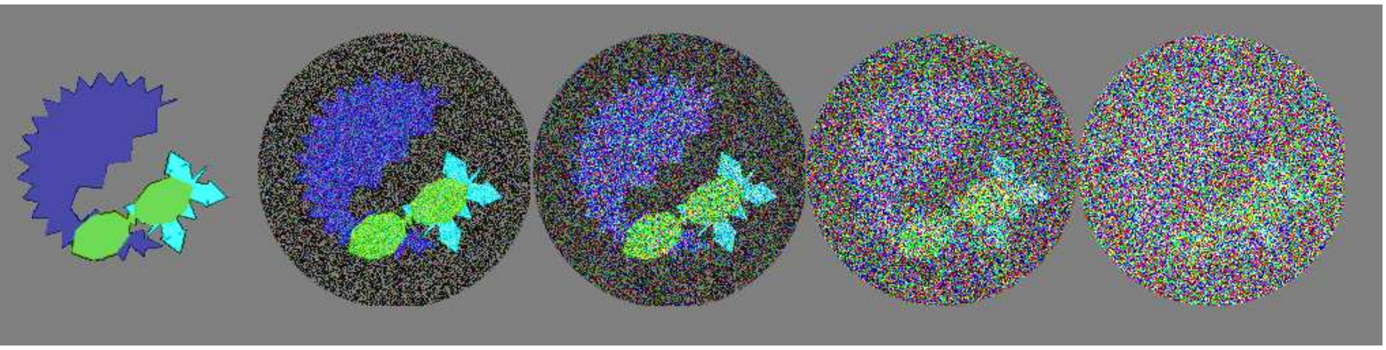


Figure 4: Image degradation (in colour) as gradual motion in stimulus space for testing attractor dynamics of delay activity. From left to right: pure 'best' image (0 noise) then 4 levels of noise on RGB map.

$$\langle x \rangle = \frac{1}{N_{st}} \sum_{i=1}^{N_{st}} x_i .$$

This coefficient, measures the extent of the correlation between the deviations from the mean over the selected set of trials, of the spike counts in each of the two intervals. It will allow us to test whether the spike counts, which vary from trial to trial, are related on a trial by trial basis across two intervals. With this tool we will monitor whether high rates during visual stimulus presentation imply high rates in the delay period and low imply low; whether high (low) counts at the beginning of a delay period imply high (low) counts in later intervals in the delay; and whether high (low) counts following a given second stimulus are related to high (low) counts in the period immediately preceding the first stimulus of the succeeding trial, its pre-stimulus interval.

The value of the MRC is in the interval [−1,1]. The standard of comparison for the magnitude of this coefficient will be established by considering sub-classes of trials in which the average rates in both intervals are both high, or both low. For instance the MRC of the subset of responses to the worst stimulus is calculated by performing the expectation of equation 1 over the trials in which the worst stimulus was presented as sample. In this case, if the fluctuations about the mean spike count are random and hence uncorrelated, the MRC coefficients will set the scale for lack of correlation. As we shall see below, large values of MRC can capture two different phenomena: (1) Trial by trial correlation of the fluctuation of the individual counts about their mean; (2) Existence of subsets of trials with very different mean counts in each, while within each subset, fluctuations about the mean may even be uncorrelated.

3 Experimental testing and the special cell

The issues raised in Section 1.4 will now be discussed in the same order in terms of the responses of a single(!) IT cell. The intention is not to claim that the data presented here give a definite answer to these questions. That would require the recording and analysis of many cells and, traditionally, in more than one monkey. The objective, rather, is to raise the questions, the problems and the potential answers, by considering in detail the record of a particularly rich cell that exhibits behaviour in relation to which we may pose most of the problems and demonstrate the form answers may take. Following continued recording from numerous neurons in this, and another monkey, we will be able to report whether this area of IT may in fact serve this memory function for DMS tasks. Alternatively, we may report that too few IT neurons have the expected significant DAD activity to be able to serve this function under the conditions used in our experiment, and it would be best to look for such DADs elsewhere. Finally, we may, of course, find that delay activity is present in IT but does not have generally the properties which we expect (and which are reported here for this single cell), so that our understanding of cortical dynamics should best be revised.

3.1 Identification of the stimulus

Delay activity distributions are likely to be sustained by the synaptic structure, learned or pre-existing. The fact that the delay activities are selective to the preceding stimulus implies that the synaptic structure can sustain a variety of distributions in the absence of a stimulus. Which of the stable delay activities is actually propagating depends on which was the last effective stimulus. But different stimuli presented on the monitor screen may not actually lead to different neuronal responses in deep structures (higher cortical areas) such as IT cortex, since the differences may well be filtered out at afferent levels. It is the representation of the stimulus as it affects the cells of IT that is relevant for which of the DADs is excited, or learned. Thus, for example, it may be that differences of scale or of color in the external image may result in very similar distributions of afferents when arriving at IT cortex. In that case they cannot elicit different DADs. In other words, stimuli and stimulus differences must be IT-observable.

What can give rise to different DADs are significantly different distributions of synaptic inputs that arrive from previous cortical areas to IT. These distributions, in turn, can be observed by recording the neuronal activity in the attractor network during presentation of the stimulus. In fact, IT cortex is particularly propitious in this respect because rates observed during the presence of the stimulus in visually responsive cells that participate in a DAD, are much higher than those observed for the same cells during inter-stimulus intervals of the trial (See e.g. [2] and responses in the underlined intervals in the histograms for 'best' stimulus in Fig. 5). This allows a very clear identification of the stimulus, for comparison with the attractor. In other cortical areas, e.g. pre-frontal cortex, this may not be as clear (see e.g.

[4, 5]).

The range of responses during the stimulus and during the ISI, as in Figs. 3 and 5, exhibits IT-observability of the stimuli as well as of their differences. Thus, the stimuli used for this experiment, the 30 fractals and Fourier descriptors are IT observable in that the responses to some of them are much higher than the background spontaneous rates, as well the rates during the delay period ISI. This difference from background will be measured quantitatively below. In addition, it is clear that our neuron also differentiates between stimuli, e.g. between the best and the worst stimuli, both during and following the stimulation.

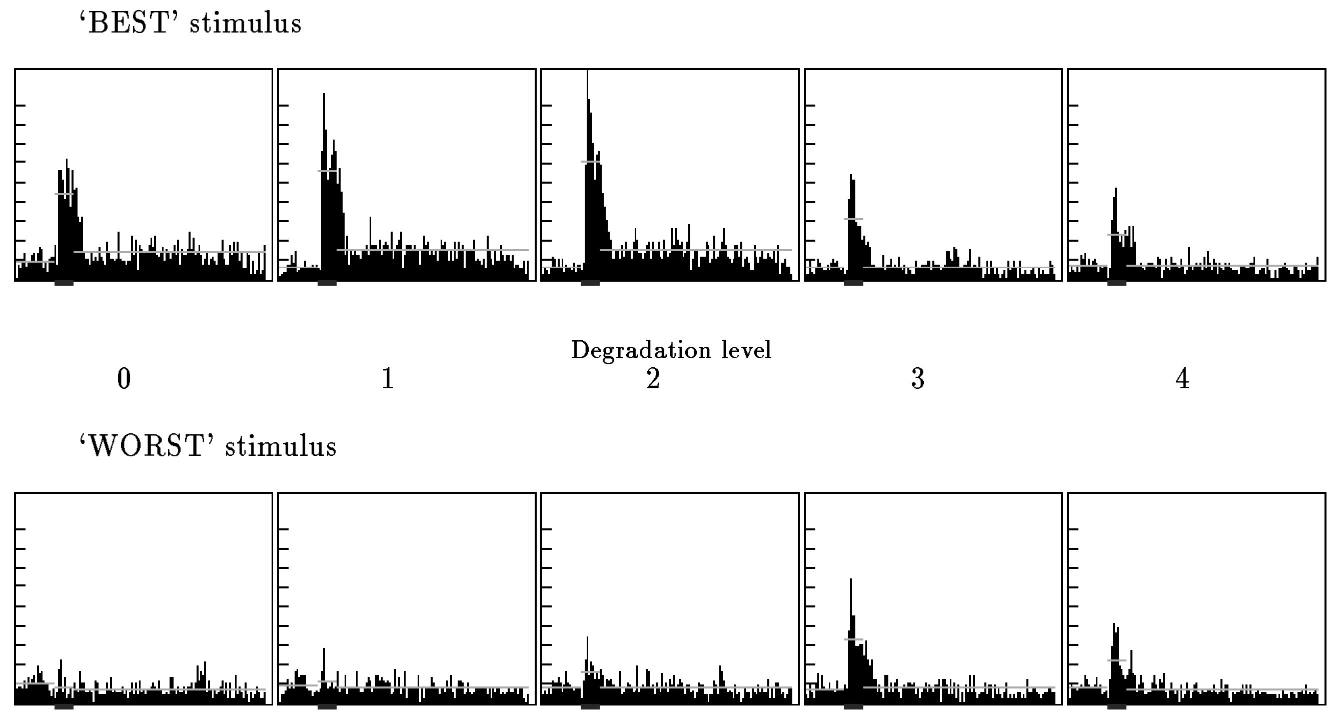
3.2 Delay activity and spontaneous activity

It is quite common to consider the ongoing activity before the first stimulus (the pre-stimulus activity) in a given cell as spontaneous activity. Delay activity is considered significant if its rate is significantly higher than the pre-stimulus rate. In this way, one observes in Fig. 5, in the first three windows on the left of the best stimulus row, that the average rate (the horizontal line) during the delay (the last, wide, interval) is higher than in the pre-stimulus interval of the current trial (see the numbers in the table below the histograms). They become equal in the two windows on the right (see below, Discussion). In the leftmost window of the 'best' row of histograms, corresponding to the undegraded best stimulus, the ratio ISI/pre-stim rates is 13/8, which may not be considered very convincing as a signal to noise ratio.

But in a context in which delay activities exist, pre-stimulus rates are not necessarily spontaneous. For example, if the second stimulus in a trial is a stimulus that evokes a DAD when presented as first stimulus, and if the neuron being observed has an elevated rate in this DAD, then one would expect that also following the second stimulus this neuron would have an elevated rate. In fact, this is observed in Fig. 6, where we plot histograms of spike rates that follow presentation and removal of the 'best' (top) and the 'worst' (bottom) stimuli as *second* stimulus. These histograms are averages over all trials at all levels of first stimulus degradation, since the second stimulus is always undegraded. What one observes is that following the 'best' second stimulus, the level of post-second-stimulus persistent rate is as high as the delay activity for the first three levels (0–2) of degradation in Fig. 5 (i.e. no degradation and the next two levels of degradation). This is true even when the first stimulus, due to its degradation, does not provoke significant delay activity. This is nicely consistent with the fact that this persistent activity is provoked by the second stimulus that is never degraded.

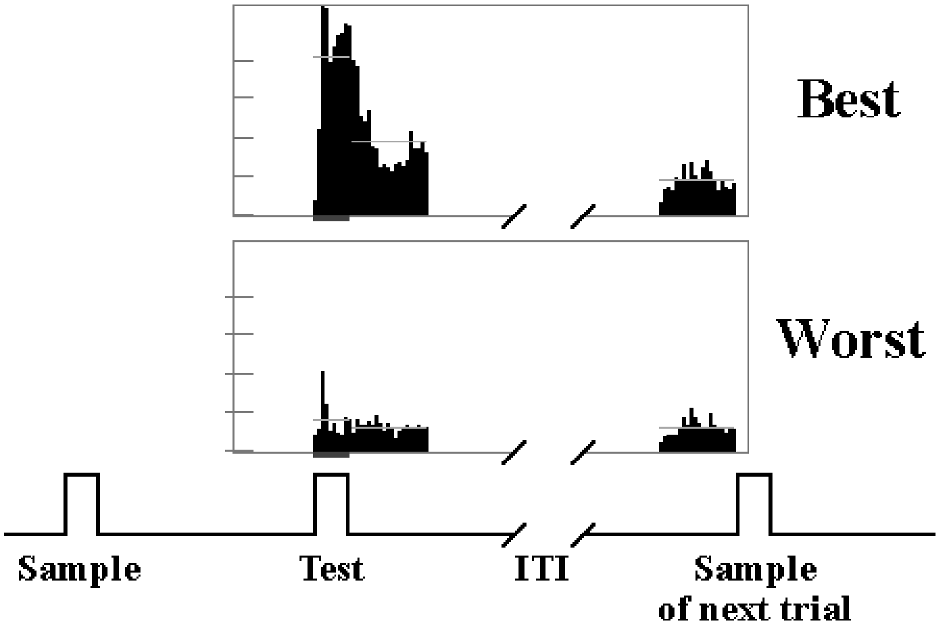
Another expression of the same fact is presented in Fig. 7. These are the trial-by-trial distributions of the post-second-stimulus average rate for 'best' and 'worst' stimuli, (plotted above and below the axis, respectively, though both are positive valued, of course).[3]

[3] In calculating the average post-stimulus rate we take an interval starting 250ms after the second stimulus is turned off, to exclude latency of response to the stimulus. See e.g. Section 3.5.



Stim	Trial no	Prestim	Stim1	ISI1	ISI2	ISI3	ISI4	ISI5	total ISI
B0	12	8.6± 1.0	46.3± 3.0	17.3± 1.9	12.3± 1.3	13.8± 0.7	12.3± 1.6	9.6± 1.9	13.1 ± 0.8
B1	12	5.7± 0.9	60.8± 5.4	20.9± 2.3	15.8± 1.1	14.4± 1.3	12.6± 1.6	8.7± 1.3	14.5 ± 0.9
B2	12	5.3± 1.2	67.8± 4.8	21.6± 1.6	15.6± 1.3	13.3± 1.9	12.0± 1.0	9.3± 1.3	14.3 ± 0.8
B3	12	5.5± 0.5	33.2± 2.7	8.8± 1.4	5.5± 0.7	8.5± 2.0	4.2± 0.6	3.0± 0.8	6.0 ± 0.6
B4	13	6.7± 0.9	25.0± 1.6	9.9± 0.8	6.5± 0.8	5.8± 0.9	4.2± 0.6	4.6± 0.7	6.2 ± 0.3
W0	13	9.6± 1.1	8.0± 1.2	6.7± 0.8	5.5± 0.7	6.5± 0.8	8.9± 1.4	4.3± 0.6	6.4 ± 0.4
W1	13	8.1± 1.1	10.6± 1.3	8.0± 0.8	9.1± 1.0	7.4± 0.5	6.2± 1.1	4.7± 1.1	7.1 ± 0.5
W2	12	7.5± 1.0	17.2± 1.9	9.2± 1.0	8.8± 1.0	6.2± 0.9	6.1± 1.0	5.7± 0.6	7.2 ± 0.5
W3	12	7.0± 1.2	34.8± 2.6	11.8± 1.3	7.3± 1.3	6.3± 1.0	4.8± 0.9	4.9± 0.8	7.0 ± 0.5
W4	12	7.4± 0.9	23.7± 2.9	10.7± 0.9	5.5± 0.8	5.5± 0.7	4.4± 1.0	4.3± 0.9	6.1 ± 0.5

Figure 5: Testing attractor dynamics: Spike rate histograms (rate scale $10\mathrm{s}^{-1}$, bins 50ms) 'best' (top row) and 'worst' (second row) first stimulus, each original and four levels of degradation. Each histogram is divided into three intervals. From left: pre-stimulus of the current trial, first stimulus, delay interval. Horizontal line over each interval is average rate in interval. The value of the average rates for each interval is reported in the table below the histograms with the mean rates and the standard errors (in s^{-1}). The delay interval (ISI) has been divided in 5 parts of 1s each. The mean rate of the activity during the entire ISI is in the last column on the right. For both 'best' and 'worst' first stimulus one observes variation of the rate during stimulation (the visual response) as function of the level of degradation. For the first three windows in 'best' row there is significant invariant delay activity. In the last two levels of degradation the delay activity drops and becomes indistinguishable from the delay rate for the 'worst' stimuli. The rate in the first second of the ISI, following a strong visual response is significantly higher than the mean rate of the delay interval. This is due to the latency of the external stimulus. See e.g. Section 3.5



Stim	Trial no	Stim2	Post-stim	Pre-stim (Next trial)
B0	62	44.7± 1.3	18.5± 0.8	8.6± 0.5
W0	61	8.2± 0.5	5.7± 0.4	5.7± 0.4

Figure 6: Testing information transmission across trials: spike rate histograms for 'best' and 'worst' second stimulus and the corresponding table of mean rates. (as in Fig. 5). The histogram windows are divided in 4 intervals: second stimulus (underlined), post-second stimulus, central part of inter-trial interval (invisible) and the pre-stimulus of the successive trial. Note that pre-stimulus activity of the trial following 'best' second stimulus is significantly higher than that following the 'worst' second stimulus. It carries the information about the second stimulus of the previous trial (see the text and Fig. 7).

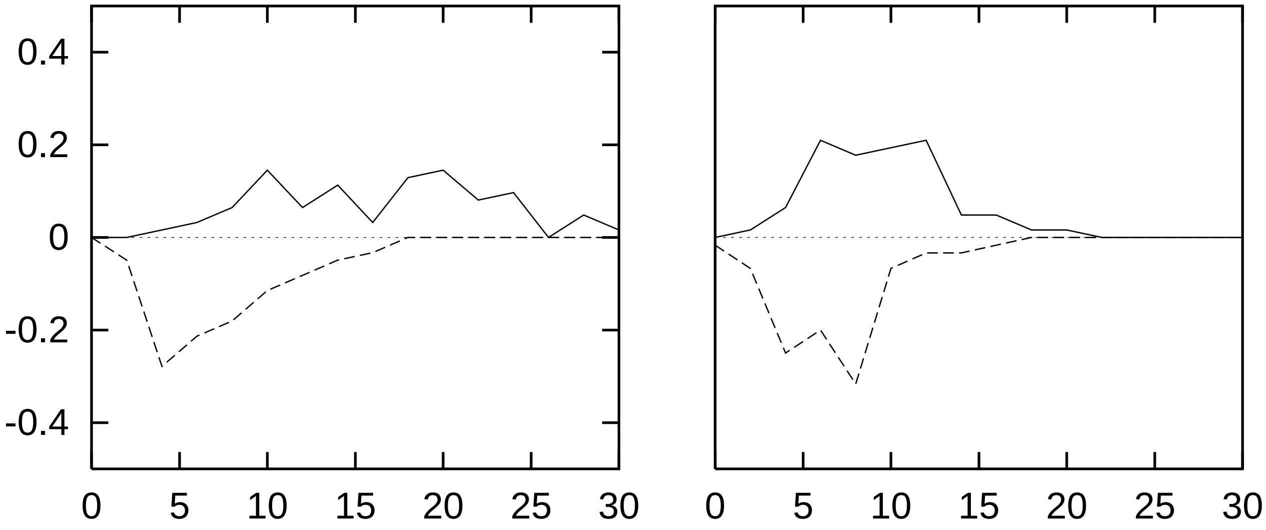


Figure 7: Distributions of the post-second-stimulus rates (left) in s^{-1}(x-axis) and the rates of the pre-stimulus of the next trial (right) for 'best' (continuous line) and 'worst' (dashed line). The two distributions differ significantly also in the second case, carrying the information about the second stimulus of the trial. The post stimulus interval is on average 710ms, following the 250ms expected latency effect. See the discussion in the text.

The distributions differ significantly according to the T-test: $P(\mathrm{Tst}) < 10^{-5}$. If this elevated rate were to continue beyond the monkey's behavioural response, it would arrive at the following trial as an elevated rate and appear as an elevated pre-stimulus activity before the next trial. There are previously reported indications that delay activity does in fact survive motor reaction events[3], though here there is a novelty in that the post reaction activity is not motivated by the behavioral paradigm. As a test of this possibility we separated the average pre-stimulus rates over the entire set of trials (N=123 trials) into two sets, according to the preceding second stimulus (see Fig. 6). It turns out that the average pre-stimulus rate in the group of trials following trials in which the second stimulus was the 'best' stimulus (N=62) is 8.6s^{-1}while for those following trials where the second stimulus was the 'worst' stimulus, (N=61) the average pre-stimulus rate is 5.7s^{-1}. (This difference is significant according to the T-test: $P(\mathrm{Tst}) < 10^{-5}$).

Yet another test of this effect will be discussed below in Section 3.6.

A simple tool suggests itself for improved on-line monitoring of the relevant pre-stimulus activity: to reduce (though not fully eliminate) the effect of persistent post-second-stimulus high rate, one should consider the average pre-stimulus activity over all trials including all stimuli. This will help, provided the number of DADs in which the cell under observation participates is relatively low. That would be the typical case.[4] (See also ref. [2]).

[4]We are indebted for this observation to Dr. Nicolas Brunel.

3.3 Attractor dynamics

Attractor dynamics (associative memory) as a description of persistent DADs implies that each DAD is excited by a whole class of stimuli. Each of these stimuli raises the rates of a sufficient number of cells belonging to the DAD so that via the learned collateral synaptic matrix they can excite the entire subset of neurons characterising the attractor and maintain them with elevated rates when the stimulus is removed. The class of stimuli leading to the same persistent distribution is the basin of attraction of that particular attractor. As one moves in the space of stimuli, at some point the boundary of the basin of attraction of the attractor is reached. Moving even slightly beyond this boundary, the network will relax either into another learned attractor, or to the grand attractor of spontaneous activity. See e.g. [16, 8].

To test the validity of the attractor picture one has to be able:

(a) to move in the space of stimuli by steps which are not too big, so that there is a choice between staying inside the basin and moving out. Recall that moving in this space does not mean only changing the stimuli, but changing them in a way that is IT-observable;

(b) to find IT-observable changes in stimuli and have the delay activity unchanged;

(c) to arrive at IT-observable changes in stimuli which will go over the 'edge', i.e. will not evoke the same DAD.

A convincing demonstration of these phenomena would again require recording a large number of cells, especially since all cells with high rates in the same DAD (corresponding to the same stimulus) are predicted to loose their rates together. The prototypical behaviour of such cells expresses itself on our single cell. First, Fig. 5 demonstrates the required motion in stimulus space at the level of IT cortex: In the five windows for the best stimulus as well as in those for the worst, going from left to right, corresponding to increasing the degradation level as presented in Fig. 4, there is a clear variation in the visual response – the neuron's spike rate during stimulation. This implies that the choice of the degradation mode is an IT-observable motion in stimulus space, as required in (a).

Second, in the top three windows in the 'best' stimulus row, the delay activity is essentially invariant, as would have been expected if the stimulus were moving within the basin of attraction of that DAD. This corresponds to (b), above.

Third, as the level of degradation increases to reach the rightmost two windows for the best stimulus, the delay activity disappears (Fig. 5). The average rate in the delay period becomes that of the reduced pre-stimulus activity discussed above, in accordance with the expectation (c) (see table in Fig. 5). This rate is also the same as that during the delay in all five windows corresponding to the 'worst' stimulus.

The discussion of point (a) calls for a comment concerning the fact that as the best image is initially degraded, the visual response increases. Only in the fourth and fifth levels of degradation (3,4) does the visual response decrease. This may seem contradictory, in that one might naively expect that the visual response should be maximal for the pure image and decrease monotonically with degradation. But

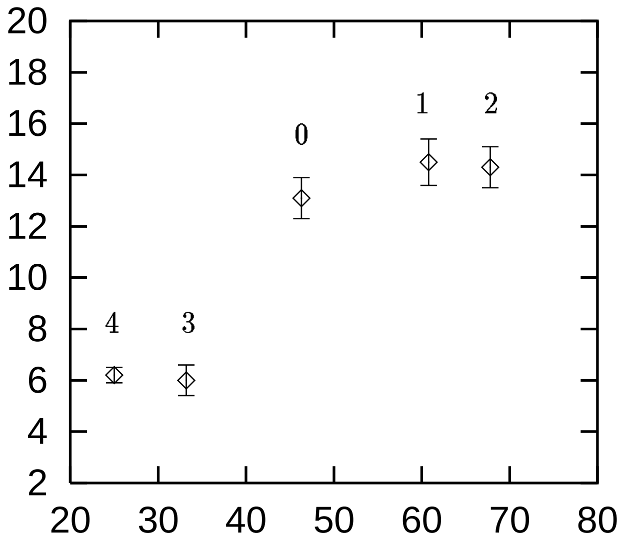


Figure 8: Average delay activity rate (in s^{-1}, y-axis) vs average stimulus response (in s^{-1}, x-axis) for five levels of degradation of 'best' stimulus. Degradation level indicated above each data point (0 for pure image). Error bars are standard errors. Note that stimulus response difference between 1 and 0, where delay activity is essentially constant, is larger than between 0 and 3, where delay activity collapses. Recall that 0 implies pure stimulus.

on second thought it should be clear that there is no reason why a particular cell should be maximally tuned to the image rather than to a degraded version of it. In fact, looking at other cells, we find that while there is not a strong systematic trend for the visual response as a function of the degradation level, on the average the visual response decreases with degradation. Nor is there any argument why there should be[17]. What is essential is that there be a set of neurons that have a strong visual response and a corresponding significant elevated delay rate, and when the images change moderately, the delay activity remain invariant and when they change enough, the delay activity change abruptly.

Our paradigmatic cell would have belonged to such a class had it existed.

This cell gives yet two more, strong signals of the attractor property. Looking at the visual response to the degraded 'best' stimulus in decreasing order of rates suggested above: 2–1–0–3–4, one observes in the table in Fig. 5 and in Fig. 8 that the rate difference from 1–0 is $14.5s^{-1}$and no noticeable change in delay activity. Yet going from 0–3 the visual response changes by $13.1s^{-1}$and the delay activity collapses. In other words, the decrease in the delay activity seems precipitous, as a function of the visual response, as would befit a watershed at the edge of the basin of attraction.

The second signal is related to the fact that in the images used for testing the attractor property (Fig. 4), as the degradation noise is introduced, a circle appears around the figure (for technical reasons). This might have been perceived by the monkey as a different stimulus. That is, as the figure gets increasingly blurred, it is the circle that is most visible. Yet, the delay activity for the first three degradation levels (0–2) of the 'best' stimulus remains unchanged with the appearance of the circle. The delay activity disappears when the degradation level crosses the critical level, despite the continued presence of the circle. Moreover, the circle produces no effect for the 'worst' stimulus, despite the fact that it is visible there as well. This may be interpreted as due to the fact that the degradation circle appears only during testing and is not seen often enough to have been learned.

3.4 Delay interval MRC

To test the attractor interpretation more closely we have calculated the MRC (see Section 2.3) between the average rates in the 1-second immediately following the first stimulus and the 1-second immediately preceding the second stimulus. The results are summarised in Table 1. In a subset of trials with 'best' and 'worst' first stimuli and the first three degradation levels (0–2) we find MRC=0.399. Note that in this subset of 74 trials there are 36 trials with elevated delay activity and 38 with low delay activity. In other words, the high MRC may be attributed to the presence of two underlying sub-populations with different means, as mentioned at the end of Section 2.3. Similarly, in the subset of trials with 'best' first stimuli and all degradation levels (second row in the table) there are 61 trials, of which 36 have elevated delay activity and 25 do not, and MRC=0.382.

The 61 trials with 'best' first stimulus are then divided into two subsets: one for the first three degradation levels (N_{st}=36) and one for the last two (N_{st}=25). In the first subset there is elevated delay activity while there is none in the second. For these two subsets we find MRC=0.080 and –0.123, respectively. One concludes that the average delay rates in the first subset of trials remain high throughout the ISI and low in the second set. The deviations from the mean in each subset, at the beginning and the end of the ISI are *uncorrelated*. This constitutes an example of the first option described at the end of Section 2.3 and provides scales for high and low correlations.

3.5 Single neurons vs attractors dynamics

An alternative scenario to the attractor picture may attribute the enhanced delay activity to some change in the internal state of single cells. Such a scenario would imply that the change in the internal state of the cell is triggered by the level of the visual response. This would be required to make the persistent delay activity stimulus selective. It would further imply a correlation between the rate during the ISI and the rate during the visual response. Table 2 summarises the MRCs between the visual response to the first stimulus and various 1-second intervals of the ISI. We find that the rate during the visual response to the first stimulus is correlated with the rate in the first 1-second interval of the ISI (MRC=0.407), even in a subset of trials with stimuli all leading to elevated delay activity, i.e. the 36 trials corresponding to 'best' stimulus at degradation levels 0–2. This high MRC in a trial set with a narrow

First Stimulus selection	Degradation levels	time intervals	number of trials	MRC
B+W	0-1-2	ISI1-ISI5	74	0.399
B	0-1-2-3-4	ISI1-ISI5	61	0.382
B	0-1-2	ISI1-ISI5	36	0.080
B	3-4	ISI1-ISI5	25	-0.123
W	0-1-2	ISI1-ISI5	38	0.089

Table 1: Correlations of delay activity in subintervals of ISI. First two columns define the subset of trials over which the correlation coefficient is computed; the third gives the the two time sub-intervals; the fourth is the number of trials in the subset followed by the MRC. In the first row, for example, the subset contains all the trials in which the first stimulus is 'best' or 'worst' with the first three degradation levels (0-1-2). The intervals are the first 1-second following the removal of the stimulus and the last 1-second of the delay interval. See the text for a discussion.

First Stimulus selection	Degradation levels	time intervals	number of trials	MRC
B	0-1-2	ST1-ISI1	36	0.407
B+W	0-1-2-3-4	ST1-ISI1	123	0.784
B	0-1-2	ST1-ISI4	36	0.085
B	0-1-2	ST1-ISI5	36	-0.098

Table 2: Single neurons vs attractor dynamics: Correlations between visual response and several 1-second intervals of the ISI. First two rows, high MRCs due to latency of visual response. No correlation in later intervals. Conventions as in Table 1.

distribution of rates indicates that the trial-by-trial fluctuations of the rate in the first 1-second of the ISI are correlated with the fluctuations in the rate during the visual response. In fact, if the effect of splitting the means is added, by computing the MRC of the visual response rates with the rates in the first 1-second of the ISI for all trials (N_{st}=123), we find MRC=0.784. These high correlations we associate with the latency of the visual response. In fact, it appears from Fig. 5 that the visual response penetrates about 250ms into the ISI.

Next we take the same set of 36 trials ('best' with degradation 0–2) which gave MRC=0.407 between the rate in the visual response and the rate of the first 1-second of the ISI. The MRCs between the rate of the visual response and the rate in the fourth and fifth 1-seconds of the ISI are 0.085 and –0.098, respectively (third and fourth rows in the table). We conclude that beyond the first second of the ISI the fluctuations of the visual response are not correlated with the fluctuations of the delay activity. This makes the single cell scenario rather implausible.

3.6 Information transmission across trials

The discussion of the meaning of pre-stimulus activity, Section 3.2, provides a partial confirmation of the scenario of information transport between first stimuli in consecutive trials, essential for the formation of the Miyashita[6] correlations, see e.g. Section 1.3. That scenario requires that information about the coding of one stimulus be able to traverse the time interval between consecutive trials. The scenario proposed ([12, 13, 8]) is that DADs when they first become stable, are uncorrelated. When they are stable, they would be provoked as much by a (same) second stimulus in a trial as by the first one. Since the second stimulus is in half of the trials the same as the first (to prevent bias in the DMS paradigm), in half of the training trials the activity in the inter-trial interval will be the same as the delay activity elicited by the first stimulus of the preceding trial. If training is done by a fixed sequence of *first stimuli*, then upon half of the presentations of a given **first** stimulus the activity distribution in the IT module, in the inter-trial interval, would be the same as the delay activity corresponding to the immediately preceding **first** stimulus. The joint presence of the delay activity of the previous stimulus and of the activity stimulated by the current first stimulus is hypothesised to be the information source for the learning, into the synaptic structure, of the correlations between the delay activities corresponding to consecutive images.

We calculated the MRCs between the average rate in the 1-second immediately following the second stimulus and the average rate in an interval of 1-second immediately preceding the successive first stimulus (Table 3). The idea being that, since (for technical reasons) we do not have recordings during the entire interval separating two trials, if the activity between trials were to be undisturbed by the monkey's response, to arrive at the next first stimulus, the MRC of the activities in the two extreme sub-intervals between trials should be similar to that between the two extreme sub-intervals in the undisturbed ISI. In fact, MRC=0.320 for the inter-trial interval performed over all the available trials (N_{st} =123). The entire set of trials is composed of two well balanced subsets with two different inter-trial mean rates. This is so because the second stimulus is never degraded, so for N_{st}=62 the second stimulus is a pure 'best' image, leading to high inter-trial persistent activity, and for N_{st}=61 of the trials the preceding second stimulus is the pure 'worst' image, leading to low inter-trial activity.

This value is compared with the MRCs of extreme ISI intervals in trial subsets of similar statistics, i.e. with balanced subsets of low and high averages. Those would correspond to the first and second row in Table 1. In fact, the numbers are close. If the 123 trials are separated into two subsets: one with preceding 'best' and one with preceding 'worst', we find MRC=0.038 and MRC=0.120, respectively (rows 2 and 3 in Table 3). These would be naturally compared with MRCs of sub-intervals of trials with 'best' and 'worst' first stimulus, with degradation levels (0–2) (third and fifth row in Table 1), that are respectively: 0.080 (N_{st}=36) and 0.089 (N_{st}=38).

Second Stimulus of previous trials	time intervals	number of trials	MRC
B+W	POST and PRE	123	0.320
B	POST and PRE	62	0.038
W	POST and PRE	61	0.120

Table 3: Information transmission across trials: MRCs of 1-second post second stimulus and 1-second pre-next first stimulus. For a few trials the time interval after the second stimulus is less than 1 second: for those trials the mean rate is calculated over the available time interval. First row: high MRC due to presence of two different means. Last two rows, absence of correlation in unimodal subsets of trials. Conventions as in Table 1.

4 Discussion

The cell we have discussed presents a rich phenomenology, naturally interpreted in the learning-attractor paradigm. As was repeatedly stated, to establish the various characteristics of attractor dynamics and of their internal structure many more good recorded cells are required. So the underlying intention had been to exploit this cell merely as an example of the kind of features that a cell participating in a presumed attractor scenario may have. Looking back at the wealth of data drawn from this cell, we feel one can say more. This collection of data can hardly be accounted for in other paradigms suggested to date.

Most prominent of these is the propagation of the persistent elevated rates following the test stimulus, after this matching stimulus is turned off, and even after the behavioural reaction is completed. The amount of data collected for this cell and the tools we have sharpened to analyse them, leave no doubt that this propagation can indeed take place. In fact we have found that this persistent activity is as robust when it crosses the inter-trial interval (including the behavioural response) as it is crossing the ISI delay interval between sample and match. What is surprising about it is that following the second stimulus and the reaction, there is no apparent need for maintaining the active memory. One might have expected that at this point the persistent activity distribution will die down. It does not. This has been observed also in ref. [3], but the novelty here is that the delay activity continues to persist even when there seems to be no functional role for it. What seriously modifies the delay activity is a new visual stimulus, either the second stimulus following the ISI, or the first in another trial, much as in Miller et al[18].

The fact that a DAD can get across the inter-trial interval is an essential pillar in the construction of the dynamic learning scenario for the generation of the Miyashita [6] correlations in the internal representations (working memory) of stimuli often experienced in contiguity. Our single cell result suggests that such correlations may find their origin in the arrival of selective persistent activity related to one stimulus at the presentation of the consecutive stimulus in the next trial.

The collective, attractor picture would have been more complete had we had a case in which there is no visual response and yet elevated delay activity. In ref. [15] such cases are shown. Yet, the mean rate correlation analysis on this cell shows that neural activity during the delay period is correlated with the visual response only in the immediate interval following the stimulus. This is very strong evidence that what determines the actual spike rate later into the ISI is beyond the single cell. It is most likely the result of the interaction of this cell with other cells which participate in the same DAD, via potentiated synapses.

Finally, there is an additional, behavioural, correlate to the activity of this cell. When the performance level (% correct responses) of the monkey is considered vs the level of stimulus degradation, averaged across experiments, it is found that the performance is similar for the first three levels of degradation. Then, for the fourth and fifth levels, performance drops abruptly to chance level, just as our cell indicates that the stimuli are not now in the basin of attraction of the DAD [17]. This effect calls, of course, for much further study. But it, too, opens new vistas toward identification of a potential functional role of the delay activity. Sakai and Miyashita [15] turned to the pair associate paradigm, because the monkeys perform the DMS task well even for stimuli that do not generate delay activity, 'new stimuli'. Here we see that in the absence of delay activity the DMS task was not performed successfully if the sample stimulus was very degraded.

Our paradigmatic cell suggests that the mechanism which allows matching to an identical new sample may not work when the monkey has to match to a very degraded image. Yet, as long as the DAD (the attractor) is effective, it corrects for the degradation, by attracting to the prototype delay activity, and the task can be performed.

5 Acknowledgement

Without the experimental and intellectual contribution of Drs. Shaul Hochstein and Ehud Zohari this paper would not have been possible. Had we had it our way, both would have figured among the authors. We have benefited from the contribution of Gil Rabinovici to the experiment. This cell was found during his stay in our lab as a summer project in his course of pre-medical studies at Stanford University. We are also indebted to Profs. J. Maunsel and R. Shapley for help in the early stages of this experiment and the detailed comments of Dr. Misha Tsodyks. This work was partly supported by grants from the Israel Ministry of Science and the Arts and the Israel National Institute of Psychobiology and by Human Capital and Mobility grant ERB-CHRX-CT93-0245 of the EEC.

References

[1] Fuster JM 1995 *Memory in the cerebral cortex* (MIT press)

[2] Miyashita Y and Chang HS 1988 Neuronal correlate of pictorial short-term memory in the primate temporal cortex, *Nature*, **331** 68

[3] Nakamura K and Kubota K 1995 Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task, *J. Neurophysiol.* **74** 162

[4] Wilson FAW, Scalaidhe SPO and Goldman-Rakic PS 1993 Dissociation of Object and Spatial Processing Domains in Primate Prefrontal Cortex, *Science*, **260** 1955

[5] Williams GV and Goldman-Rakic PS 1995, Modulation of memory fields by dopamine D1 receptors in pre-frontal cortex, *Nature* **376** 572

[6] Miyashita Y 1988 Neuronal correlate of visual associative long-term memory in the primate temporal cortex, *Nature* **335** 817

[7] Miyashita, Y. Higuchi, S., Sakai, K. and Masui, N. (1991) Generation of fractal patterns for probing visual memory. *Neurosci Res.* **12** 307

[8] Amit DJ 1995 The Hebbian paradigm reintegrated: local reverberations as internal representations, *Behavioural and Brain Science* December

[9] D.J. Amit 1994 Persistent delay activity in cortex: a Galilean phase in neurophysiology? *NETWORK* **5** 429

[10] Brunel N 1994 Dynamics of an attractor neural network converting temporal into spatial correlations, *Network*, **5** 449

[11] Amit DJ and Brunel N 1995 Learning internal representations in an attractor neural network with analogue neurons, *Network* **6** 39

[12] Griniasty M, Tsodyks MV and Amit DJ 1993 Conversion of temporal correlations between stimuli to Spatial correlations between attractors, *Neural Computation* **5** 1

[13] Amit DJ, Brunel N and Tsodyks MV 1994 Correlations of cortical Hebbian reverberations: experiment *vs* theory, *J. Neurosci.* **14** 6435

[14] Brunel N, 1995 Hebbian learning of context in recurrent neural networks, Neural Computation, to be published

[15] Sakai K and Miyashita Y 1991 Neural organisation for the long-term memory of paired associates, *Nature*, **354** 152.

[16] Amit DJ and Brunel N 1995 Global spontaneous activity and local structured (learned) delay activity in cortex, submitted to *J. Neurophysiol.*

[17] Yakovlev V, Fusi S, Amit DJ, Hochstein S, Zohary E 1995, An experimental test of attractor network behavior in IT of the performing monkey, 4th Annual Meeting of the Israel Society for Neuroscience, Eilat, *Israel J. of Medical Sciences* **31** 765

[18] Miller E.K., Li L. and Desimone R 1993, Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J. Neurosci* **13** 1460