Progress in Neurobiology 103 (2013) 214–222

Contents lists available at SciVerse ScienceDirect

Progress in Neurobiology

journal homepage: www.elsevier.com/locate/pneurobio

From fixed points to chaos: Three models of delayed discrimination

Omri Barak [a,*], David Sussillo [b], Ranulfo Romo [c,g], Misha Tsodyks [d,a], L.F. Abbott [a,e,f]

[a] Center for Theoretical Neuroscience, Columbia University, New York, NY 10032, USA
[b] Department of Electrical Engineering, Neurosciences Program, Stanford University, Stanford, CA 94305, USA
[c] Instituto de Fisiología Celular-Neurociencias, Universidad Nacional Autónoma de México, 04510 México, D.F., Mexico
[d] Department of Neurobiology, Weizmann Institute of Science, Rehovot 76100, Israel
[e] Department of Neuroscience, Columbia University, New York, NY 10032, USA
[f] Department of Physiology and Cellular Biophysics, Columbia University, New York, NY 10032, USA
[g] El Colegio Nacional, 06020 México D.F., Mexico

ARTICLE INFO

Article history:
Received 6 May 2012
Received in revised form 23 November 2012
Accepted 7 February 2013
Available online 21 February 2013

Keywords:
Working memory
Neural networks
Model
Nonlinear dynamics

ABSTRACT

Working memory is a crucial component of most cognitive tasks. Its neuronal mechanisms are still unclear despite intensive experimental and theoretical explorations. Most theoretical models of working memory assume both time-invariant neural representations and precise connectivity schemes based on the tuning properties of network neurons. A different, more recent class of models assumes randomly connected neurons that have no tuning to any particular task, and bases task performance purely on adjustment of network readout. Intermediate between these schemes are networks that start out random but are trained by a learning scheme. Experimental studies of a delayed vibrotactile discrimination task indicate that some of the neurons in prefrontal cortex are persistently tuned to the frequency of a remembered stimulus, but the majority exhibit more complex relationships to the stimulus that vary considerably across time. We compare three models, ranging from a highly organized line attractor model to a randomly connected network with chaotic activity, with data recorded during this task. The random network does a surprisingly good job of both performing the task and matching certain aspects of the data. The intermediate model, in which an initially random network is partially trained to perform the working memory task by tuning its recurrent and readout connections, provides a better description, although none of the models matches all features of the data. Our results suggest that prefrontal networks may begin in a random state relative to the task and initially rely on modified readout for task performance. With further training, however, more tuned neurons with less time-varying responses should emerge as the networks become more structured.

Contents

Abbreviations: PFC, prefrontal cortex; LA, line attractor model; RN, random network model; TRAIN, trained network model.

* Corresponding author. Present address: Rappaport Faculty of Medicine, Technion – Israeli Institute of Technology, Haifa, Israel.
E-mail address: omri.barak@gmail.com (O. Barak).

0301-0082/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.pneurobio.2013.02.002

1. Introduction

Working memory is used to hold and manipulate items mentally for short periods of time, which is crucial for many higher cognitive functions such as planning, reasoning, decision-making, and language comprehension (Baddeley and Hitch, 1974; Baddeley, 1986; Fuster, 2008). Lesion and imaging studies have identified the prefrontal cortex (PFC) as an essential area for working memory performance. To explore the neural underpinnings of this facility, experimental paradigms have been developed to record neural activity while monkeys performed working-memory tasks, among them delayed discrimination. In these experiments, monkeys have to retain the memory of a briefly presented first stimulus (visual image, location of the target, etc.) during a delay period of several seconds in order to perform a comparison with a subsequently presented stimulus. A key observation was the discovery of neurons in several cortical areas, including PFC, that exhibit stimulus specific persistent firing activity during the delay when no stimulus is present (Fuster and Alexander, 1971; Miyashita and Chang, 1988; Funahashi et al., 1989, 1990; Romo et al., 1999, 2002). It is commonly believed that this persistent selective activity maintains the memory of the stimulus.

Because no stimuli are presented during the delay, persistent activity must be internally generated. A common theoretical framework for this is the attractor neural network, which exhibits many intrinsically stable activity states sustained by mutual excitation between neurons coding for a particular stimulus or its behaviorally relevant attribute (Hebb, 1949; Hopfield, 1982; Amit and Brunel, 1997; Seung, 1998; Wang, 2001, 2009). When a stimulus is briefly presented, the corresponding attractor is evoked and remains active until the behavioral task is performed and the network returns to its baseline state. In this way, the neuronal activity encodes a memory trace during the delay.

If the features kept in working memory are of a discrete nature, such as one of a collection of visual objects, the paradigmatic network is of the Hopfield type (Hopfield, 1982) with a discrete set of attractors. If the features are continuous, such as the spatial location of a stimulus, the network dynamics should possess a continuous set of attractors (Ben-Yishai et al., 1995; Seung, 1998). In both situations, connections in the network have to be chosen as a function of the selectivity properties of pre- and postsynaptic neurons (e.g. increased mutual excitation between neurons with similar tuning properties). Because the attractor states of the network are stationary, the corresponding neural selectivity to stimulus features is also stationary over the delay period.

Maintaining the information about stimulus attributes with stationary persistent activity appears to be a natural and robust mechanism of working memory (see e.g. Wang, 2008). However, a closer look at experimental recordings reveals much greater variability in neuronal response properties than can be accounted for by standard attractor neural networks. In particular, a majority of the cells exhibit firing frequency and selectivity profiles that vary markedly over the course of the delay period (see e.g. Brody et al., 2003; Shafi et al., 2007). These observations indicate that elucidating the neuronal mechanisms of working memory is still an open issue requiring further experimental and theoretical research.

In this contribution, we consider a tactile version of the working memory task (Romo et al., 1999), in which two vibrating stimuli separated by a delay of 3 s are presented to a monkey who then has to report whether the frequency of the first stimulus is larger or smaller than that of the second (Fig. 1A and B). The delayed tactile discrimination task requires three computational elements: encoding of a stimulus parameter (the first frequency), maintenance of its value in working memory, and comparison with the second stimulus. Single neurons that correlated well with these features were recorded in the PFC (Romo et al., 1999). Fig. 1C shows a neuron with a firing rate during the first stimulus that increases as a monotonic function of the stimulus frequency, a tendency that is then maintained throughout the delay period. The neuron depicted in Fig. 1D exhibits a negative monotonic dependence on stimulus frequency, suggesting that a subtractive comparison might be implemented by combining responses of these two types of neurons.

These striking properties prompted the formulation of network models that elegantly implement the three required computational elements (Miller et al., 2003; Machens et al., 2005; Miller and Wang, 2006). Many neurons in the PFC have less regular responses than those described above (e.g. Fig. 1E and F) and, across the population, response profiles are extremely heterogeneous (Brody et al., 2003; Singh and Eliasmith, 2006; Joshi, 2007). A recent analysis trying to ascertain the degree to which two models of this type fit the recorded data concluded that "Neither model predicted . . . a large fraction of the recorded neurons . . . suggesting that the neural representation of the task is significantly more heterogeneous than either model postulates" (Jun et al., 2010). While it seems natural to suppose that a neural circuit holding a fixed value of a stimulus parameter in short-term memory would do so by representing it in a time-invariant manner, the data do not support this view. The "large fraction of recorded neurons" that failed to match these models did so because they had highly time-dependent activity. Indeed, the dominant quantity being encoded by the recorded PFC neurons is not the stimulus parameter required for the task, but instead time (Machens et al., 2010).

To study the role of time-dependent neural activity in the storage of static stimulus parameters, we compare three models to the recorded data in the delayed tactile discrimination task. One of these is the line attractor, or LA, model of Machens and Brody (Machens et al., 2005). The second is a randomly connected network model, called RN, exhibiting chaotic activity with weight modification restricted solely to readout weights. The third is a recurrent network trained to perform the task by unrestricted modification of its connection weights, called TRAIN. Individual units in these models span the range seen in the data, from structured (Fig. 1C and D) to more complex (Fig. 1E) and highly irregular (Fig. 1F).

In addition to differing in the time-dependence of their stimulus representations, the LA, TRAIN and RN models also vary over a range of what might be called model orderliness, or model structure. The LA model was designed to perform the task by assuring that it contained a line of fixed points or line attractor that could statically represent different stimulus values. The TRAIN model was developed from an initially random network by applying the recently developed "Hessian-Free" learning algorithm (Martens and Sutskever, 2011). This constructs a network that is less structured than the LA model, although it performs the task in a somewhat similar manner. The RN model is also based on a randomly connected network, but in this case the only modified element is the readout of network activity; the internal connectivity, which defines the network dynamics, remains random and unrelated to the task. This is a novel application of an echo-state type of network (Jaeger, 2001; Maass et al., 2002) that operates in the chaotic rather than in the transient decaying regime typically used for such networks (Sussillo and Abbott, 2009). The result is a highly unstructured network with chaotic activity. Thus, the RN model is far removed from the LA model, both because its structure is essentially random rather than designed, and because it exhibits chaotic rather than fixed-point dynamics. The TRAIN model is intermediate between these extremes.

The ultimate goal is, of course, to figure out where PFC circuits performing the delayed tactile discrimination task lie on the spectrum from structured to unstructured and dynamically static