Reinforcement Learning with Multiple Representations in The Basal Ganglia Loops for Sequential Motor Control

Hiroyuki Nakahara† Kenji Doya‡, Okihide Hikosaka*, Saburo Nagano§
† Lab. For Info. Synthesis, Brain Science Institute, RIKEN
‡ Kawato Dynamic Brain Project, JST
* Dept. of Physiology, Juntendo Univ., School of Med.
§ Dept. of General Systems Studies, Univ. of Tokyo

Abstract—**The basal ganglia (BG) have been hypothesized to perform reinforcement learning by use of reinforcement signals provided by dopamine neurons. It is well known that there exist multiple BG-thalamocortical loops, but their functions are poorly understood. Here, we propose a computational model of how different BG loops are employed in visuomotor sequence learning using different representations of sequence. The central idea of the model is that a visuomotor sequence is easier to *learn* in spatial representation (e.g. visual coordinates) but is easier to *control* in body-based representation (e.g. joint angle coordinates). The results of simulations of the model replicated both behavioral and neurophysiological findings in recent experimental studies using "2x5 task".**

Keywords— **reinforcement learning, basal ganglia, procedural memory, visuomotor learning**

I. INTRODUCTION

Schultz and his colleagues has shown in their experiments [1] that the response tuning of dopamine (DA) neurons in the substantia nigra pars compacta (SNc) in the basal ganglia (BG) shifts from primary reward to conditioned stimuli that predicts reward as the conditioning establishes. This fact has led to a hypothesis that a specific form of reinforcement learning (RL), temporal difference (TD) learning [2], occurs in the BG with TD error of reward prediction provided by DA neurons as reinforcement signal (e.g. [3], [4], [5]). This hypothesis provides a computational basis for investigating the BG function in sequential motor control, which has been suggested by studies on brain lesions and neuronal recordings.

The BG-thalamocortical circuit is one of the major cortico-subcortical circuits for motor control, with the other being the cerebellocortical circuit. A striking feature of the BG circuit is its separate, closed loop organization. The BG receives projections from almost the entire cerebral cortex but each part of BG projects to specific area of the frontal cortex. At least four BG loops have been identified, including motor, oculomotor and dorsolateral prefrontal loops [6]. Functions of some of these loops have been proposed, but it is very poorly understood how these BG loops work *together* in sequential motor control tasks.

Any correspondence should be sent to H. Nakahara, Lab. For Information Synthesis, Brain Science Institute, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0106, Japan, Email: hiro@brain.riken.go.jp.

In recent experiments of sequential arm reaching task called "2x5 task"[7], Hikosaka and his colleagues have found that different parts of the BG and the frontal cortex are involved differentially in acquisition and execution of sequential movement [8], [9]. Motivated by their results, we propose a computational model of how different BG-cortical loops are involved in learning of sequential movement at different stages. The key idea is that a visuomotor sequence like that in the 2x5 task is easier to *learn* in spatial representation (e.g. visual coordinates) but is easier to *control* in body-based representation (e.g. joint angle coordinates) [10]. Specifically, we propose that the dorsolateral prefrontal (DLPF) loop learns a sequence in spatial representation and the supplementary motor area loop learns a sequence in body-based representation. Both loops learn concurrently using the reinforcement signal carried by DA neurons, but relative ease of learning and effectiveness in control causes the differential involvement of these two loops. Simulation of the proposed model replicated both behavioral and neurophysiological findings of the 2x5 task experiments.

II. 2x5 TASK AND ITS FINDING

Figure 1 shows an example of the sequence of events in a single trial of the 2x5 task. When the animal pressed the home key at the start of a trial, two out of the 16 LED buttons were turned on simultaneously, which is called a 'set' of stimulus. The animal had to press the illuminated buttons in a predetermined order, which he had to find out by trial-and-error. If successful, another pair of LEDs, the second set, was illuminated which the monkey had to press again in a predetermined order. A fixed sequence of 5 sets, called a 'hyperset,' was presented in a trial. When the animal pressed a wrong button, all LED buttons were illuminated briefly with an unpleasant beep sound, and the trial was aborted without any reward. The animal then had to start over a new trial by pressing the home key. After each successful set, the animal was given a liquid reward.

The same hyperset was used throughout a 'block' of experiments until completing a certain number of successful

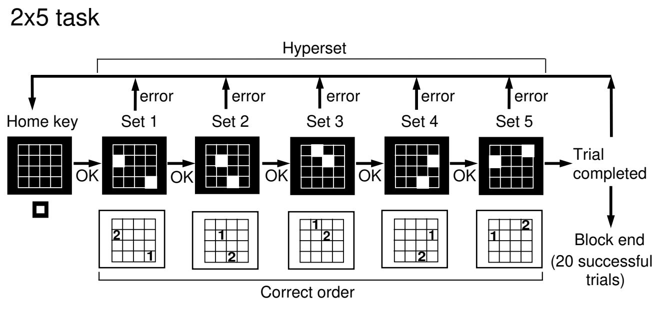


Fig. 1. Procedure of 2x5 task with an example of a hyperset . To complete a trial, a monkey has to press 10 buttons (2 buttons x 5 sets) in a correct (predetermined) order.

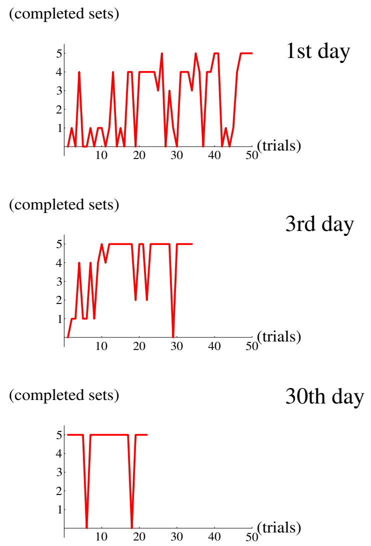


Fig. 2. Experimental result of learning a hyperset across days. The change in the number of completed sets (ordinate) across trials (abscissa) is compared among the 1st day (top), the 3rd day (middle) and the 30th day (bottom). Taken from Hikosaka et al.

trials (criterion). On each day during a training period, the monkey performed several blocks of trials with different hypersets. Some hypersets were used every day, called 'learned' hypersets, and others were randomly generated and used only once, called 'new' hypersets. Order of presentation of learned and new hypersets are randomized everyday.

Learning in 2x5 tasks is measured by the decrease in the number of trials to reach a criterion to complete a block and/or the decrease in the performance time. Hikosaka et al. [7] observed both *short-term* learning and *long-term* learning. Short-term learning is indicated by improved performance during a block of experiment and long-term learning is indicated by improved performance across days (see Figure 2).

Functional differentiation was observed in blockade experiment by muscimol injection between anterior BG (caudate head) and posterior BG (putamen) [8], as well as the presupplementary motor area (pre-SMA) and the supplementary motor area (SMA) [9]. Number of error trials to criterion was significantly increased by blockade of the anterior BG or the pre-SMA for new hypersets but not for learned ones. Number of error trials to criterion was significantly increased by blockade of the posterior BG for learned hypersets but not for new ones. Blockade of the SMA affected both learned and new ones, but only mildly. In brief, these lines of experimental evidence suggest that the anterior BG and the pre-SMA is more involved in early acquisition process of sequences, and that the posterior BG and, possibly, the SMA are more involved in maintenance and retrieval of acquired sequential memory.

III. Hypothesis on Multiple Representations in the Basal Ganglia Loops

In visuomotor tasks such as visually-guided reaching, problems in inverse kinematics and inverse dynamics must be solved to reach a target based on visual information [11]. We propose that it is easier to learn a visuomotor sequence in visual coordinates particularly when the sequence is learned by trial-and-error, whereas it is faster and easier to execute the sequence in body-based coordinates once it is acquired. We further propose that it is advantageous to have multiple learning processes with different speeds concurrently. A quick acquisition of a sequence in visual coordinates would help the animals solve new problems that they encounter in everyday life; a slow acquisition of a sequence in body coordinates would help the animals store the memory of the frequently-used sequence robustly and retrieve it quickly. Thus, it is plausible to consider from a computational viewpoint that a sequence in visual coordinates is more suitable for early acquisition stage, whereas a sequence in body-based coordinates is more suitable for execution with robust maintenance and quick retrieval of its memory.

The dorsolateral prefrontal loop: the projection from DLPF to the BG is primarily to anterior striatum, including head of the caudate nuclei (CD) and the rostral putamen. The DLPF is well known to be involved in visuospatial memory and is considered to play a role in control and planning of sequential movements [12]. Posterior parietal cortex that is connected with the DLPF also projects to the dorsolateral head of the CD [6]. Based on these facts, it is likely that the DLPF loop, including the DLPF and the anterior striatum, engages computation in visual coordinates (Figure 3).

The motor loop: in motor loop, most of the projections to the BG originate from the cortical motor areas, among which the SMA is of particular interest, and principally terminates in the posterior striatum (the bulk of putamen) [6]. The SMA is well connected with other motor cortical areas and has been known long as involved in sequential movements, particularly internally-generated complex ones [13]. Thus, the motor loop, including the SMA and the posterior striatum, may engage computation in body-based coordinates (Figure 3).

It is noteworthy that not the SMA but the pre-SMA

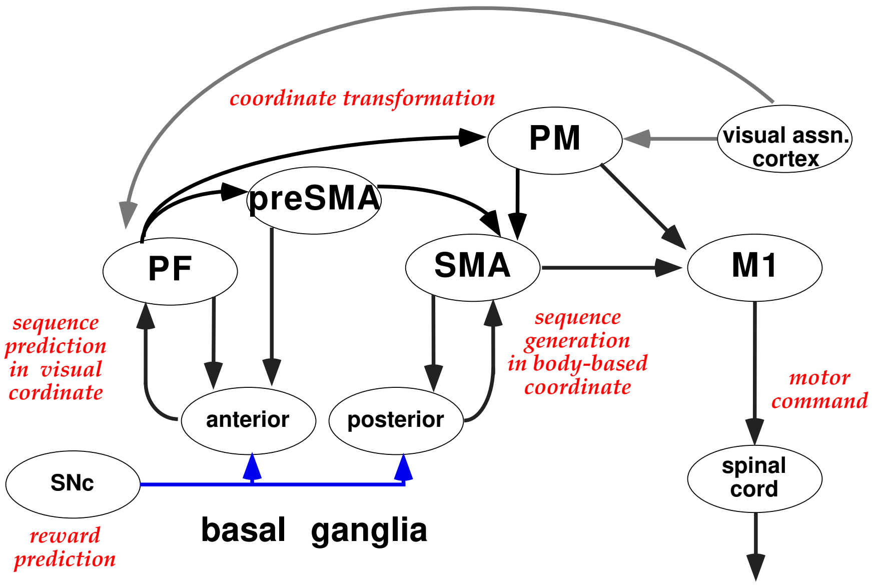


Fig. 3. Hypothesized scheme of the basal ganglia-thalamocortical loops for sequential motor control.

is connected with the DLPF [13]. The pre-SMA has the projections to the anterior striatum. Hence, the pre-SMA is heavily interacted with the DLPF loop. It is also experimentally shown that the pre-SMA is generally more activated in the period after receiving sensory inputs and before starting movements [14], [13].

We hypothesize that the DLPF loop, perhaps together with the pre-SMA, learns the sequence using visual coordinates. The DLPF loop, hence, is more critical in early acquisition stage of sequences. In contrast, the motor loop learns the sequence using body-based coordinates. The motor loop, hence, is more involved in execution of well-learned sequences. A unique feature of our model is that both loops concurrently learn sequences based on reinforcement signal provided by DA neurons in SNc (Figure 3).

IV. Simulation on 2x5 Task

In order to test the behavior of a model based on our hypothesis, we built a neural network of 2x5 task learning. The specific questions we asked were whether the model would have the short-term and long-term learning effects in performance and whether the model could replicate the results of blockade experiments for the DLPF loop.

A. Network Architecture

Figure 4 illustrates the overall structure of the network model, which consisted of the visual network (DLPF-BG loop), the motor network (SMA-BG loop), the critic network (DA system), and inverse kinematics modules.

Both the visual and motor networks had the same structure and had an output layer and a context layer, each corresponding to the BG and the cortex. First, the state y_i of the output layer was updated by the softmax function of the weighted sum of the input u_j and the context x_j with a tuning parameter for sharpness of the softmax, denoted by ζ.

$$s_i(t) = \sum_j w^I_{ij} u_j(t) + w^C_{ij} x_j(t) \qquad (1)$$

$$y_i(t) = \frac{\exp[\zeta s_i(t)]}{\sum_k \exp[\zeta s_k(t)]} \qquad (2)$$

where w^I_{ij} and w^C_{ij} are weights for the input and the context, respectively.

Then, one of the output units was stochastically set as $z_i(t) = 1$ with a probability

$$\mathrm{Prob}(z_i(t) = 1) = y_i(t). \qquad (3)$$

The context layer was updated, with a time constant, τ, by the following equation.

$$x_i(t+1) = x_i(t) + \frac{1}{\tau}(z_i(t) - x_i(t)) \qquad (4)$$

The weight matrices w^I and w^C were updated by a Hebbian rule, weighted by the reinforcement signal $\hat{r}(t)$, as defined below;

$$w^I_{ij}(t+1) = w^I_{ij}(t) + \eta^I \hat{r}(t) z_i(t) u_j(t) \qquad (5)$$

$$w^C_{ij}(t+1) = w^C_{ij}(t) + \eta^C \hat{r}(t) z_i(t) x_j(t) \qquad (6)$$

Initially, the input weights were set by an identity matrix so that the default behavior is to press one of the lit buttons.

In the visual network, the state was encoded by the 16 units corresponding to the positions of 16 buttons in the Cartesian space. The input u^V to the visual network was a 16 dimensional vector of 0 and 1, encoding the corresponding lit buttons as 1s and otherwise as 0s. Its output z^V represented the button to be pressed, or the movement target in the visual space.

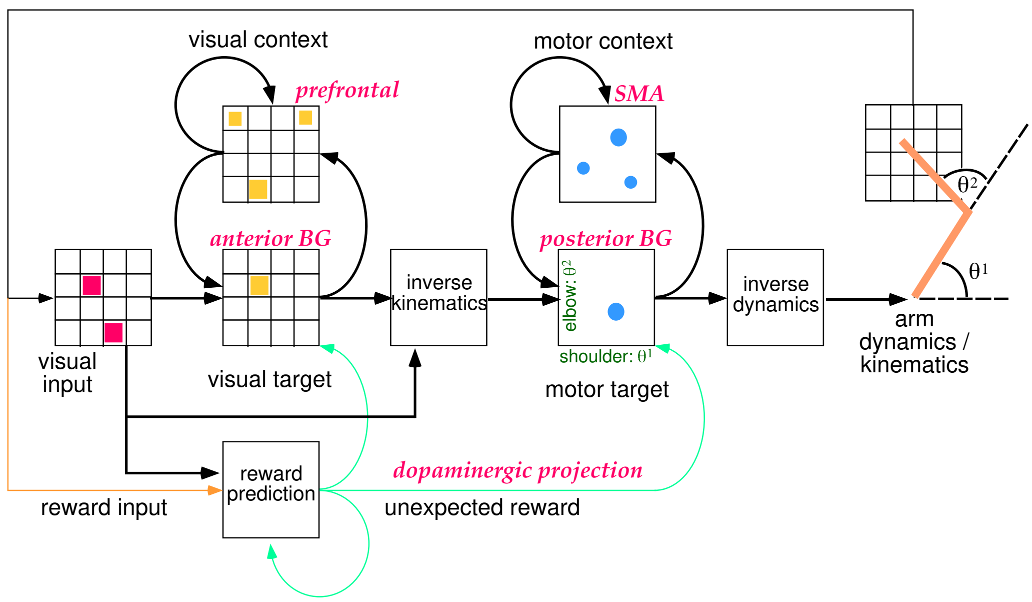


Fig. 4. Diagram of the proposed model in context of the 2x5 task

In the motor network, state was encoded by a population vector of joint angles. Each unit had a Gaussian softmax activation function

$$a_j(\theta) = \exp[-\frac{1}{2}\sum_{i=1}^{2}\left(\frac{\theta^i - \theta^i_j}{\delta^i}\right)^2], \tag{7}$$

$$b_j(\theta) = \frac{a_j(\theta)}{\sum_k a_k(\theta)} \tag{8}$$

where θ^1 and θ^2 denote shoulder and elbow angles and θ^i_j denotes the preferred joint angle for the j-th unit.

An inverse kinematic model, which was derived analytically from the geometry of the arm, was used to transform the visual representation into corresponding motor representation. In the normal operation, the input to the motor network was a sum of two activation vectors, one calculated from the current visual input and the other from the output of the visual network.

The critic network provides the estimate of value function that stands for the value of the current state, or the current sensory input, denoted by u^R. Its encoding is the same as the encoding of the state in the visual network. The estimated value of a state, $P(t)$, is defined by

$$P(t) = \Sigma_j w^R_j u^R_j + b^R \tag{9}$$

where w^R_j and b^R are the weight matrix and bias for the critic. Using this estimated value function, the reinforcement signal, or TD error, $\hat{r}(t)$, is computed by

$$\hat{r}(t) = r(t) + \gamma P(t+1) - P(t) \tag{10}$$

The critic weight matrix, w^R, and bias, b^R, is updated by use of TD error as defined below:

$$w^R_j(t+1) = w^R_j(t) + \eta^R \hat{r}(t) u^R_j \tag{11}$$

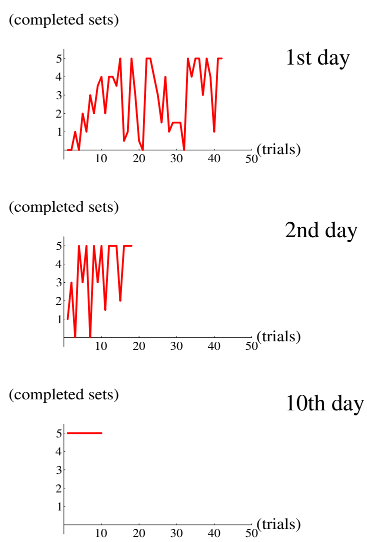


Fig. 5. Simulation result of learning a hyperset across days. The change in the number of completed sets (ordinate) across trials (abscissa) is compared among the 1st day (top), the 2nd day (middle) and the 10th day (bottom).

$$b^R(t+1) = b^R(t) + \eta^b \hat{r}(t) \tag{12}$$

Thus, note that we used TD(0) learning in simulation we report below for sake of simplicity and that the overall architecture of the network above is based on actor-critic scheme [2].

B. Simulation Setup

We let the model learn 2 learned hypersets and 1 new hyperset per a simulated day, approximately keeping ratio between learned and new hypersets in experiment. Simulation is run for 10 days to train the model networks. Learn-

TABLE I

Results on the blockade of visual network: number of trials to criterion for learned hypersets

	NORMAL	BLOCKADE
Learned	11.4(±1.07)	11.7(±0.95)

ing process during this period is examined to test whether the model exhibits different learning levels. Then, using the trained network parameters, we tested the performance of the network for learned and new hypersets in case of the blockade of the DLPF loop, or the visual network. In the simulation, the blockade of the visual network was realized by inhibiting input from the visual network to the motor network.

C. *Results*

C.1 Different learning levels

Hikosaka et al. [7] indicated *short-term* learning by improved performance during a block and *long-term* learning by improved performance across days. It is clear in Figure 6 (right) that the model makes more errors in the first half of total trials than in the second half, in particular for the 1st and 2nd days. This indicates short-term learning (see also Figure 5). Figure 6 (left) shows that the model improved its performance for learned hypersets across days particularly for first few days, indicating long-term learning as observed in the experiments (see also Figure 5).

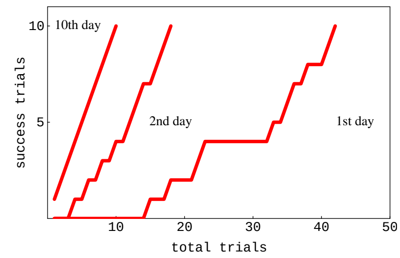
C.2 Blockade of the anterior basal ganglia and the presupplementary motor area

It is observed that the performance for the learned hyperset was not deteriorated by the blockade of the visual network (Table I), similarly to the results of the blockade experiments in the 2x5 task [8], [9]. In the present network architecture, it was observed in the training period that the network could not learn some types of new hypersets at all and the same phenomena was sometimes observed in the blockade condition as well, which will be discussed in the next section.

V. Discussion

We have hypothesized that the basal ganglia contribute to sequential motor control based on reinforcement learning with multiple representations. The dorsolateral prefrontal loop uses visual coordinates and contributes to learning sequences particularly in early stage. The motor loop uses body-based coordinates and provides sequential memory, once acquired, with robustness, even though the motor loop learns them much slower than the dorsolateral prefrontal loop. The results of the model based on this hypothesis resembled some of the experimental results in the 2x5 task: different levels of learning at behavior level, and functional differentiation at neurophysiological level.

As noted above, it was observed in the present simulation that the acquired sequential memory severely interfered with learning for some types of new hypersets in the training period as well as in the blockade condition. This is probably because, in the present architecture, the output from the visual network is forwarded only as an input to the motor network so that the output from the visual network cannot directly influence hand movements. We believe that the present architecture should be extended to include a mechanism to let the output from the visual network directly influence the generation of a final output. For example, a mechanism to weight or gate outputs from both of the visual and motor networks can be incorporated in the present architecture [10]. The recent study on neural activities in the pre-SMA suggests that such mechanism may exist and that the pre-SMA may be a part of it [15], [16]. This point is currently being investigated. In addition, the functions of the cerebellum was not explicitly addressed in the present study, however, there are experimental results on the cerebellum in the 2x5 task, indicating that the dentate nucleus is involved in maintenance of sequential memory [17]. Integrating the functions of the cerebellum into the proposed hypothesis must be a fruitful future research.



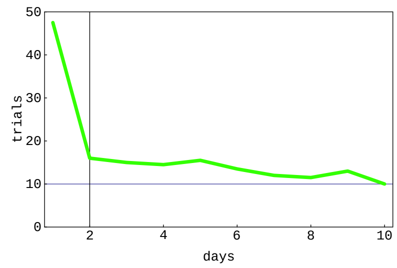


Fig. 6. Performance of the proposed model: (left) example of learning a learned hyperset across days by the model. The number of successful trials is plotted against the total number of trials for the 1st, 2nd, and 10th days. (right) the averaged performance for learned hypersets across 10 days. The mean number of trials to criterion is plotted across days.

Acknowledgments

Supported by JSPS and SPRF of RIKEN to HN, Uehara Memorial Foundation and JSPS Research for the Future Program to OH. HN acknowledges many insights from S. Miyachi, K. Miyashita, S. Sakai, X. Lu, and P. Pedroso in development of this study.

References

[1] W. Schultz, R. Romo, T. Ljungberg, J. Mirenowicz, J. R. Hollerman, and A. Dickinson, "Rewared-related signals carried by dopamine neurons," in *Models of Information Processing in the Basal Ganglia*, pp. 231–248, Cambridge, Massachusetts: MIT Press, 1995.

[2] A. Barto, R. Sutton, and C. Anderson, "Neuronlike adaptive elements that can solve difficult learning control problems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 13, pp. 834–846, 1983.

[3] J. C. Houk, J. L. Adams, and A. G. Barto, "A model of how the basal ganglia generate and use neural signals that predict reinforcement," in *Models of Information Processing in the Basal Ganglia*, pp. 249–270, Cambridge, Massachusetts: MIT Press, 1995.

[4] K. Doya, "Efficient nonlinear control with actor-tutor architecture," in *Advances in Neural Information Processing Systems* (D. Touretzky, M. Mozer, and M. Hasselmo, eds.), vol. 9, 1997.

[5] W. Schultz, P. Dayan, and R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, pp. 1593–1599, 1997.

[6] G. E. Alexander, M. D. Crutcher, and M. R. DeLong, "Basal ganglia-thalamocortical circuits: Parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions," in *Progress in Brain Research* (H. Uylings, C. Van Eden, J. De Bruin, M. Corner, and M. Feenstra, eds.), vol. 85, ch. 6, pp. 119–146, Elsevier Science Publishers B.V., 1990.

[7] O. Hikosaka, M. Kato, S. Miyachi, and K. Miyashita, "Learning of sequential movements in the monkey: Process of learning and retention of memory," *Journal of Neurophysiology*, vol. 74, no. 4, pp. 1652–1661, 1995.

[8] S. Miyachi, O. Hikosaka, K. Miyashita, Z. Karadi, and M. Kato, "Different roles of monkey straitum in learning of sequential movements," *Experimental Brain Research*, vol. in press, 1996.

[9] K. Miyashita, K. Sakai, and O. Hikosaka, "Effects of SMA and pre-SMA inactivation on learning of sequential movements in monkey," in *Proceedings of 26th Annual Meeting, Society for Neuroscience.*, p. 731.3, 1996.

[10] H. Nakahara, *Sequential Decision Making in Biological Systems: The Role of Nonlinear Dynamical Phenomena in Working Memory and Reinforcement Learning in Long-Term Memory.* PhD thesis, Univ. of Tokyo, 1997.

[11] M. Kawato, "Computational schemes and neural network models for formation and control of multijoint arm trajectory," in *Neural Networks for Control*, Cambridge, MA: MIT Press, 1990.

[12] P. S. Goldman-Rakic, "Circuitry of primate prefrontal cortex and regulation of behavior by representational memory," in *Handbook of Physiology - The Nearvous System V* (F. Plom and V. Mountcastle, eds.), vol. 5, ch. 9, pp. 373–417, 1987.

[13] J. Tanji, "The supplementary motor area in the cerebral cortex.," *Neuroscience Research*, vol. 19, pp. 251–68, May 1994.

[14] Y. Matsuzaka, H. Aizawa, and J. Tanji, "A motor area rostral to the supplementary motor area (presupplementary motor area) in the monkey: neuronal activity during a learned motor task," *Journal of Neurophysiology*, vol. 68, no. 3, pp. 653–662, 1992.

[15] K. Miyashita and O. Hikosaka, "Activity of pre-SMA neuron for the performance of sequential movement in monkeys," in *Proceedings of 27th Annual Meeting, Society for Neuroscience.*, (submitted) 1997.

[16] K. Shima, H. Mushiake, N. Saito, and J. Tanji, "Role for cells in the presupplementary motor area in updating motor plans.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, pp. 8694–8, Aug 6 1996.

[17] X. Lu, O. Hikosaka, and S. Miyachi, "Role of monkey cerebellar dentate nucleus in procedural memory," in *Proceedings of 26th Annual Meeting, Society for Neuroscience.*, p. 546.13, 1996.