REINFORCEMENT DRIVEN DIMENSIONALITY REDUCTION - A MODEL FOR INFORMATION PROCESSING IN THE BASAL GANGLIA

Izhar Bar-Gad[1] Eytan Ruppin[2] and Hagai Bergman[1,3]

[1] Center for Neural Computation, Hebrew University, Jerusalem, ISRAEL

[2] Departments of Computer Science and Physiology, Tel-Aviv University, Tel-Aviv, ISRAEL

[3] Department of Physiology, Hebrew University-Hadassah Medical School, Jerusalem, ISRAEL.

Correspondence To:

Izhar Bar-Gad

Department of Physiology

The Hebrew University – Hadassah Medical School

P.O. Box 12272

Jerusalem, ISRAEL 91120

Tel: 972-2-6757388

Fax: 972-2-6439736

E-mail: izharb@alice.nc.huji.ac.il

The basal ganglia are part of a loop connecting the entire cortex to the frontal cortex. Despite a large body of clinical and experimental data, the processing they perform remains obscure. Recent physiological findings of uncorrelated activity of basal ganglia neurons contradict available anatomical data showing the existence of extensive convergence and lateral inhibitory connections. This discrepancy leads us to propose that the basal ganglia use a combination of unsupervised and reward driven learning to reduce the dimensionality of cortical information. Simulations implementing key aspects of the cortico-striato-pallidal circuitry predict that during learning the efficacy of the lateral synapses diminishes and neural activity becomes uncorrelated. Through this process the basal ganglia achieve efficient extraction of cortical information which is probably used by the frontal cortex for planning upcoming action.

The critical role played by the basal ganglia in the pathogenesis of various movement disorders such as Parkinson's and Huntington's diseases has been known for many years [1]. Later research has indicated that the basal ganglia participate in everyday complex behaviors that require coordination between cognition, motivation and movements [2]. The basal ganglia are comprised of many nuclei with complex interactions between the neurons within each nucleus and between the nuclei [3]. A major pathway in the basal ganglia circuitry is from most cortical areas to the striatum and then to the output stages of the basal ganglia, e.g., the internal segment of the globus pallidus (GPi). The GPi output projects back to the frontal cortex through thalamic relay stations (Fig. 1a). This pathway is characterized by a high degree of anatomical convergence. The number of cortical neurons projecting to the striatum is two orders of magnitude greater than the number of striatal neurons [3] and an additional decrease of the same magnitude occurs from the striatum to the GPi [4]. Although quantitative studies of the increase in the neuronal populations at the pallido-thalamic and thalamo-cortical levels are still lacking, most anatomical studies indicate that the bottle-neck is at the pallidal level [5,6].

The GABAergic projection neurons compose the vast majority of the basal ganglia neurons [3]. The GABAergic projections also form massive collateral anatomical connections in the striatum [3] and in the pallidum [7,8]. However, the prediction of strong collateral inhibition has been thwarted by recent physiological studies in which no evidence was found for functional synaptic interactions between striatal neurons [9]. Cross-correlation tests of the firing synchrony in the striatum [10,11] and pallidum [12] have also revealed no correlation

between the recorded neurons in contrast to the generally correlated activity of neighboring neurons in the cortex [13].

This anatomical and physiological data enforces three major constraints on any model of basal ganglia function: Massive stepwise funneling architecture, extensive network of lateral inhibitory connections devoid of functional significance, and the absence of intra-nuclear correlation of striatal and pallidal firing. Previous models of the basal ganglia have been motivated primarily by the anatomical evidence of the strong lateral connectivity to assume mutual inhibition between striatal neurons [14], and view the basal ganglia as an action selection network governed by mutual inhibitory domains [15-17]. Their key idea being that the cortex generates many possible actions and that the basal ganglia select a single action among those actions. However, these models do not incorporate the recent physiological data concerning the intra-nuclear interactions, and erroneously predict strong lateral interactions and negative correlation of intra-nuclear neuronal firing. Moreover they do not address the massive stepwise funneling structure typical to the basal ganglia.

In this study we report on a new hypothesis that explains the above-mentioned discrepancies. The hypothesis is based on the proposal that the basal ganglia perform dimensionality reduction of the large and complex information space spanned by the activity of cortical neurons. We show that a neural network featuring key aspects of the known anatomical and physiological facts that characterize the basal ganglia is ideally suited to perform such a process. Finally, we use this computational model to provide explicit predictions for future in-vitro and in-vivo experiments.

The term dimensionality reduction describes the process of projecting inputs from a high dimensional data space to a considerably smaller one. Efficient reduction is achieved when all or most of the information contained within the original space is preserved. Dimensionality reduction in the nervous system can be depicted as compression of the information encoded by a large neuronal population to a smaller number of neurons. This process is eminently useful because it allows the transmission of large amounts of information within a limited number of axons. Further, the process allows exposure of neurons in the final (output) layer to maximal incoming information using the anatomically limited number of synapses that each neuron can receive. Dimensionality reduction has been proposed as an important part of processing in sensory systems, enabling efficient identification by removal of redundancies in the sensory inputs. A classical method for performing dimensionality reduction is principal component analysis (PCA). PCA retrieves

the orthogonal axes of the input data that display maximal variance and can be used to form its best linear approximation. Theoretical studies demonstrate that neural networks can extract principal components using a competitive Hebbian learning rule and lateral inhibitory connectivity [18-20]. As will be shown the properties of these networks closely resemble the functional properties of the basal ganglia network.

To examine the hypothesis that the basal ganglia circuitry performs dimensionality reduction, we studied a simulated neural network consisting of a feed-forward network of neurons with linear activation function and lateral connectivity within the layers (Fig. 1b). Learning is Hebbian for the feed-forward weights and anti-Hebbian for the lateral weights. The synaptic weights are constrained according to the known physiology and anatomy of the basal ganglia. Thus, positive weights are applied at glutamatergic synapses and negative values at GABAergic ones. The network is presented with a series of input patterns. The elements of each pattern are correlated (see Methods section) modeling cortical population vector inputs [21]. Initially, the network performs sub-optimal information compression and the output neurons are correlated. This correlation causes an increase in the absolute value of the efficacies of the inhibitory lateral synapses (anti-Hebbian learning, Fig. 2a) and changes in the efficacies of the feed-forward connections (according to Hebbian learning rules, Fig. 2b). These changes, in turn, result in decorrelation of neuronal activity within the output layer (Fig. 2c) and improvement in information compression (Fig 2d). Overall, the modification of the neural circuit leads to the formation of optimal information compression and manifests important features of the basal ganglia: uncorrelated activity of the output neurons and a diminished efficacy of lateral synaptic interactions.

Dimensionality reduction in a behaving animal should be affected not only by the statistical properties of the input patterns but also by their behavioral significance. The relative significance of the input is determined by its ability to predict reward to the animal. Performing such reward-related dimensionality reduction in the basal ganglia can be achieved through interaction of the aforementioned feed-forward network with a reinforcement signal. A candidate signal is received in the basal ganglia from striatal cholinergic interneurons [2] and from dopaminergic neurons in the substantia nigra pars compacta (SNc) which are known to respond to reward-related events [22]. The SNc terminals in the striatum are part of a complex architecture comprised of a cortical glutamatergic projection which terminates on the head of a dendritic spine of a striatal projection neuron, and a nigro-striatal dopaminergic synapse located on the neck of the same dendritic spine

[23]. Through this complex synaptic structure the reinforcement modulates the access of striatal neurons to cortical inputs [24]. The mechanism of the reinforcement signal formation and its interaction with the cortico-striatal transmission has been modeled extensively [14] to explain the role of the basal ganglia in the expression of learned responses. We added these reward-related properties to the model to generate Reinforcement Driven Dimensionality Reduction (RDDR). The simulation utilized a multi-layer feed-forward network similar to the one described previously, but now the feed-forward network interacts with a reinforcement signal provided at the intermediate (striatal) layer. The extraction becomes discriminative, performing better for reward related inputs and worse for events not related to reward prediction (Fig. 3).

The RDDR mechanism also offers new explanations to some open questions in the pathophysiology of movement disorders, especially Parkinson's disease. The model clarifies why the effects of focal lesions in the normal basal ganglia are minimal [25] while the effects of abnormal levels of dopamine are overwhelming. In response to local lesions the network adapts and reorganizes its connections losing only the minor components while maintaining the principal ones, thereby minimizing the information loss (Fig. 4a). On the other hand dopamine depletion (a negative reinforcement signal [22]), as in Parkinson's disease, substantially damages the RDDR process since no discrimination is possible between important and negligible information (Fig. 4b). Conventional dopamine replacement therapy restores the background level of dopamine. However, the pulsatile nature of the treatment causes inevitable random fluctuations in dopamine levels in the striatum resulting in the generation of random encoding and the development of dyskinesia.

The RDDR model emphasizes the role of the basal ganglia in extraction and pre-processing of information from the whole cortex. It provides an interesting explanation for the apparent lack of evident physiological function of the lateral inhibitory connections and uncorrelated activity observed in the previous studies of the basal ganglia. These finding were observed in studies carried out in adult animals that were not engaged in learning of new skills and situations, where the RDDR model maintains that the lateral connections are functional only during the learning phase, after which their efficacy vanishes. The changes in the lateral connections are augmented by the parallel changes in the feed-forward projections to cause the output to become uncorrelated.

Major facts that were not implemented yet in the basic RDDR model can be integrated into it. Neurons in GPi are almost linear in their I-f (input current/firing rate) curve

[26], whereas striatal [27] and thalamic [28] projection neurons are highly non-linear in their response. Interestingly, multi-layer PCA networks containing two such non-linear layers and a bottleneck linear layer perform better information extraction than linear PCA [29]. The complexity of the model is further increased since the cortico-basal ganglia-cortico circuit is not merely a feed forward network but a partially closed loop. Major sources of input to the striatum are the intralaminar thalamic nuclei and the frontal cortex that receive basal ganglia output. We hypothesize that the basal ganglia perform dimensionality reduction of the cortical neural activity representing the present state of the animal. The reduced information is projected to the frontal cortex that uses it for planning future actions. This recurrent processing may therefore explain the major role of the basal ganglia circuitry in sequential behavior [30].

The RDDR model, like any model, encompasses only those elements that are assumed to be the most significant to the function of the actual neural network. However, the model provides a novel perspective on basal ganglia function, suggesting answers to fundamental questions in the field and yielding specific testable predictions. From a high-level functional perspective, the RDDR model has two main advantages: It enables the efficient transfer of information from all over the cortex to executive regions in the frontal cortex via a very small number of connections. It also provides a vehicle by which reinforcement learning (requiring a complex tri-synaptic structure to carry out the multi-Hebbian learning rules involved) may be carried out in the brain in a central, parsimonious location. The model predicts an increase in correlated activity and increased synaptic efficacies of lateral connections within the basal ganglia during periods of network reorganization. Such reorganization occurs in young animals, in adult animals following focal lesions or in animals during intensive learning periods. The model can also be tested in vitro by examining the pertaining cellular learning rules and the effects of dopamine upon them. We believe that combining these future experiments with further theoretical insights will shed new light on the basic functions of the basal ganglia in health and disease.

Methods

The simulations are based on a feed-forward neural network for performing PCA using lateral inhibition [19,20]. The neural network is comprised of three layers: the first layer representing the cortical input, an intermediate layer corresponding to the striatum and an output layer representing the GPi (Fig. 1b). The network weights are constrained to either positive or negative values to reflect the known neurotransmitter physiology. The feed-

forward weights between the input and intermediate layers are limited to positive values (corresponding to glutamatergic excitatory synapses). The feed-forward weights between the intermediate and output layers and the lateral weights in both processing layers are constrained to negative values (corresponding to GABAergic inhibitory synapses). To implement reinforcement driven dimensionality reduction, a reinforcement signal is combined with the feed-forward input at the intermediate layer to create a multi-Hebbian learning rule, simulating the complex cortico-SNc-striatal synapses. The reinforcement signal is positive for reward-related events and zero for non reward-related events (baseline dopamine levels). Dopamine depletion as in Parkinson's disease is modeled by negative reinforcement values.

Each input example pattern (*c*) is an N dimensional real valued vector. Each element of the input vector is generated from a linear transformation of a K dimensional source vector (K<N), whose components are normally distributed i.i.d. variables. This procedure generates N dimensional input patterns that essentially lie within a K dimensional subspace, thus enabling the complete reduction of the inputs into a K dimensional output space. For the study of selective reinforcement in the RDDR model, several distinct subsets of input patterns were generated. Each subset is characterized by a distinct source to input transformation matrix, receiving a specific reward level.

The intermediate layer activity (*s*) is

$$(1)\quad s_i = \sum_{j=1}^{N} w_{ij} \cdot c_j + \sum_{j=1}^{M} a_{ij} \cdot s_j \ .$$

The learning rule for the feed forward weights (*w*) between the input and the intermediate layer is a competitive multi-Hebbian rule, combining feed-forward and reinforcement signals

$$(2)\quad \Delta w_{ij} = \eta \cdot r \cdot [s_i \cdot c_j - s_i^2 \cdot w_{ij}] \quad , \qquad w_{ij} \geq 0$$

and the learning rule for the intermediate layer lateral weights *(a)* is a competitive anti-Hebbian rule

$$(3)\ \Delta a_{ij} = -\eta \cdot [s_i s_j + s_i^2 \cdot a_{ij}] \qquad , \qquad a_{ij} \leq 0 \qquad a_{ii} = 0 \, .$$

The output layer activity (*g*) is

$$(4)\quad g_i = \sum_{j=1}^{M} u_{ij} \cdot s_j + \sum_{j=1}^{K} b_{ij} \cdot g_j \ .$$

The learning rule for the feed forward weights (*u*) between the intermediate and the output layer is a competitive Hebbian rule

(5) $\Delta u_{ij} = \eta \cdot [g_i \cdot s_j - g_i^2 \cdot u_{ij}] \quad , \quad u_{ij} \le 0$

and the learning rule for the output layer lateral weights *(b)* is competitive anti-Hebbian rule, analogous to (3)

(6) $\Delta b_{ij} = -\eta \cdot [g_i g_j + g_i^2 \cdot b_{ij}] \quad , \quad b_{ij} \le 0 \quad b_{ii} = 0.$

To measure the information loss of the network due to the RDDR process, the K dimensional representation of the input patterns in the output layer is expanded back to a N dimensional space to create the reconstructed, decompressed pattern. The reconstruction error is the mean squared difference between the original and reconstructed elements over all input patterns.

The simulations employ an input layer of 16 neurons, an intermediate layer of 8 neurons and an output layer of 4 neurons. The learning rate (η), for both feed-forward and lateral weights is 0.0002. The reinforcement signal was 1 for positive reinforcement and -0.1 for negative reinforcement. The feed-forward weights were initialized to random values and the lateral weights were initialized to zero.

(a) (b)

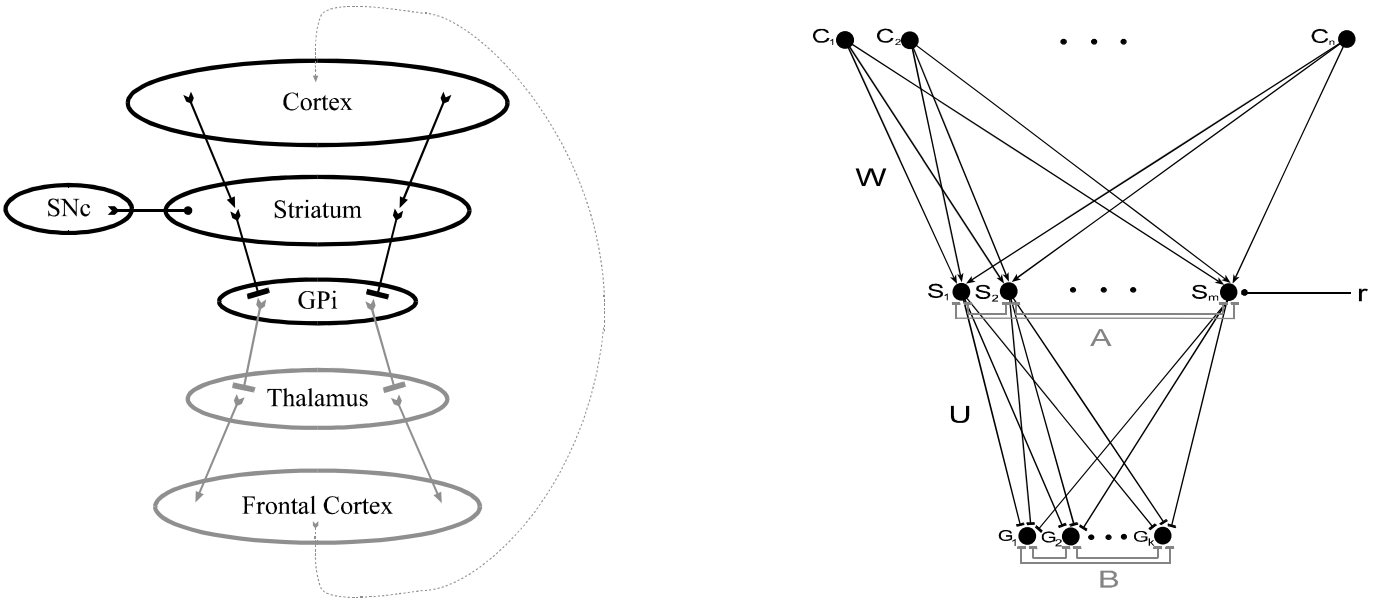

Figure 1. **Structure of the reinforcement driven dimensionality reduction network**

(a) Schematic diagram of the cortico-basal ganglia-cortical circuit (black - layers incorporated in the model, gray- layers not included in the model).

(b) The model is composed of a three-layered feed-forward network simulating the cortico-striato-pallidal circuit with lateral inhibitory connections at the intermediate (striato) and output (pallidal) layers. A reinforcement signal is provided at the intermediate layer.

Arrow-head connections represent glutamatergic excitatory synapses, square-head connections represent GABAergic inhibitory synapses and round-head connections represent dopaminergic synapses.

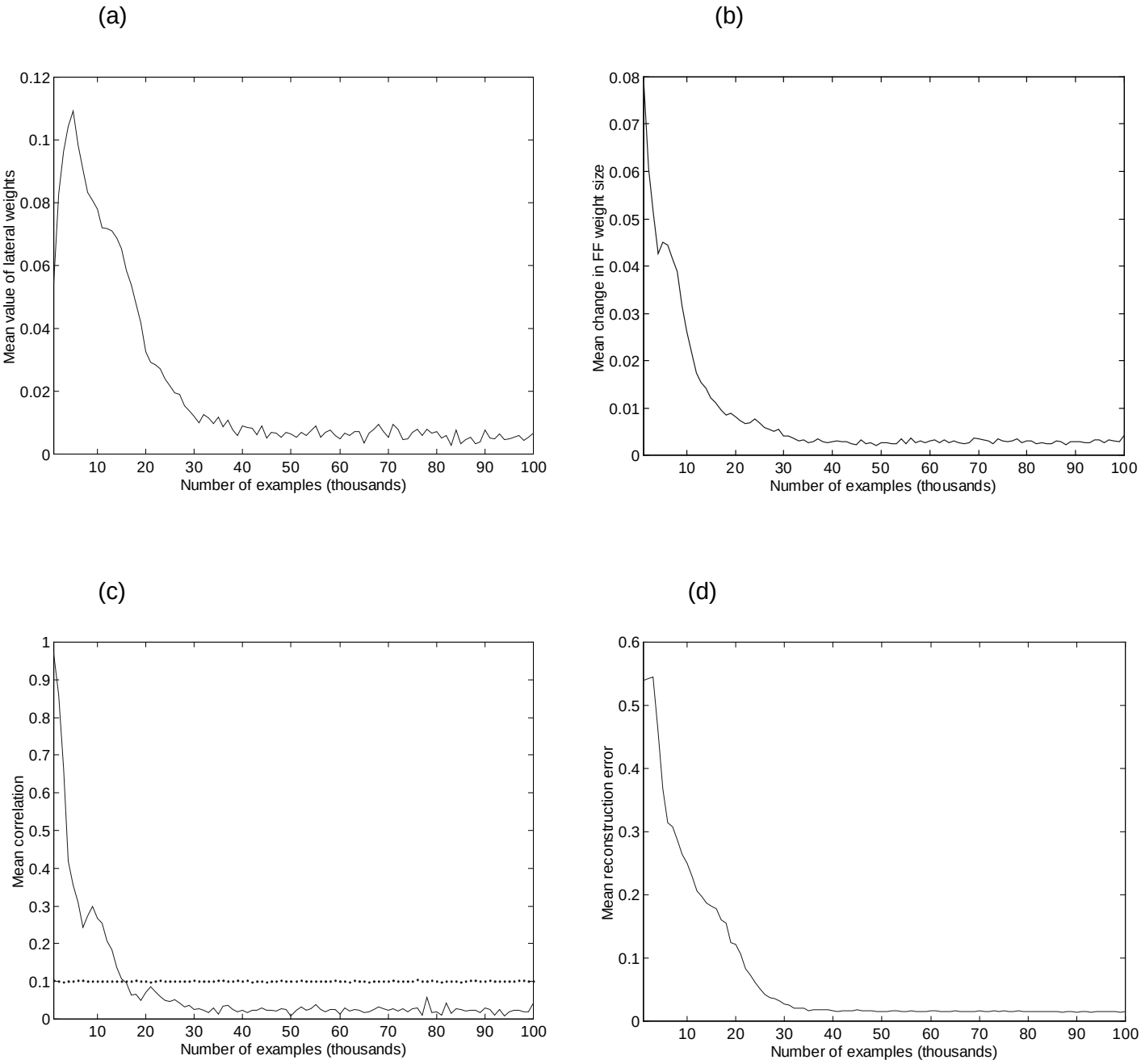

Figure 2. **The learning phase of the dimensionality reduction network.**

The following variables are displayed as a function of the number of input examples presented to the network.

(a) Mean value of the lateral weights of the output (pallidal) layer.

(b) Mean change in the values of feed forward intermediate to output (striato-pallidal) weights.

(c) Correlation between neurons of the output (pallidal) layer (solid line) and correlation of the neurons of the input (cortical) layer (dotted line).

(d) Information loss due to the compression (reconstruction error).

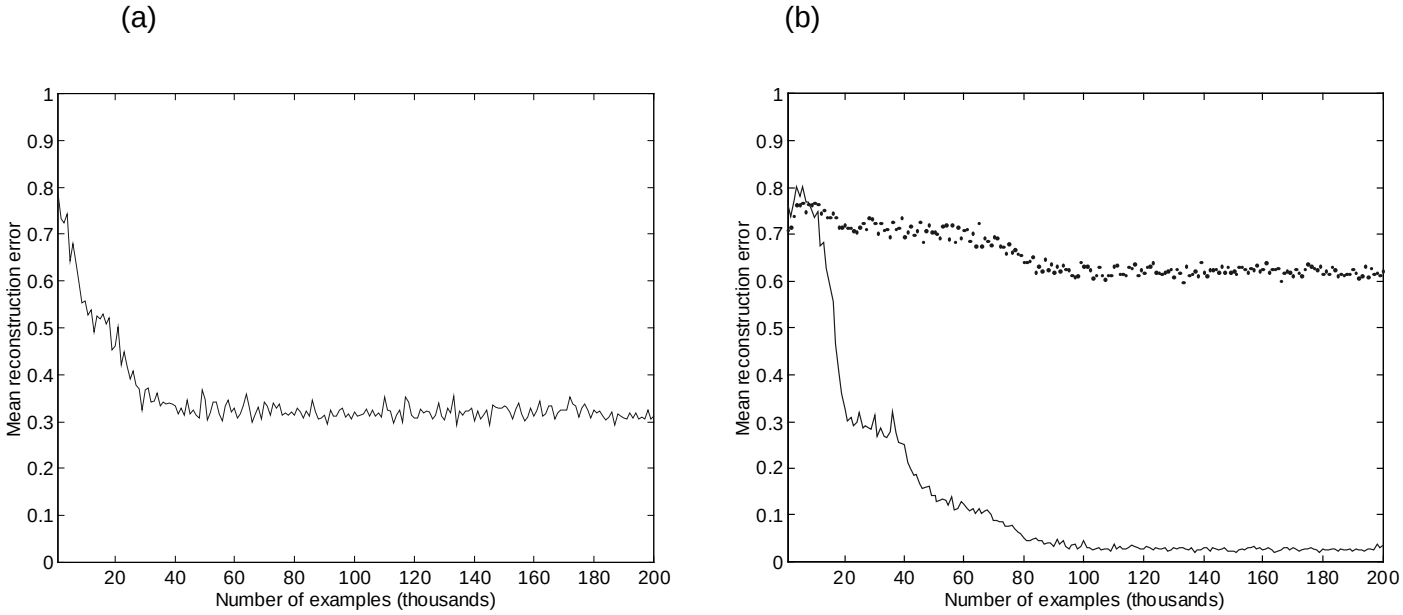

Figure 3. **Selective reinforcement driven dimensionality reduction.**

The mean reconstruction error of input patterns in a network receiving inputs from four different sources.

(a) All patterns receive equal reinforcement.

(b) Selective reinforcement. Reinforced patterns (solid line) and non-reinforced patterns (dotted line).

(a) (b)

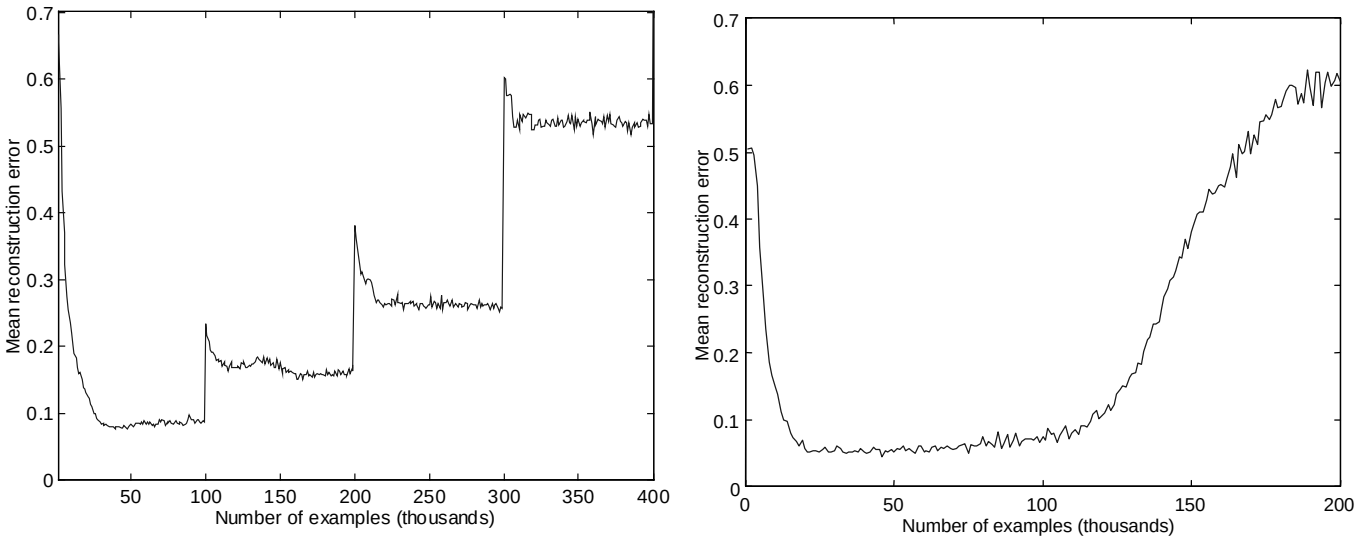

Figure 4. **Effects of simulated lesions and dopamine depletion on information compression.**

(a) Effects of focal lesions in the output layer (GPi) on the reconstruction error. 25% of the neurons are removed after every 100,000 examples. Each lesion is followed by a sharp increase in the reconstruction error that decreases gradually as the network rearranges to extract the principal components. The increase in the steady state level of the reconstruction error is hyper-linear since the minor components are the first to be lost.

(b) Effects of dopamine depletion on the reconstruction error. The network receives constant negative reinforcement after 100,000 examples, causing a loss of its compression capabilities in an increasing manner.

Reference List

1. Wilson, S.A.K. *Brain* **36**, 427-492 (1914).

2. Graybiel, A.M., Aosaki, T., Flaherty, A.W. & Kimura, M. *Science* **265** , 1826-1831 (1994).

3. Gerfen, C.R. & Wilson, C.J. in *Handbook of Chemical Neuroanatomy, Vol 12: Integrated Systems of the CNS, Part III* (eds Swanson, L.W., Bjorklund, A. & Hokfelt, T.) 371-468 (Elsevier Science, 1996).

4. Percheron, G., Francois, C., Yelnik, J., Fenelon, G. & Talbi, B. in *The basal ganglia IV* (eds Percheron, G., McKenzie, J.S. & Feger, J.) 3-20 (Plenum Press, New York, 1994).

5. Arecchi Bouchhioua, P., Yelnik, J., Francois, C., Percheron, G. & Tande, D. *.Neuroreport.* **7**, 981-984 (1996).

6. Sidibe, M., Bevan, M.D., Bolam, J.P. & Smith, Y. *.J.Comp.Neurol.* **382**, 323-347 (1997).

7. Kita, H. in *The Basal Ganglia V* (eds Ohye, C., Kimura, M. & McKenzie, J.S.) 77-94 (Plenum Press, New York, 1996).

8. Yelnik, J., Francois, C. & Tand, D. *3rd Congress of European Neuroscience Society.Bordeaux* 104 (1997).(Abstract)

9. Jaeger, D., Kita, H. & Wilson, C.J. *J.Neurophysiol.* **72**, 2555-2558 (1994).

10. Jaeger, D., Gilman, S. & Aldridge, J.W. *.Brain Res.* **694**, 111-127 (1995).

11. Stern, E.A., Jaeger, D. & Wilson, C.J. *Nature* **394**, 475-478 (1998).

12. Bergman, H., Feingold, A., Nini, A., et al. *TINS* **21**, 32-38 (1998).

13. Eggermont, J.J. *The Correlative Brain. Theory and experiment in neuronal interaction* (Springer-Verlag, Berlin, 1990).

14. Beiser, D.G., Hua, S.E. & Houk, J.C. *Current Opinion in Neurobiology* **7**, 185-190 (1997).

15. Wickens, J. *A Theory of the Striatum* (Pregamon Press, Oxford, 1993).

16. Mink, J.W. *Prog.Neurobiol.* **50**, 381-425 (1996).

17. Berns, G.S. & Sejnowski, T.J. *J.Cogn.Neurosci.* **10**, 108-121 (1998).

18. Oja, E. *J.Math.Biol.* **15**, 267-273 (1982).

19. Foldiak, P. *Proceedings, International Joint Conference on Neural Networks* **1**, 401-405 (1989).

20. Kung, S.Y. & Diamantaras, K.I. *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing* **2**, 861-864 (1990).

21. Georgopoulos, A.P. *Trends.Neurosci.* **18**, 506-510 (1995).

22. Schultz, W. *J.Neurophysiol.* **80**, 1-27 (1998).

23. Freund, T.F., Powell, J.F. & Smith, A.D. *Neuroscience* **13**, 1189-1215 (1984).

24. Calabresi, P., Pisani, A., Mercuri, N.B. & Bernardi, G. *Trends.Neurosci.* **19**, 19-24 (1996).

25. Kato, M. & Kimura, M. *J.Neurophysiol.* **68**, 1516-1534 (1992).

26. Kita, H. & Kitai, S.T. *Brain Res.* **564**, 296-305 (1991).

27. Nisenbaum, E.S. & Wilson, C.J. *J.Neurosci.* **15**, 4449-4463 (1995).

28. Steriade, M., McCormick, D.A. & Sejnowski, T.J. .*Science* **262**, 679-685 (1993).

29. Oja, E. in *Artificial Neural Networks* (eds Kohonen, T., Makisara, K., Simula O. & Kangas J.) 737-745 (1991).

30. Aldridge, J.W. & Berridge, K.C. *J.Neurosci.* **18**, 2777-2787 (1998).

Acknowledgements

This study was supported in part by the Israeli Academy of Science, AFIRST, Alon Fellowship and the US-Israel Bi-national Science Foundation. We thank Thomas Wichmann, Eilon Vaadia and James Reggia for their critical reading and helpful suggestions.