Computational examination on the dynamics of recall activity in the inferior temporal cortex

Atsuo SUEMITSU* Shigemitsu MOROKAMI** Kazuhiko MURATA** Masahiko MORITA***

*Doctoral Program in Engineering, University of Tsukuba
**Doctoral Program in System Information Engineering, University of Tsukuba
***Institute of Engineering Mechanics and Systems, University of Tsukuba

1-1-1 Ten-nodai, Tsukuba, Ibaraki 305-8573, Japan

Abstract - In the inferior temporal cortex of the monkey, 'pair-recall' neurons which exhibit prospective activity for the target during the delay period of a pair-association task have been found. To explain this and other physiological findings, we previously constructed a model of pair-association memory consisting of two interactive networks. The present paper reports that recent empirical data on the time course of the pair-recall activity accord very well with the prediction of our model. This strongly suggests that trajectory attractors are formed in area TE, implying that the learning signal necessary for forming them is sent backward from the perirhinal cortex.

I. INTRODUCTION

In the visual system of the brain, information on the shape of stimuli is sent from area V1 via areas V2, V4 and TEO to area TE (TE) in the inferior temporal cortex (IT), where the shape is thought to be recognized, and thence further transmitted toward the medial part, namely, the perirhinal cortex (PRh), entorhinal cortex and hippocampal body (Fig. 1). Among these, TE and PRh are known to be deeply involved in stimulus–stimulus association memory.

With regard to this, Sakai and Miyashita [1] reported a key finding on the memory-related activity of IT neurons. They trained monkeys on a delayed pair-association (DPA) task, in which one of the pictures is presented as a cue and the monkey must judge whether a test picture presented after a delay interval is the paired associate (target) of the cue or not, and recorded neuronal activity during the task. One of the most important observations was that some neurons, termed pair-recall neurons, exhibited prospective activity; they do not response to the cue, but gradually increase activity during the delay period, showing the maximum activity when the target is presented.

Such activity is not explained by conventional memory models using point attractor networks (e.g. [2]), and the underlying mechanism was not clear. We have investigated this problem and constructed a computational model of pair-association (PA) memory based on a trajectory attractor network [3,4,5]. This model can not only learn and perform a PA task in a biologically feasible manner, but can also reproduce the activity of the pair-recall neurons well. Moreover, assuming that the two components of the model correspond to TE and PRh, it accords with physiological findings, and makes some predictions on the activity of TE neurons. These predictions, however, were not able to be verified by the empirical data obtained from Sakai and Miyashita [1].

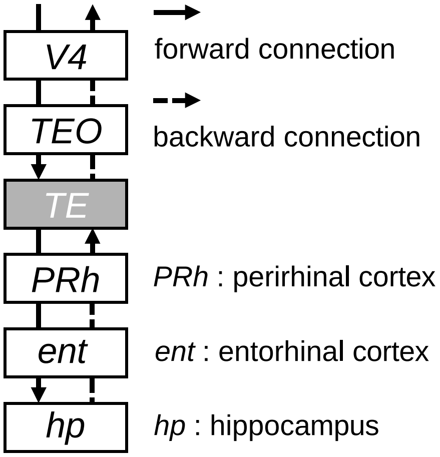


Fig. 1: Forward and backward pathways of visual signals in the temporal lobe.

Recently, Naya et al. [6] recorded pair-recall neurons more extensively and analyzed their temporal characteristics to obtain striking data. In the present report, we apply the same analysis to our model and compare the results to examine the formation and recall mechanisms of PA memory in IT.

II. THE MODEL OF PAIR-ASSOCIATION MEMORY

In this section, we briefly describe our model of PA memory (refer to [5] for a detailed description).

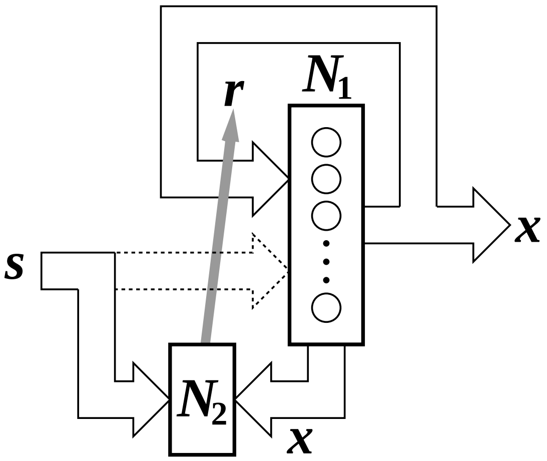


Fig. 2: Block diagram of a model of pair-association memory.

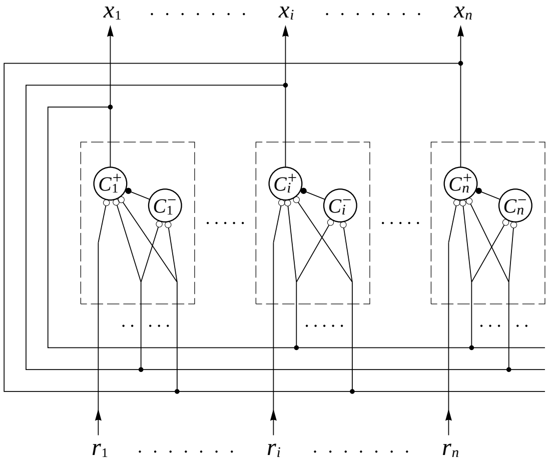

x_1 .... x_i .... x_n
C_1^+
C_1^-
C_i^+
C_i^-
C_n^+
C_n^-
r_1 .... r_i .... r_n

Fig. 3: Structure of network N_1.

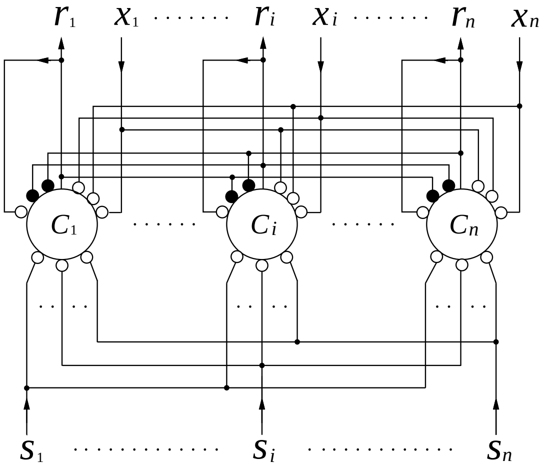

r_1 x_1 .... r_i x_i .... r_n x_n
C_1
C_i
C_n
s_1 .... s_i .... s_n

Fig. 4: Structure of network N_2.

A. Structure and Dynamics

The model is composed of two interconnected neural networks, association network N_1 in which memories are stored and trainer network N_2 which generates the learning signal required for memory formation (Fig. 2). The output pattern $\boldsymbol{x} = (x_1, \ldots, x_n)$ of N_1 is sent to N_2 and also fed recurrently into N_1, and the output pattern $\boldsymbol{r} = (r_1, \ldots, r_n)$ of N_2 is fed back to N_1 as the learning signal. The external input pattern $\boldsymbol{s} = (s_1, \ldots, s_m)$ is fed into N_2; although N_2 should receive $\boldsymbol{s}$ via N_1, the direct input path to N_1 is omitted for simplicity.

The structures of networks N_1 and N_2 are shown in Fig. 3 and Fig. 4, respectively, and their dynamics are described mathematically by

$$
\begin{aligned}
y_i &= f\left(\sum_{j=1}^{n} w_{ij}^{-} x_j - \theta\right), && (1)\\
\tau \frac{du_i}{dt} &= -u_i + \sum_{j=1}^{n} w_{ij}^{+} x_j - w_i^{*} y_i + \lambda r_i, && (2)\\
x_i &= f\left(u_i\right), && (3)\\
\tau \frac{dv_i}{dt} &= -v_i + \sum_{j=1}^{n} p_{ij} s_j + \sum_{j=1}^{n} q_{ij} x_j, \\
&\quad - \rho \sum_{j \neq i} r_j + \sigma r_i + h, && (4)\\
r_i &= f\left(v_i\right), && (5)\\
f(u) &= \frac{1}{1+e^{-cu}}, && (6)
\end{aligned}
$$

where w_{ij}^{+} and w_{ij}^{-} are the synaptic weights from the j-th unit to excitatory and inhibitory cells C_i^{+} and C_i^{-} of the i-th unit, respectively, w_i^{*} represents the efficiency of the inhibitory synapse from C_i^{-} to C_i^{+}, and p_{ij} and q_{ij} are random weights of input synapses to the i-th cell C_i of N_2; u_i and v_i are the potentials of C_i^{+} and C_i, respectively, λ denotes the input intensity of $\boldsymbol{r}$, h is an offset, and θ, τ, ρ, σ and c are positive constants.

B. Learning

Learning of this model is performed by modifying the synaptic weights of N_1 according to

$$
\begin{aligned}
\tau' \frac{dw_{ij}^{+}}{dt} &= -w_{ij}^{+} + \alpha r_i x_j, && (7)\\
\tau' \frac{dw_{ij}^{-}}{dt} &= -w_{ij}^{-} - \beta_1 r_i x_j + \beta_2 x_i x_j + \gamma, && (8)
\end{aligned}
$$

while each cell of the model is running according to Eqs. (1–6) [7]. Here τ' is a time constant ($\tau' \gg \tau$), α, β_1, and β_2 are learning coefficients ($\beta_1 < \beta_2$), and γ is a positive constant representing the lateral inhibition between units. Coefficient α may be a positive constant, but because the learning performance is better when α is a decreasing function of x_i, we adopt

$$\alpha = \begin{cases} \alpha'(\kappa - x_i) & (x_i < \kappa) \\ 0 & (x_i \geq \kappa), \end{cases} \tag{9}$$

where $\kappa \equiv \beta_1/\beta_2$ and α' is a positive constant.

By repeating this synaptic modification several times, a pattern 'close' (in the sense of the angle between pattern vectors) to learning signal $\boldsymbol{r}$ becomes a point attractor in the state space when $\boldsymbol{r}$ stands still; however, a successive attractor is formed along the track of $\boldsymbol{r}$ if $\boldsymbol{r}$ varies continuously. In addition, by $\boldsymbol{r}$ leading the move of $\boldsymbol{x}$, a gentle flow from $\boldsymbol{x}$ toward $\boldsymbol{r}$ or in the movement direction of $\boldsymbol{r}$ is produced. A string-shaped attractor with such a flow is called a trajectory attractor [8].

A pair of patterns is associated by sequentially feeding them to the model and performing the above described learning. For example, when a pattern A is fed to N_2 ($\boldsymbol{s}$ = A), N_2 outputs a randomly transformed pattern a ($\boldsymbol{r}$ = a); if the pair pattern B is then fed, interactions between N_1 and N_2 enable the output $\boldsymbol{r}$ of N_2 to change successively, drawing a track as schematically shown in Fig. 5, to a pattern b' near to a pattern b encoding B. In the same way, when we input B and A, in that order, a spatiotemporal pattern varying from b to a' is sent from N_2 to N_1. Consequently, trajectory attractors are formed in N_1 along these tracks and the output pattern $\boldsymbol{x}$ comes to shift automatically from a to b' and b to a' with the input of A and B, respectively.

C. *Correspondence to Physiological Data and Predictions*

The model after learning can not only perform the DPA task, but also reproduce the activity of pair-recall neurons in IT; in addition, it possesses two important properties.

First, it is presumed that networks N_1 and N_2 correspond to TE and PRh, respectively. This presumption agrees with the anatomical structure as shown in Fig. 1, and also with the results of lesion studies. For example, the monkey with lesions of the perirhinal and entorhinal cortices cannot learn a PA task at all, although recognition of visual stimuli is intact and re-learning of a stimulus set learned before the lesion is possible [9]. This suggests that PRh is essential for PA learning but PA memories are finally formed in TE, as assumed by the model.

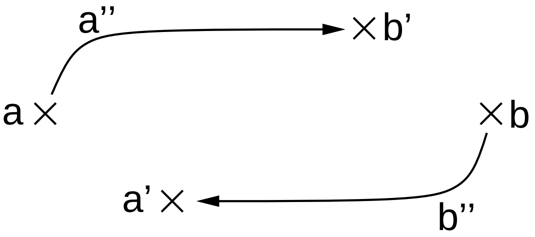


Fig. 5: Paths of the learning signal.

Second, the following characteristics of the pair-recall neurons in TE are predicted.

(1) The onset time of the pair-recall activity during the delay period will be widely diverse. This is a direct consequence of the continuous state transition of network N_1 along a trajectory attractor, since if the state jumps, many units will change their output synchronously.

(2) The activity pattern of neurons when a picture is presented as a cue will be substantially different from that recalled during the delay period after presentation of the paired picture, in other words, the code of the target will be recalled only incompletely. This is derived from the discrepancy between a and a' in Fig. 5, which is necessary for forming two trajectory attractors in opposite directions with avoiding mutual interference.

In the following, we try to verify these predictions on the basis of the empirical data from Naya et al. [6].

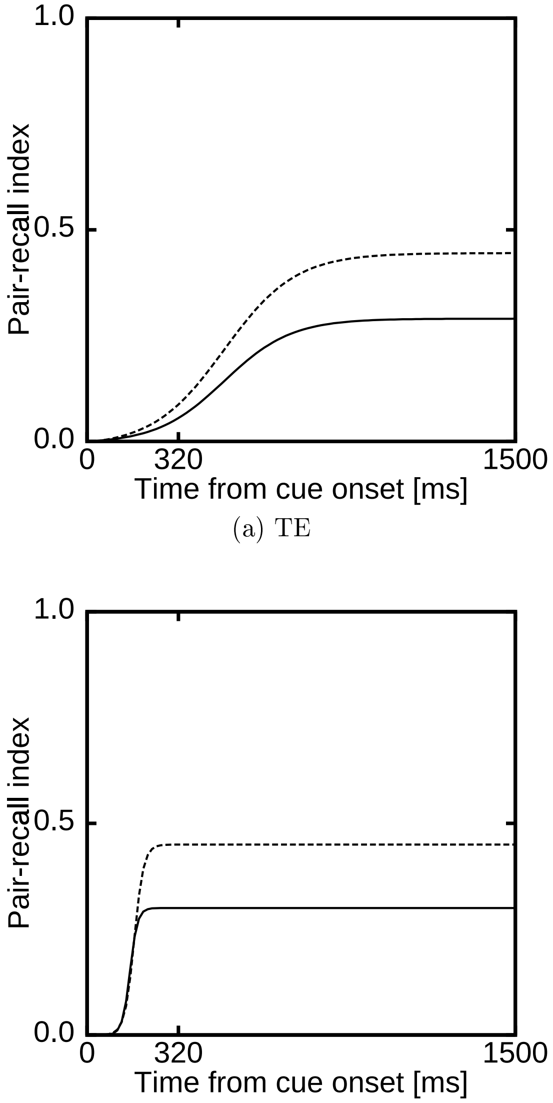
III. ANALYSIS OF THE TIME COURSE OF RECALL ACTIVITY

The method and results of the analysis by Naya et al. [6] are summarized as follows.

They trained monkeys on the same DPA task as Sakai and Miyashita [1], and recorded neuronal activity during the task from TE and area 36 (A36), a part of PRh adjacent to TE. Subsequently, for each neuron exhibiting a selective response during the delay period, a pair-recall index (PRI) was calculated according to

$$PRI(t) \;=\; \frac{\langle C_p|F(t)\rangle - \langle C|C_p\rangle\langle C|F(t)\rangle}{\sqrt{(1-\langle C|F(t)\rangle^2)(1-\langle C|C_p\rangle^2)}}.$$

Here, $F(t)$, C and C_p denote vectors $[f_1(t), \cdots, f_l(t)]$, $[c_1, \cdots, c_l]$ and $[c_{p(1)}, \cdots, c_{p(l)}]$, respectively, l being the number of learned pictures ($l = 24$ in their experiment), $f_k(t)$ the activity (firing rate) at time t when the picture k is presented as a cue, c_k and $c_{p(k)}$ the cue activities when the picture k or its paired associate $p(k)$ is presented as



(a) TE

(b) A36

Fig. 6: Time courses of the averaged PRI for population of pair-recall neurons in TE and A36 (adapted from Naya et al. [6]).

a cue, and $\langle X|Y\rangle$ indicates the correlation coefficient between X and Y. This index indicates how similar the activity at time t is to the response to the target, and is normalized so that it may not exceed 1 and may be nearly zero during the cue period; also it is not affected by the difference in the activity level between the cue period and the delay period.

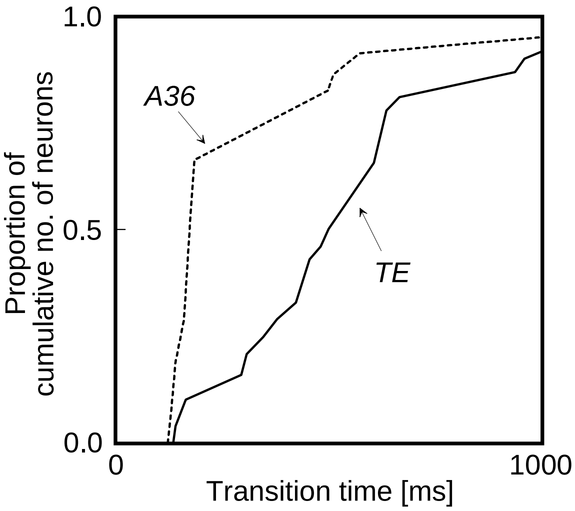


Fig. 7: Distribution of the TRT values for IT neurons (adapted from Naya et al. [6]).

Fig. 6 shows the time course of the population average of PRI, where the best-fit Weibull functions for the averaged $PRI(t)$ of the (a) TE and (b) A36 neurons are plotted; solid lines indicate the average of all neurons, whereas broken lines indicate the average of the neurons exhibiting a significant increase in PRI (5% significance level). Here, a notable point is that the averaged PRI for TE increased gradually for more than 500 ms. It is also a very novel and important finding that although the cue response is later in A36 than in TE (see Fig. 1), the increase of PRI is considerably earlier, which we will discuss in Section IV.

They subsequently obtained the transition time (TRT) for each neuron which is defined as the time when $PRI(t)$ reaches 50% of its maximum. Fig. 7 shows a cumulative frequency histogram of the TRT values, where the solid line and broken line indicate TE and A36, respectively. The graph for TE increases gradually with TRT, indicating that the onset time for the $PRI(t)$ increase is distributed nearly uniformly over a wide range.

We analyzed our model using the same method. Specifically, after training the model with $n = 1000$ on 20 pairs of randomly generated patterns, we applied the DPA task to the model and calculated $PRI(t)$ for 100 randomly selected units of N_1. Methods of the simulation and parameters of the model were the same as those in our previous study [5], except that the input and delay periods were slightly modified to facilitate the comparison with the empirical data.

The results are displayed in Figs. 8 and 9. Fig. 8 shows the averaged PRI in the same manner as Fig. 6, but time (the abscissa) is scaled by the time constant τ in Eq. (1) and curve fitting is not applied. Fig. 9 shows a cumulative frequency histogram of the TRT in the same manner as Fig. 7.

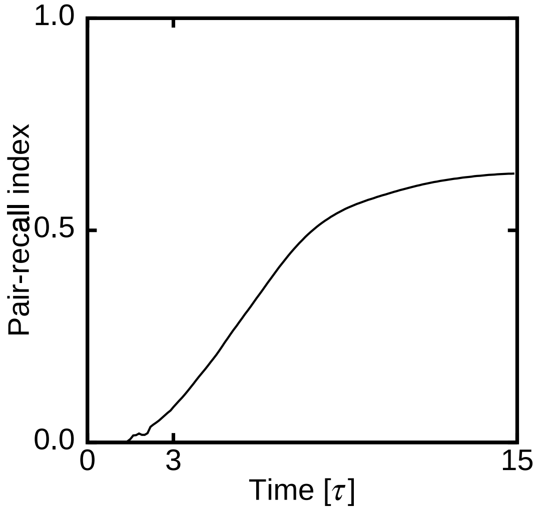


Fig. 8: Time course of the averaged PRI for the units of N_1 of the model.

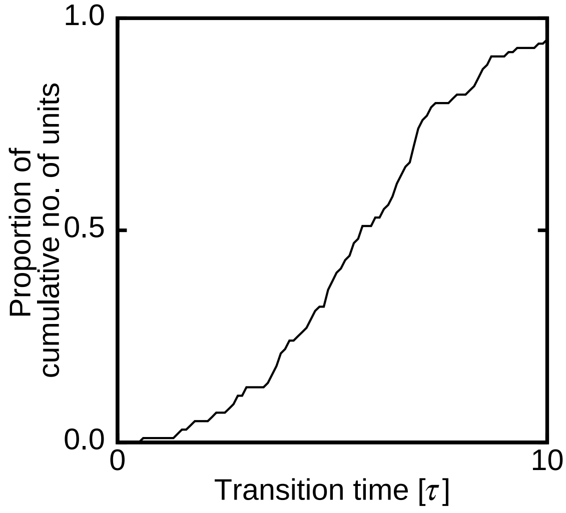


Fig. 9: Distribution of the TRT values in the simulation.

IV. DISCUSSION

Comparing Fig. 6(a) and Fig. 8, we see that the curves of graphs for TE and the model are similar in form, but the value of averaged $PRI(t)$ in the former is about half of that in the latter. This difference seems partly because neurons that do not contribute to recall of the target may possibly be included in the empirical data whereas such units are not in the model. In fact, the graphs are quantitatively more similar if TE neurons without significant change in $PRI(t)$ are omitted (Fig. 6(a), broken line).

At the same time, the fact that the averaged PRI in TE does not exceed 0.5 even after such an operation indicates that the activity pattern, not only its level, recalled during the delay period is substantially different from that elicited by presentation of the target. Such a difference has been inferred from the finding that recognition errors occur most frequently when a nontarget test is similar to the correct target [10], but the above data shows it more directly and verifies the prediction (2) described in Section II.

Next, as evident from a comparison between Fig. 7 and Fig. 9, the TRT distributions for TE and the model agree very closely. This is thought to be not simply a coincidence but a reflection of the working principle of the model, since a similar distribution was obtained when simulations were repeated with different conditions. Accordingly, the empirical data is not only consistent with prediction (1) in Section II, but also suggests that the activity pattern of TE neurons changes successively in the same course as that of the model, and furthermore, that the change is brought about by the same mechanism, namely, trajectory attractors.

The argument that trajectory attractors are formed in TE has another important ground. As described above, to form a trajectory attractor in a network, a learning signal which leads, and thus should precede, the state transition of the network is required, and it is a main point of our model that this learning signal is sent from N_2 to N_1. Since it is reasonably presumed that N_2 corresponds to PRh, the marked finding shown in Figs. 6 and 7 that the activity shift in A36 toward the target-coding state is earlier than that in TE supports the model.

Here an argument may arise that it is not necessary to form trajectory attractors in TE if the activity in A36 leads the state transition in TE. However, a view that pair-association memories are stored not in TE but in PRh only does not agree with the above lesion study of PRh. Also, if TE contains only point attractors encoding individual pictures and the target recall depends completely on A36 or other regions, the TRT distribution for TE should not be so wide as shown in Fig. 7, since the state during transitions between attractors is generally unstable and transient; in addition, the discrepancy between target and recalled codes is not explained. We believe, therefore, that it is the most reasonable interpretation that trajectory attractors underlie the pair-recall activity in TE.

V. CONCLUDING REMARKS

As described above, the empirical data on the pair-recall neurons in TE agreed well with calculations performed using our computational model. This strongly suggests that trajectory attractors are formed in TE and that the learning signal necessary for forming them is sent backward from PRh.

However, the present model was constructed for the purpose of explaining the activity of TE neurons but not of PRh neurons, and is insufficient for a model of PRh. As a matter of fact, curves in similar form to those in Fig. 6(b) and Fig. 7 (broken line) are obtained if we calculate PRI and TRT for cells of N_2 using their outputs in learning. Nevertheless, they are quite different in terms of the origin of abscissa; that is, the averaged PRI for A36 begins increasing immediately after the cue presentation, whereas PRI of N_2 cells increases after the input of the target pattern because N_2 is currently not endowed with any mechanism of target recall.

In connection with this, Erickson and Desimone [11] demonstrated that the delay activity of PRh neurons reflects the cue at an early stage of training but it comes to reflect the prospective target with long-term training. Also, Tokuyama et al. [12] reported that BDNF (brain-derived neurotrophic factor), which is thought to mediate synaptic plasticity, was selectively induced in PRh during PA learning. These findings imply that learning in PRh is performed previous to or in parallel with the memory formation in TE.

To fully explain the above data on A36, therefore, modification of the structure of N_2 with introduction of synaptic plasticity will be necessary; also, the addition of another network corresponding to the inner part of A36 (e.g., the entorhinal cortex) will possibly be required. Such improvements of the model are a subject for future study.

References

[1] Sakai, K. and Miyashita, Y. (1991) : Neural organization for the long-term memory or paired association, Nature, Vol. 354, pp. 152–155

[2] Amit, D.J. and Fuji, S. (1997) : Paradigmatic working memory (attractor) cell in IT cortex, Neural Computation, Vol. 9, pp. 1071–1092

[3] Suemitsu, A. and Morita, M. (1999) : A neural network model of pair-association memory in the inferotemporal cortex, Proceedings of the 1999 International Conference on Neural Information Processing, Vol. 2, pp. 790–794

[4] Suemitsu, A. and Morita, M. (2000) : A model of memory formation in the pair-association task, Proceedings of the 2000 International Conference on Neural Information Processing, Vol. 2, pp. 915–919

[5] Morita, M. and Suemitsu, A. (in press) : Computational modeling of pair-association memory in inferior temporal cortex, Cognitive Brain Research

[6] Naya, Y., Yoshida, M. and Miyashita, Y. (2001) : Backward spreading of memory-retrieval signal in the primate temporal cortex, Science, Vol. 291, pp. 661–664

[7] Morita, M. (1996) : Computational study on the neural mechanism of sequential pattern memory, Cognitive Brain Research, Vol. 5, pp. 137–146

[8] Morita, M. (1996) : Memory and learning of sequential patterns by nonmonotone neural networks, Neural Networks, Vol. 9, pp. 1477–1489

[9] Murray, E.A., Gaffan, D. and Mishkin, M. (1993) : Neural substrates of visual stimulus–stimulus association in rhesus monkeys, J. Neuroscience, Vol. 13, pp. 4549–4561

[10] Rainer, G., Rao, S.C. and Miller E.K. (1999) : Prospective coding for objects in primate prefrontal cortex, J. Neuroscience, Vol. 19, pp. 5493–5505

[11] Erickson, C.A. and Desimone, R. (1999) : Responses of macaque perirhinal neurons during and after visual stimulus association learning, J. Neuroscience, Vol. 19, No. 23, pp. 10404–10416

[12] Tokuyama, W., Okuno, H., Hashimoto, T., Li, Y.X. and Miyashita, Y. (2000) : BDNF upregulation during declarative memory formation in monkey inferior temporal cortex, Nature Neuroscience, Vol. 3, No. 11, pp. 1134–1142