Sequence Representation in Animals and Networks: Study of a Recurrent Network Trained with Reinforcement Learning

Raju S. Bapi **Kenji Doya**
Computational Neurobiology Group
Kawato Dynamic Brain Project, ERATO, JST
2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan

Abstract

Neural encoding for sequence identification, memory and production is studied using an Elman-style recurrent network (Elman, 1990) is studied. Novel feature of this network is that learning is implemented using biologically plausible reinforcement learning paradigm. Findings from Tanji & Shima's (1994) experiments on monkeys indicate that there is sequence-specific activity in the supplementary motor area (SMA) and change of motor-plan related activity in the preSMA. A network with anatomical correspondence to known cortico-basal ganglionic structure is designed to understand how sequence-specific activity arises due to training. Simulations demonstrate that context switching in the preSMA layer facilitates learning of multiple sequences and the hidden unit activity in the resultant network resembles what Tanji & Shima observed in monkeys. It is proposed that the basal ganglia (BG) implement an actor and a critic which learn sequences with the help of internal reinforcement signals calculated as predictions of future reinforcement and context inputs from the preSMA and SMA circuitry. Further, flexible shifting among various sequences requires contextual inputs from preSMA.

1 INTRODUCTION AND BACKGROUND

Serial order is an important aspect of human and animal behavior and continues to receive attention. In spite of research over many years, it is not yet completely understood how brain solves the serial order of motor behavior problem (Rosenbaum, 1991). However, in recent years there has been rapid progress in neuroscience, especially with the availability of newer and better technology.

In the present paper, we will review the experimental findings of Tanji & Shima (1994) on monkeys. The aim of the current work is to give a computational account of these experimental findings. In particular, the aim is to understand how a network of neuron-like elements can learn a sequence of actions using very coarse reinforcement signals and how in such a network context (short and long term) is represented. The structure of the network is designed based on known neuroanatomical information (Houk, Davis & Beiser, 1995) and insights from computational models of temporal processing in recurrent networks (Elman, 1990). Learning is implemented using biologically plausible reinforcement learning (RL) scheme (Barto & Jordan, 1987). To our knowledge, RL scheme has not been implemented in recurrent nets and we present a novel preliminary approach.

2 TANJI'S EXPERIMENTS

Tanji and colleagues devised a sequence learning paradigm and made single cell recordings in behaving monkeys while they performed sequential arm movements. The recordings have been made in various areas of the brain, including the basal ganglia (BG) which comprises the striatum (STR) and the globus pallidus (GP), the supplementary motor area (SMA), the pre-supplementary motor area (preSMA), the premotor cortex (PM), and the motor cortex (MI). After describing the experimental setup, a summary of results from various experiments will be given below.

2.1 EXPERIMENTAL SETUP

Monkeys were initially trained to respond with a push, pull or turn movement on a handle when presented with a red, yellow, or green color light, respectively, on a display panel. After establishing the color-to-action mapping, they were required to sequence these elementary actions by observing a succession of color lights and recalling the appropriate movement at every step. The trials when color lights preceded movements were called, 'Visual Trials (VT)' and the other trials when there were no colored lights were called 'Memory Trials (MT)'. In a block of trials, monkeys experienced 5 VT followed by 6 MT. They were taught four such sequences till they achieved 95% success rate on each sequence. After this stage, single cell recordings were made while monkeys performed several blocks of trials. Each block used a different sequence and the order of appearance of different sequences varied semi-randomly. The end of a block is signaled by random flashing of lights which also indicated a change in the sequence.

2.2 EXPERIMENTAL RESULTS

In the present study, we focus on results from the recordings of SMA, PreSMA, PM and parts of the BG. SMA recordings (Tanji & Shima, 1994) revealed a significant activation exhibiting two main properties: i) *sequence-specific activity*: activity of some cells preceded specifically before the performance of a particular sequence (say Push-Pull-Turn) and ii) *transition-specific activity*: some cells were tonically active after performing a specific movement and before initiating a particular next movement (say, Pull followed by Push). In contrast, the activity of majority of cells in the MI was selective to single movements but not sequence-specific. While SMA cells were mainly active in the MT phase, cells in PM were active only during the VT phase and cells in MI were active during all phases of the task. Some cells in preSMA were active only during the very first VT, possibly signifying the start of a new sequence and thus specifying the need for a change in motor plan (Shima *et al*, 1996).

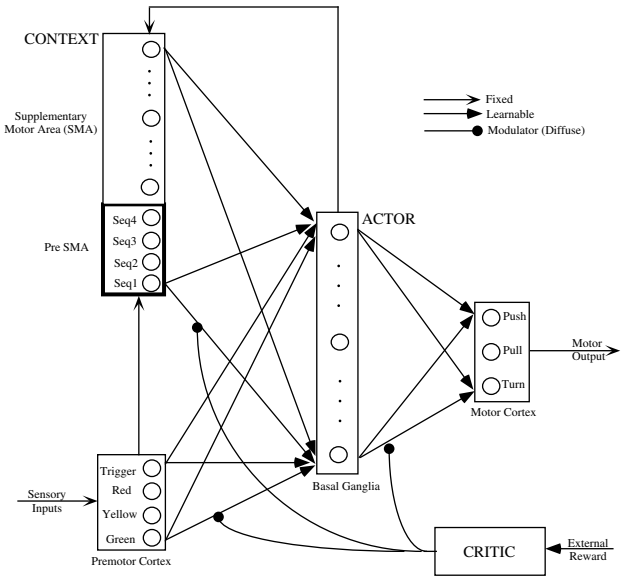

Figure 1: Elman-style Sequence Processing NN Architecture

Schultz *et al* (see Houk *et al*, 1995) reviewed, among other things, motor sequence-specific activity in putamen neurons. Putamen forms the input side of BG. GP forms the output side of BG and some neurons here showed more activity during the MT phase and specifically during particular transitions in some sequences or specifically to a particular sequence of arm movements (Mushiake & Strick, 1996). Some of these experimental findings will be used to formulate hypotheses about functional roles played by various areas involved in sequential arm movements. Further these hypotheses will be tested in simulations to see if the whole scheme works as expected.

3 NEURAL NETWORK ARCHITECTURE

Design hypotheses are listed and then a description of the network architecture and the correspondence between brain areas and network components is given.

3.1 DESIGN HYPOTHESES

It is hypothesized that BG learns by trial and error during the VT phase. Once learning is robust, network can recall the sequence during the MT phase also. In order to learn the sequential transitions, it is proposed that BG and cortex implement a recurrent network to learn the short term context of the ongoing sequence. Thus, given the previous movement, context from cortex enables BG to perform the correct next action. It is proposed that preSMA supplies long term context information to enable switching among multiple sequences.

3.2 DESCRIPTION OF THE ARCHITECTURE

Network comprises an input module (PM), an output module (MI), an action prediction module (Actor), a reward generation module (Critic), and a context module

(SMA and preSMA) as illustrated in figure 1.

Input Module: This module located in the premotor cortex takes in sensory input of color and trigger information (4 nodes). During VT, correct motor output is given by the mapping established between inputs and outputs via the hidden layer. However, during the MT only a fixed nonspecific trigger signal is fed as PM efferent signal and the correct movement has to be specified from elsewhere (using SMA and BG structures).

Actor: The hidden layer with 20 nodes forms the actor which is hypothesized to be in BG. Actor receives sensory inputs from input module and context information from context module. Given this information, it generates the correct movement at the output module. Hidden units have sigmoid transfer function and the learning rate for input-to-hidden layer weights is set at $\frac{1}{size(Inputlayer)}$.

Critic: In the present simulations a hand-crafted fixed critic based on the associative reward-penalty scheme of Barto and Jordan (1987). The implementation of an adaptive critic is left as future work. Critic generates reward signal (+1) if the move made is correct or punishment signal (-0.2), if the move is wrong. Trial is terminated upon a wrong move.

Context Module: As will be shown in the results section, context nodes situated in the SMA structures serve two fold purposes – they implement a i) short term (SMA layer with 16 nodes) and a ii)long term context enforcement (preSMA layer with 4 nodes, one for each sequence). Hidden unit activity from the previous time step is copied onto the context layer, with the exception of the preSMA context units. Only one of the preSMA units is turned on during a block and the same unit is turned on during all subsequent presentations of that sequence. Although currently the preSMA switching is done algorithmically, a scheme with two-stage competition has been devised successfully (results not reported here).

Output Module: Output module consisting of 3 nodes, one for each action (Push, Pull Turn), is located in the motor cortex. It receives BG output from the hidden layer and generates a motor movement by using a winner-take-all strategy. A softmax transfer function is used for output nodes and the output layer also has bias nodes. Learning rate for hidden-to-output weights and bias weights is set at $\frac{1}{size(hiddenlayer)}$ and that of context-to-hidden layer weights at $\frac{3}{size(contextlayer)}$.

4 SIMULATION RESULTS

Monkeys had extensive training in order to establish the color-to-action mapping before starting the sequence task. Accordingly the model is initially trained on single color inputs and made to learn the corresponding action. These trials termed, 'Stimulus- Response (SR)' trials, are given before sequence training and a good mapping is established in 5-10 SR trials. In order to aid the correct development of the mapping, initial weights between input-to-hidden layer and hidden-to-output layer are set randomly but with an optimal transformation. As a result of the preparatory training and due to the availability of visual inputs, superior performance is seen during VT phase. The aim of the network is also to recall sequences accurately in the MT phase when visual guidance is absent.

4.1 LEARNING OF SEQUENCES

Simulations closely followed the event sequence of the actual experiment. At first a set of three sequences (say, Seq1, Seq2, Seq3) is chosen from a pool of four sequences

used in Tanji's experiment. Different orderings of items are denoted here as – Seq1:123, Seq2:132, Seq3:312, and Seq4:321. Network was trained with one sequence (say, Seq1) per block for 400 blocks. In each block a sequence is chosen randomly from the three sequence-set and the same set of three sequences (say, Seq1, Seq2, Seq3) is used throughout a simulation. Each block consisted of 5 VT followed by 6 MT. During each VT, color and trigger inputs are given and the network is required to generate the appropriate action. A 3-bit binary vector is used as color input, with each bit representing one of the three colors (red, yellow, or green). Trigger signal is modeled as a fixed signal (+1) injected nonspecifically to all the hidden units. Motor output at MI is also 3-bit binary vector, with each bit representing one of the three actions (push, pull, or turn). During the MT, input consisted of only a nonspecific trigger signal presented successively and the network is required to recall appropriate actions at each time step in the correct order. All the node activities are reset to 0 at the beginning of each trial.

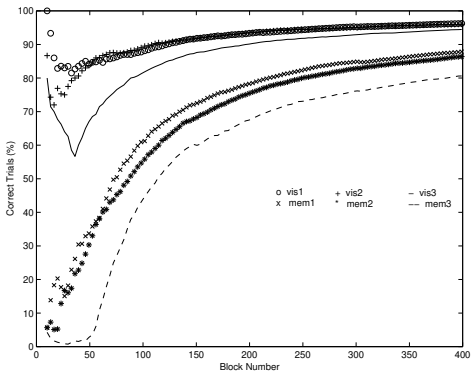

Figure 2: Learning Curve in a typical simulation with random presentation of three Sequences (Seq1, Seq2,and Seq3), one per block.

Figure 2 shows the learning profile of the network during the simulation with sequences (Seq1, Seq2, Seq3). Each data point is obtained by calculating the correct number of VT and MT in the preceding ten blocks of trials. Performance for each sequence is shown individually. It can be seen that performance on VT is far superior throughout as the mapping due to the availability of visual guidance signals. Performance on MT increases and reaches 60% level in 200 blocks on all the sequences. Four such simulations, each one using a different set of three sequences, are performed and the overall averaged results are shown in figure 3. Overall learning profile is similar to that seen in single simulation results. It is also observed that without the context switching scheme we described above operating in the preSMA module, network had great difficulty learning multiple sequences (results not shown here).

4.2 Sequence and Transition Specific Activity

In this section, evidence is presented for sequence and transition specific activity developing in the hidden layer of the network as a result of training. Figure 4 shows activity of a particular node (unit 5) that increases with training specifically for sequences 2 and 3. Activity in response to sequence 4 gets suppressed. On the other hand, figure 5 shows activity of another unit (20) which is transition specific – from move3-to-move2. Both the sequences 2 and 3 share this particular transition.

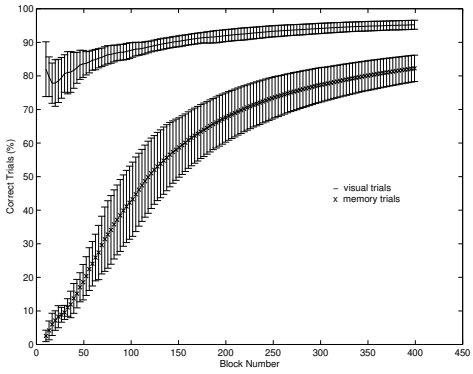

Figure 3: Overall Learning Profile (average and standard deviation over 4 simulations)

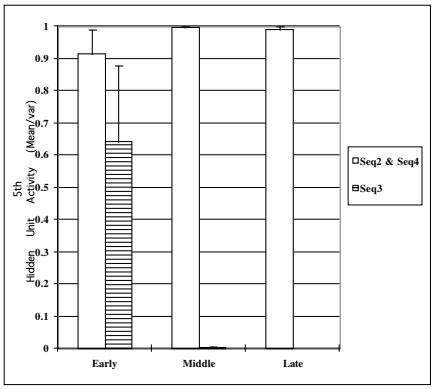

Figure 4: Activity of 5th Hidden Unit Showing Specific Activity for Sequences 2 and 3 only.

It is also observed that transition specificity develops at a much faster time course than sequence specificity and that is why the unit 20's activity is already high in the early stage. Performance of the network on sequences 2 and 4 increases from 70% in the early stage (blocks 50-70), to 96% in the middle stage (blocks 150-170), and to 97% in the late stage (blocks 380-400). As can be seen from figure 4 that this is in pace with unit 5's acquisition of sequence specificity. Further, it is observed that errors can be reliably predicted from activity of these units – inactivation leads to errors (results not shown here).

5 DISCUSSION AND CONCLUSIONS

The proposal that BG and cortex recurrent loops learn and perform sequences successfully by RL has been demonstrated in simulations. The novelty here is the training of Elman-style network with RL. The chaining of successive movements is enabled by context inputs from SMA structures. It has been demonstrated that SMA structures enable switching among multiple sequences. To our knowledge,

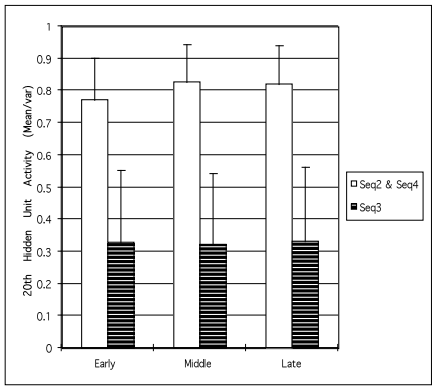

Figure 5: Activity of 20th Hidden Unit Showing Transition Specific Activity for Sequences 2 and 3 only.

this is first attempt at a computational account of tanji's findings.

In the simulations reported here switching is done externally by setting specific preSMA nodes and hence is not adaptive. The selection of a context node by actively monitoring the sequence performed during VT will be explored in future. Also, a fixed critic is implemented in the current work and an adaptive critic implementation in recurrent nets is left for future work.

References

Barto, A. G., & Jordan, M. I. "Gradient following without back-propagation." In M. Caudill & C. Butler (eds.), *IEEE First International Conference on Neural Networks*, San Diego, CA, June 21-24, **vol.II**, pp.629-636 (1987).

Elman, J. L. "Finding structure in time," *Cog. Sci.*, **14**, pp.179–211 (1990).

Houk, J. C., Davis, J. L., & Beiser, D. G. (Eds.) *Models of information processing in the basal ganglia*, Cambridge, MA: MIT Press, 1995.

Mushiake, H. & Strick, P. L. "Pallidal neuron activity during sequential arm movements," *J. Neurophys.*, **74**, 6, pp.2754-2758 (1996).

Rosenbaum, D. A., *Human Motor Control*, San Diego, CA: Academic, 1991.

Schultz, W., Dayan, P., & Montague, P.R. "A neural substrate of prediction and reward," *Science*, **275**, pp.1593–1599 (1997).

Shima, K., Mushiake, H., Saito, N., & Tanji, J. "Role for cells in the presupplementary motor area in updating motor plans," *Proc. Natl. Acad. Sci. USA*, **93**, pp.8694-8698 (1996).

Tanji, J. & Shima, K. "Role for supplementary motor area cells in planning several movements ahead," *Nature*, **371**, 29, pp.413–416 (1994).

[Submitted for NIPS*98, Denver, CO, USA, Nov 30 – Dec 5, 1998]