On memories, neural ensembles and mental flexibility

Dimitris A. Pinotsis[a,b,*], Scott L. Brincat[a], Earl K. Miller[a]

[a] The Picower Institute for Learning & Memory and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] The Wellcome Trust Centre for Neuroimaging, University College London, WC1N 3BG, UK

ARTICLE INFO

ABSTRACT

Memories are assumed to be represented by groups of co-activated neurons, called neural ensembles. Describing ensembles is a challenge: complexity of the underlying micro-circuitry is immense. Current approaches use a piecemeal fashion, focusing on single neurons and employing local measures like pairwise correlations. We introduce an alternative approach that identifies ensembles and describes the effective connectivity between them in a holistic fashion. It also links the oscillatory frequencies observed in ensembles with the spatial scales at which activity is expressed. Using unsupervised learning, biophysical modeling and graph theory, we analyze multi-electrode LFPs from frontal cortex during a spatial delayed response task. We find distinct ensembles for different cues and more parsimonious connectivity for cues on the horizontal axis, which may explain the oblique effect in psychophysics. Our approach paves the way for biophysical models with learned parameters that can guide future Brain Computer Interface development.

Introduction

Memories are assumed to be represented by groups of co-activated neurons, called neural ensembles. How to identify and describe neural ensembles has long been a central issue in neuroscience (Hebb, 1949). It is not an easy task: one has to deal with an immensely complex system where billions of neurons are linked to each other through trillions of connections. A further complication is that neurons can have multiple functions, especially in higher level cortex (Fusi et al., 2016; Rigotti et al., 2013). Thus, the same neurons may participate in many different ensembles and, conversely, different ensembles might share some of the same neurons. Clearly, the structure of these ensembles cannot be described in terms of anatomical connectivity only: if anatomical connectivity was all there was to ensembles, then activating one would activate others leading to a jumble of ensembles. Further, anatomy alone seems to preclude a hallmark of higher cognition: flexibility. Ensembles should be able to break apart and re-form from moment to moment without changing the underlying anatomy. Finally, ensembles are functional units and thus defining them based on anatomy alone is not possible in a behavioral context.

Previous work has attempted to identify neural ensembles using electrophysiological measures (Brown et al., 1998; Diba and Buzsáki, 2007; Johnson and Redish, 2007) and more recently optogenetics and immediate early gene (IEG) labelling (Ryan et al., 2015). However, due

to slow dynamics, these approaches can only provide limited insights into fast activity and neural oscillations that are thought to play a key role in memory function and ensemble formation (Buschman et al., 2012; Fries et al., 2007; Fusi et al., 2016; Haegens et al., 2011; Miller and Buschman, 2013). Thus far, electrophysiology studies have only considered neural ensembles in a piecemeal fashion, that is, using pairwise correlations. They have focused on single neurons and/or functional connectivity between pairs of neurons and/or recording sites; the existence of an ensemble is thus inferred indirectly (Buschman et al., 2012; Gray, 1999; Modi et al., 2014).

Here, we suggest an alternative approach to identifying ensembles based on effective connectivity. We describe ensemble properties using neurophysiological data combined with ideas from biophysical modeling, unsupervised learning and complex systems theory. We analyzed multiple-electrode recordings obtained during a classic test of working memory: spatial delayed response (Funahashi et al., 1990; Fuster et al., 1985). We examined LFPs between simultaneously recorded electrodes in dorsolateral prefrontal cortex (PFC), supplementary eye field (SEF), and frontal eye fields (FEF). Our goal was to identify neural ensembles carrying spatial information in a *holistic*, not piecemeal, fashion and describe the connections that form them.

We used brain decoding algorithms, graph theory and spectral analysis to understand the structure of neural ensembles that give rise to observed patterns of LFP responses. This allowed us to treat neural

ensembles as complex networks and to describe properties of the underlying connectivity. We obtained the estimates of effective connectivity underlying the ensembles by training a biophysical neural field model as a particular type of deep neural network called an autoencoder. We found that we could describe ensemble properties and use them to decode the spatial location held in working memory using only a few parameters, which makes this approach computationally tractable. Further, it also revealed ensemble properties that cannot be observed using pairwise correlations. For example, using topological measures, we found that network connectivity in the spatial delayed response task was different for different cued locations. Cues on the horizontal axis had shorter characteristic path lengths (the least number of steps between different network nodes) than others. This could explain the oblique effect (psychophysics performance is better for stimuli on than off the horizontal/vertical axes). We also found connectivity and corresponding oscillatory dynamics across different spatial scales and different frequencies within cortical areas, which gives a new dimension to cortical network interactions.

Materials and methods

Experimental data and recording setup

Two adult male monkeys (monkey C, Macaca fascicularis, 9 kg; monkey J, Macaca mulatta, 11 kg) were handled in accordance with National Institutes of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care. They were trained to perform an oculomotor spatial delayed response task. This task required the monkeys to hold the location of one of six randomly chosen visual targets (at angles of 0°, 60°, 120°, 180°, 240° and 300°, 12.5° eccentricity) in memory over a brief (750 ms) delay period and then saccade to the remembered location. If a saccade was made to the cued location, the target was presented with a green highlight and a water reward was delivered otherwise the target was presented with a red highlight and reward was withheld. Three 32-electrode chronic arrays were implanted unilaterally in PFC, SEF and FEF in each monkey (Fig. 1A). Each array consisted of a 2 × 2 mm square grid, where the spacing between electrodes was 400 μm. The implant channels were determined prior to surgery using structural magnetic

resonance imaging and anatomical atlases. From each electrode, we acquired both threshold-crossing spike waveforms and local field potentials (extracted with a fourth order Butterworth low-pass filter with a cut-off frequency of 500 Hz, and recorded at 1 kHz) using a multichannel data acquisition system (Cerebus, Blackrock Microsystems). We analyzed local field potentials (LFPs) during the delay period when monkeys held the cued locations in memory. We assumed that each electrode sampled LFP activity from a neural population in its proximity and modelled each brain area as a cortical area sampled at $N_S = 32$ locations. LFP activity was modelled by a mathematical model of wave dynamics known as a neural field. Electrodes were numbered in a monotonic fashion; neighbouring electrodes had adjacent numbers.

Neural ensembles for memory maintenance

Using a neural field model allowed us to use patterns of LFP activity across recording sites to infer the underlying effective connectivity for each of the cued locations. Neural fields provided a quantitative way to describe each ensemble's network interactions and make predictions about patterns of activity that correspond to different attractor states, see Fig. 1B. Each attractor state can be considered to reflect an ensemble or engram (Liu et al., 2012). This is also related to chimera states and metastability (Martens et al., 2016). Our goal was to obtain learned connectivity parameters that can describe the structure of neural ensembles activated while remembering different stimuli. The spacing between electrodes was larger (400 μm) than what is thought to be the origin of the LFP signal (250 μm, see Katzner et al., 2009). Using brain decoding algorithms and graph theoretic measures (centrality) we quantified the separability of neural ensembles in SEF and FEF (see below). However, volume conduction could in principle introduce confounds. These can be accommodated by using the effective connectivity parameters obtained here as priors to fit a more complicated biophysical model that accounts for volume conduction effects (Pinotsis et al., 2014). As it is common in computational neuroscience and modern machine learning approaches, we describe neural activity using a one dimensional model. In this model, space is defined along the line traced out by the electrodes. This is similar to ring models (Ben-Yishai et al., 1997; Somers et al., 1995), recurrent

A. Electrode arrays

B. Neural Field Model and Connections
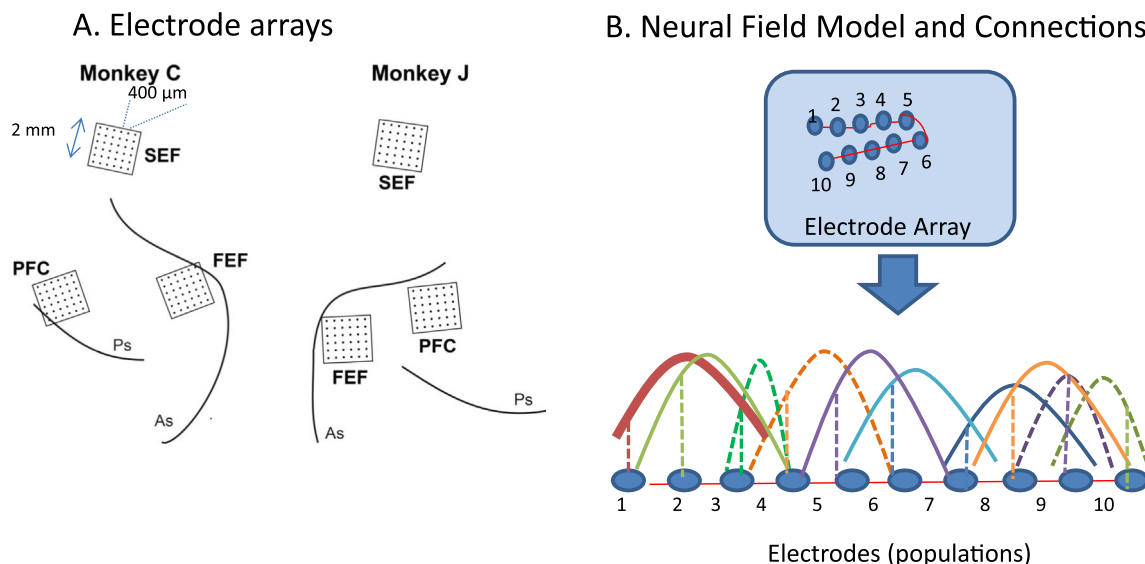
Fig. 1. A. Recording setup. Three 32-electrode chronic microelectrode arrays were implanted in dorsolateral prefrontal cortex (PFC), supplementary eye field (SEF), and frontal eye field (FEF) in each monkey. Each array consisted of a 2 × 2 mm square grid, where the spacing between electrodes was 400 μm. Ps: principal sulcus; As: arcuate sulcus. One monkey received implants in the left hemisphere and the other in the right hemisphere. These were located near Ps and As in both monkeys. **B.** Neural field model and connections. Neural fields provided a quantitative way to describe each ensemble's network interactions and patterns of activity across simultaneously recorded sites. The same model can describe different ensembles. Each electrode occupies a position on a cortical manifold (line) W parameterized by the variable v and is connected to all other electrodes with connections whose strength follows a Gaussian profile (coloured solid and dashed lines), see also Eq. (4).