Distinguishing Causal Interactions in Neural Populations

Anil K. Seth
A.K.Seth@sussex.ac.uk

Gerald M. Edelman
edelman@nsi.edu
The Neurosciences Institute, San Diego, CA 92121, U.S.A.

We describe a theoretical network analysis that can distinguish statistically causal interactions in population neural activity leading to a specific output. We introduce the concept of a causal core to refer to the set of neuronal interactions that are causally significant for the output, as assessed by Granger causality. Because our approach requires extensive knowledge of neuronal connectivity and dynamics, an illustrative example is provided by analysis of Darwin X, a brain-based device that allows precise recording of the activity of neuronal units during behavior. In Darwin X, a simulated neuronal model of the hippocampus and surrounding cortical areas supports learning of a spatial navigation task in a real environment. Analysis of Darwin X reveals that large repertoires of neuronal interactions contain comparatively small causal cores and that these causal cores become smaller during learning, a finding that may reflect the selection of specific causal pathways from diverse neuronal repertoires.

1 Introduction

How do neuronal interactions in complex nervous systems cause specific outputs, and how are these causal interactions modulated during learning? Addressing these issues requires a theoretical analysis of the principles governing causal interactions in large and richly interconnected neuronal populations during behavior. Here, we describe a novel network analysis that distinguishes causal interactions in the neural activity in relation to a given neural reference. We use the term neural reference (NR) to refer to the activity of a particular neuron (e.g., a motor neuron) at a particular time (e.g., when an output is generated). As part of this analysis, we introduce the concept of a causal core to designate those neural interactions that are causally significant for a given NR. This concept provides a novel means of linking neuronal interactions in large populations to specific outputs.

Our analysis derives from the perspective that the brain is a selectional system and that neuronal interactions reflect processes of selection operating on diverse repertoires of neuronal groups (Edelman, 1978, 1987, 1993;...
Distinguishing Causal Interactions in Neural Populations

Izhikevich, Gally, & Edelman, 2004; Seth & Baars, 2005). Our approach here elaborates on three features of this view in particular. First, variability in population activity may be a necessary substrate for the formation of diverse and dynamic repertoires of neuronal interactions and may reflect not only neuronal noise. Second, population activity patterns are interpreted as dynamic, causally effective processes giving rise to specific outputs, and not as information-bearing representations underlying computations. Finally, synaptic plasticity is suggested to contribute to behavioral learning by shaping the selection of causal pathways within populations and not by reinforcing connections among neural representations of sensory and motor variables. We emphasize that selectionist theory provides an interpretive framework for this analysis and is not itself tested by our analysis.

The general framework for the analysis is illustrated in Figure 1. First, an NR is selected from among many possible neuronal events—in this example, by virtue of its relationship to a specific behavioral output (see Figure 1A; yellow neuron at time t). Second, a network of neuronal interactions is identified, incorporating only those interactions that could have potentially contributed to the occurrence of the NR (Krichmar, Nitz, Gally, & Edelman, 2005). This network corresponds to the “context network” of the NR (see Figure 1A; network of green neurons). In order to sample the most salient components for a particular NR, these networks generally reflect a relatively short time period. Finally, a subnetwork, which we define as the causal core, is selected from the context network. The causal core comprises those interactions that are causally significant as assessed by Granger causality (see Figure 1B and below; red arrows show causally significant connections), and which form part of a causal chain leading to the NR (see Figure 1C; red arrows indicate the causal core, and the blue arrow indicates a causally significant interaction that is excluded from the core because it is a causal dead-end). The concept of Granger causality is based on prediction: if a signal A causes a signal B, then past values of A should help predict B with greater accuracy than past values of B alone (Granger, 1969). Because the statistical nature of Granger causality requires large sample sizes to make robust inferences, it is assessed over a time period considerably longer than that used for identifying the context network.

In practice, the above analysis requires extensive knowledge of anatomical and functional connectivity during behavior. Unfortunately, such data are not currently available in vivo. Therefore, to provide an illustrative example, we analyze a brain-based device (BBD), a physical device whose behavior in a real-world environment is controlled by a simulated nervous system based on vertebrate neuroanatomy and neurophysiology (Krichmar & Edelman, 2002; Seth, McKinstry, Edelman, & Krichmar, 2004b, 2004a; Krichmar, Seth, Nitz, Fleischer, & Edelman, 2005). Unlike animal models, use of a BBD enables precise recording during behavior of the state of every neuronal unit and every synaptic connection. The BBD analyzed here, Darwin X, incorporates a detailed model of the hippocampus and
Figure 1: Distinguishing causal interactions in neuronal populations. (A) Select a neural reference (NR), that is, the activity of a particular neuron (yellow) at a particular time (t). The context network of the NR corresponds to the network of all coactive and connected precursors, assessed over a short time period (e.g., t, t − 1, . . . , t − 6). (B) Assess the Granger causality significance of each interaction in the context network, based on extended time series of the activities of the corresponding neurons (0 − T; this time period is considerably longer than that used to identify the context network itself). Red arrows indicate causally significant interactions. (C) The causal core of the NR (red arrows) is defined as that subset of the context network that is causally significant for the activity of the corresponding neuron (excluding both noncausal interactions—black arrows—and dead-end causal interactions such as 5 → 4, indicated in blue). The concept of a causal core provides a novel means of linking neuronal interactions in large populations to specific outputs.

surrounding cortical areas. By integrating visual and self-motion cues, it learns to navigate to an arbitrarily chosen location in a room (Krichmar, Nitz et al., 2005; Krichmar, Seth et al., 2005).

To analyze Darwin X, we identified NRs at various stages of the learning process, selecting those corresponding to hippocampal control of motor
output. By recording in detail the responses of Darwin X’s nervous system during behavior, we were able to construct context networks and identify causal cores for the chosen NRs. We found that causal cores generally comprised small subsets of the corresponding context networks. Moreover, causal cores were not composed solely of strong synaptic interactions and were not random samples of a context network. We also found that causal cores became smaller with experience, a process we refer to here as refinement.

Although BBDs are necessarily gross simplifications of biological systems, their analysis can provide potentially useful heuristics for the interpretation of neurobiological data (Krichmar & Edelman, 2002). Accordingly, as with previous analyses of Darwin X (Krichmar, Nitz, et al., 2005; Krichmar, Seth, et al., 2005), the results presented here can be related to mammalian hippocampal function. However, our primary purpose here is to elaborate several heuristics for analyzing and interpreting activity patterns within neuronal populations. The timeliness of developing such heuristics is emphasized by recent methodological advances that promise detailed new data revealing anatomical and functional architectures of neuronal populations at microscopic scales (Kelly & Strick, 2004; Konkle & Bielajew, 2004; Raffi & Siegel, 2005; Cossart, Ikegaya, & Yuste, 2005; Ohki, Chung, Kara, & Reid, 2005).

Figure 2: Darwin X in its environment with visual landmarks hung from the walls and with the hidden platform indicated. Darwin X has a CCD camera, which responds to landmarks that appear within its field of view. While Darwin X could not see the hidden platform, it could detect an encounter with the platform by means of a downward-facing IR sensor. This figure has been adapted from (Krichmar, Seth, et al., 2005).
We emphasize that this study differs fundamentally from previous analyses of Darwin X (Krichmar, Nitz, et al., 2005; Krichmar, Seth, et al., 2005). One previous study introduced the recursive analysis of anatomical and functional connectivity that is used in this study to identify context networks (Krichmar, Nitz, et al., 2005). However, this previous study did not analyze causal interactions or examine changes in neural dynamics during learning. A second study applied a different Granger causality analysis to small circuits of 8 to 10 neuronal units each (Krichmar, Seth, et al., 2005). However, these circuits were selected without reference to anatomy, and the small size of the selected circuits precluded analysis of population-level neuronal dynamics.
2 Materials and Methods

2.1 Darwin X. Darwin X is the most recent of a series of autonomous brain-based devices (BBDs) that have been developed over the past 12 years (Edelman et al., 1992; Krichmar & Edelman, 2002; Seth et al., 2004a; 2004b; Krichmar, Nitz, et al., 2005; Krichmar, Seth, et al., 2005). Details of the construction and behavior of Darwin X are available elsewhere (Krichmar et al., 2005a, 2005b) and only those necessary for comprehension will be repeated here. Figure 2 shows the device in its environment, which consisted of an enclosed arena within which paper stripes of different widths and colors were hung on different walls. Darwin X was equipped with a CCD camera, enabling it to detect these stimuli. Figure 3 provides a schematic diagram of the simulated nervous system, in which different sensory streams converged onto a hippocampal region in which there were several levels of signal processing before the processed signals projected divergently back to sensory cortex (Lavenex & Amaral, 2000). The sensory streams included visual “where” signals that respond to stripe width (parietal area Pr), visual “what” signals that respond to stripe color (inferotemporal area IT), and head-direction signals (anterior thalamic area ATN). These areas were reciprocally connected to entorhinal area EC (subdivided into ECin and ECout). Both trisynaptic pathways (EC → DG → CA3 → CA1) and perforant shortcuts (e.g., EC → CA1) were included in the simulated hippocampal anatomy (Witter, Naber, & van Haeften, 2000). The full nervous system of Darwin X contained 50 neural areas, approximately 90,000 mean firing-rate neuronal units, and approximately 1.4 million synaptic connections.

Darwin X was trained on a dry version of the Morris water maze task (Morris, 1984). A hidden platform was placed in one quadrant of the arena (see Figure 2). This platform could be detected by Darwin X only when the device was directly overhead. Each encounter with the platform stimulated a value system that modulated synaptic plasticity at various loci in the simulated nervous system (see Figure 3). Synaptic plasticity in Darwin X was based on a modified BCM learning rule (Bienenstock, Cooper, & Munro, 1982; Krichmar, Seth, et al., 2005) in which synapses between neuronal units with strongly correlated firing rates are potentiated and synapses between neuronal units with weakly correlated firing rates are depressed.

As a result of the above features, Darwin X learned to navigate to the platform from arbitrary starting locations. Darwin X was trained over 16 separate trials, each beginning from one of four different starting locations. As reported in detail in Krichmar, Seth, et al. (2005), initially Darwin X searched randomly for the platform, but after about 10 trials, the device reliably took a comparatively direct route to the platform from any starting location.

2.2 Context Networks. To identify context networks in Darwin X’s neural activity, we selected 15 neuronal units in area CA1 of the simulated
hippocampus, whose activity directly affected motor output. We refer to these units as reference units. For each reference unit, we then identified various time steps, drawn from a variety of experimental trials, at which the reference unit was active and at which Darwin X changed its direction. This combination of reference units and time steps provided 96 neural references (NRs; see Figure 1A) that spanned the learning process. It is important to emphasize here these distinctions:

- **Neuronal unit**, which make up all parts of Darwin X’s simulated nervous system.
- **Reference unit**, a neuronal unit in CA1 chosen for its influence on behavior.
- **Neural reference**, the activity of a particular reference unit at a particular time. In this analysis, each reference unit was associated with multiple NRs. While it may be natural to select NRs that correspond to behavioral outputs, NRs can in principle correspond to the activity of any neuronal unit at any time.

We then recursively examined the activity of all neuronal units (throughout the simulated nervous system) that led to each NR (see Figure 1A). This procedure has previously been referred to as a backtrace (Krichmar, Nitz, et al., 2005). For each NR, we identified the list of neuronal units that were both anatomically connected to the corresponding reference unit and were active, above a threshold, at the previous time step. The threshold sets the minimum value of the product of presynaptic activity and synaptic strength for a synaptic interaction to be included; it was set to be very low (0.01). A second iteration was carried out by repeating this procedure for each neuronal unit in this list. In total, six successive iterations were carried out for each NR, reflecting six simulation cycles of neural activity, which corresponded to approximately 1.2 s of real time. This limit was chosen in order to isolate the most salient neural interactions for a particular NR while avoiding a combinatorial explosion in the size of successive iterations. The resulting context networks consisted of multiple paths of functional interactions leading, through time, to the corresponding NRs (see Figure 1A). We emphasize that the term *context network* is used here to refer specifically to the networks described and does not connote alternative uses of the term *context* pertaining to, for example, invariances in environmental, emotional, or cognitive conditions (Mackintosh, 1983; Bouton, 2004; Fanselow, 2000; Baars, 1988).

In principle, this procedure could be extended to NRs consisting of multiple coactivated neuronal units, in which case the first iteration of the backtrace would be carried out using a list of neuronal units rather than a single neuronal unit. However, the corresponding context networks would likely increase in size very rapidly, leading to combinatorial problems.
2.3 Causal Cores. To identify the causal core of each NR, we applied a Granger causality analysis to its context network (Granger, 1969; Seth, 2005). The concept of Granger causality is based on prediction, and its application usually involves linear regression modeling. To illustrate Granger causality, suppose that the temporal dynamics of two time series, \( X_1(t) \) and \( X_2(t) \) (both of length \( T \)), can be described by a bivariate autoregressive model:

\[
X_1(t) = \sum_{j=1}^{p} A_{11,j} X_1(t - j) + \sum_{j=1}^{p} A_{12,j} X_2(t - j) + E_1(t)
\]

\[
X_2(t) = \sum_{j=1}^{p} A_{21,j} X_1(t - j) + \sum_{j=1}^{p} A_{22,j} X_2(t - j) + E_2(t),
\]

where \( p \) is the maximum number of lagged observations included in the model (the model order, \( p < T \)), \( A \) contains the coefficients of the model (i.e., the contributions of each lagged observation to the predicted values of \( X_1(t) \) and \( X_2(t) \)), and \( E_1, E_2 \) are residuals (prediction errors) for each time series. If the variance of \( E_1 \) (or \( E_2 \)) is reduced by the inclusion of the \( X_2 \) (or \( X_1 \)) terms in the first (or second) equation, then it is said that \( X_2 \) (or \( X_1 \)) Granger-causes \( X_1 \) (or \( X_2 \)). In other words, \( X_2 \) Granger-causes \( X_1 \) if the coefficients in \( A_{12} \) are jointly significantly different from zero. This can be tested by performing an F-test of the null hypothesis that \( A_{12} = 0 \), given assumptions of covariance stationarity on \( X_1 \) and \( X_2 \). The magnitude of a Granger causality interaction can be estimated by the logarithm of the corresponding F-statistic (Geweke, 1982).

We used Granger causality to assess the causal significance of every connection in each context network (see Figure 1B). For each connection, \( X_1 \) and \( X_2 \) corresponded to the activity time series of the presynaptic and postsynaptic neuronal units, respectively. To ensure a robust estimation of causal significance, \( X_1 \) and \( X_2 \) contained neuronal unit activities for the entire trial in which the corresponding NR was located. The length of these time series ranged from 450 to 4992 time steps (in contrast to the 6 time steps incorporated into each backtrace; see above), and each contained many separate periods of neuronal activity and inactivity. To achieve covariance stationarity, first-order differencing was applied to \( X_1 \) and \( X_2 \). The analyzed time series therefore represented changes in neural activity rather than absolute activity levels. After differencing, only 3 of 96 context networks contained more than 1% of time series that were not covariance stationary (Dickey-Fuller test, \( P < 0.01 \)). These networks were subsequently excluded from the analysis.

For each connection in each of the 93 remaining context networks, we constructed a bivariate autoregressive model of order \( p = 4 \). This model order was chosen according to the Bayesian information criterion (BIC) (Schwartz, 1978) (the mean minimum BIC, computed from all bivariate
models, was 3.46). To verify that the order-4 models sufficiently described the data, we noted that the mean adjusted residual-sum-square \( \text{RSS}_{\text{adj}} \) was sufficiently high (0.35 \( \pm \) 0.168). From each bivariate model, we calculated the statistical significance of a Granger causality interaction from the presynaptic unit \( (X_1) \) to the postsynaptic unit \( (X_2) \). To take account of the large number of connections in each context network, a Bonferroni correction was applied to the significance thresholds (corrected \( P < 0.01 \)). The resulting networks of significant Granger causality interactions are referred to here as Granger networks.

The causal core of each NR was identified by extracting the subset of the corresponding Granger network consisting of all causally significant connections that led, via other causally significant connections, to the NR (see Figure 1C). We emphasize the importance of this step. In order to be causally relevant to an NR, a causally significant interaction must be part of a chain of causally significant interactions leading to the NR.

To assess the robustness of our results, we repeated the Granger causality analysis of each context network using a range of Bonferroni-corrected significance thresholds (\( P < 0.02 \) to \( P < 0.05 \)). The resulting Granger networks and causal cores were very similar in all cases to those obtained using the default threshold (\( P < 0.01 \)). Robustness to model order \( p \) was assessed by repeating the analysis using order-8 models, as indicated by the Akaike information criterion, an alternative to the BIC (Akaike, 1974). The resulting Granger causalities were identical to those identified by the order-4 models.

### 2.4 Network Measures

Throughout this letter, we use the term network to provide a useful and commensurable representation, in terms of nodes and edges, of both real physical interactions (as in context networks) and virtual statistical interactions (as in Granger networks and causal cores). In both cases, nodes correspond to neuronal units. In context networks, edges correspond to synaptic interactions, weighted by the product of synaptic strength and presynaptic activity. In Granger networks and causal cores, edges correspond to statistically significant causal interactions, weighted by the estimated strength of the causal interaction.

In order to measure the similarity between two networks \( A \) and \( B \), we define the following network similarity index,

\[
\eta_{AB} = \frac{2f(A \cap B)}{f(A) + f(B)},
\]

where \( A \cap B \) is the network consisting of the edges jointly present in networks \( A \) and \( B \), and the function \( f(X) \) returns the size (number of edges) of the network \( X \). This index is equal to 1 if \( A \) and \( B \) are identical and is equal to 0 if \( A \) and \( B \) are nonoverlapping. It is important to emphasize that
while topologies can be compared among different network types using the $\eta$ index, the specific weight assigned to an edge in a context network cannot be directly compared to the corresponding value in a Granger network or causal core.

Networks of each type were visualized using either an energy-minimization method to emphasize network structure or, in a standardized circular arrangement, to aid comparison among networks. In the former case, we used the Pajek program (http://vlado.fmf.uni-lj.si/pub/networks/pajek/), which implements the Kamada-Kawai energy minimization algorithm, with the result that nodes that are strongly interconnected tend to be bunched together. Because this algorithm was applied separately for each visualized network, spatial arrangements cannot be compared between networks.

3 Results

3.1 Causal Cores in Darwin X. Figure 4 shows the context network for an example NR (reference unit 6 during trial 13). The thickness of each line and size of each arrowhead is determined by the product of synaptic strength and presynaptic activity. Figure 5 shows the corresponding Granger network—the subset of interactions of the context network in Figure 4 calculated to be causally significant. Here, the thickness of each line reflects the strength of the corresponding Granger causality interaction. Figure 6 shows the corresponding causal core—the subnetwork of Figure 5 following removal of all dead-end edges that did not participate in causal chains leading to the NR. As in the Granger network (see Figure 5), the thickness of each line in the causal core reflects the strength of the corresponding Granger causality interaction.

Comparing the networks in Figures 4, 5, and 6, it is apparent that the context network is much larger than the Granger network, which is itself much larger than the causal core. For this example NR, the causal core has several trisynaptic pathways in which signals from sensory cortex causally influence entorhinal cortex, which in turn causally influences dentate gyrus and CA3 before reaching the NR. One example of a such a pathway is shown by the gray shading in Figure 6.

In the context network (see Figure 4), there is a large central cluster of entorhinal interactions interspersed with entorhinal-CA1 interactions. The Granger network (see Figure 5) has a comparatively sparse architecture, with considerably fewer intra-entorhinal interactions. Intra-entorhinal interactions are almost totally absent in the corresponding causal core (see Figure 6).

Figure 7 shows the average composition and size of all 93 context networks, Granger networks, and causal cores. Consistent with the above observations, context networks were significantly larger than Granger networks, which were themselves significantly larger than causal cores.
Intra-entorhinal interactions dominated context networks, were less prominent in Granger networks, and were comparatively rare in causal cores. The expanded segments of each chart show the components of the trisynaptic and perforant pathways. These components were much more prominent in causal cores than in context networks or Granger networks, suggesting that these pathways had greater causal influence, in relation to the NRs, than entorhinal interactions.

How does the size of a causal core depend on the amount of time that is integrated by a context network? As noted previously (in section 2.2), each context network was constructed by iterating six time steps back from a given NR (a backtrace). Figure 8 shows the ratio of causal core size to the corresponding context network size, for each NR separately, as the backtrace procedure is iterated. As additional time steps were incorporated in the analysis (i.e., as the context network reached further back in time from the NR), the majority of causal cores constituted a progressively smaller fraction of the corresponding context networks. This raises the possibility that in Darwin X, even if the construction of a context network were to be iterated through many more time-steps than is feasible in practice, the corresponding causal core may not change substantially.

Can causal cores in Darwin X be identified on the basis of synaptic strengths, without recourse to a statistical causality analysis? To address this question, we examined the relation between synaptic strength and Granger causality for each NR separately, using the network similarity index \( \eta \) defined in section 2.4. From each context network, we selected \( j \) subnetworks by removing those edges corresponding to \( j\% \) of the weakest synaptic weights, for all \( j \in [1, 100] \). From each subnetwork, we also removed those dead-end edges that did not lead to the NR. We then calculated the network similarity \( \eta \) between each subnetwork and the corresponding causal core. The resulting values represent, for a particular NR, the similarity between the causal core and a series of subnetworks identified by narrowing the corresponding context network on the basis of synaptic strength.

Figure 9A shows the mean network similarity \( \eta \) across all NRs as a function of \( j \). For low \( j \), \( \eta \) is low, as expected, because the context networks are still large. \( \eta \) is also low for high \( j \), which indicates that causal cores do not consist solely of the strongest synaptic weights. Figure 9B shows the same analysis repeated using synaptic interaction strength (i.e., the product of synaptic weight and presynaptic activity) instead of synaptic weight. The results are very similar. The maximum \( \eta \) averaged across all NRs was approximately 0.3 for both analysis methods, and we found the maximum value of \( \eta = 1 \) for only 1/93 NRs, in the second analysis only. These observations indicate that causal cores cannot in general be identified on the basis of synaptic weights or synaptic interaction strengths.

In relation to the above point, it bears emphasizing that whereas a Granger-causality analysis offers a natural threshold, based on
Figure 4: Example context network from Darwin X, obtained from CA1 reference unit 6 during trial 13. Each circle represents a single neuronal unit; different colors represent distinct brain areas, as defined in section 2.1 and Figure 3. The thickness of each line reflects the product of synaptic strength and presynaptic activity.

Figure 5: The Granger network corresponding to the context network shown in Figure 4. This network comprises those connections in Figure 4 that are causally significant. The thickness of each line reflects the strength of the Granger causality interaction (the logarithm of the corresponding F-statistic; see section 2.3).
Figure 6: The causal core corresponding to the context network in Figure 4 and the Granger network in Figure 5. This network was derived by removing from the corresponding Granger network all dead-end edges that did not participate in causal chains leading to the NR. The gray shading shows one of the several trisynaptic pathways in the causal core. As in Figure 5, the thickness of each line reflects the strength of the Granger causality interaction.

Figure 7: Network composition and average network size (number of edges) for context networks, Granger networks, and causal cores. Different colors represent different connection types (see section 2.1 and Figure 3). Expanded segments show components of trisynaptic pathways (EC→DG, DG→CA3, and CA3→CA1) and perforant pathways (EC→CA1). Context networks had a much higher proportion of intra-entorhinal connections than causal cores.
Figure 8: The ratio of causal core size to context network size, in terms of connections ($K$) and neuronal units ($N$), for all NRs. This ratio is shown as a function of the iteration depth of the backtrace procedure (see section 2.2).

Figure 9: Network similarity $\eta$ between causal cores and a series of subnetworks extracted from the corresponding context networks. For each NR, 100 subnetworks were extracted by removing $j\%$ of the weakest edges in the corresponding context networks, as well as dead-end edges (see the text for details). (A) Mean $\eta$ across all 93 NRs for subnetwork extraction based on synaptic weight. (B) Mean $\eta$ based on synaptic interaction strength. Shaded areas represent standard errors.

statistical significance, for selecting a subnetwork, thresholds based on synaptic weight or synaptic interaction strength are necessarily arbitrary. Comparisons of size and composition would therefore be more difficult to interpret in these latter cases.

3.2 Modulation of Causal Interactions During Learning. We turn now to the modulation of causal interactions during learning of the hidden-platform task by Darwin X. Figure 10 shows causal cores from two reference
Figure 10: Causal cores for reference units 5 (top) and 10 (bottom) from a selection of trials spanning the learning period. To facilitate comparison between networks, they are arranged using a circular template in which units from the same brain area are placed next to each other. The causal cores diminished in size over time.

Inspection of Figure 10 shows that the causal cores diminished in size over time, mainly due to the pruning of intra-entorhinal interactions. We refer to this reduction in size during learning as refinement.

Causal core refinement could arise in three different ways: (1) as a correlate of reduction in size of the corresponding context networks; (2) as a result of a general reduction in the likelihood with which interactions are causally significant (in which case, Granger networks, but not context networks, would be similarly refined); and (3) as the result of selection of particular causal pathways from a diverse and dynamic repertoire of neural interactions (in which case only causal cores would show refinement). To distinguish these possibilities, Figure 11 shows network sizes as a function of experimental trial for each type of network and separately for each of the 15 reference units. In agreement with the third account, only causal cores showed significant refinement (see the inset in Figure 3C). Further, causal core size appeared to reach an asymptote after about 10 trials, which corresponds to the number of trials required for Darwin X to learn to move more
or less directly to the hidden platform from any starting point (Krichmar, Seth, et al., 2005).

These observations raise the possibility that causal core refinement may be a correlate of learning at the level of neuronal populations. Since causal cores in Darwin X were not composed solely of the strongest synaptic interactions (see Figure 9) and since neither context networks nor Granger networks showed similar reductions in size during learning (see Figure 11), causal core refinement in Darwin X may best be understood as resulting from the modulation, via synaptic plasticity, of causal pathways within neuronal populations.

4 Discussion

We have developed a theoretical network analysis that can distinguish statistically causal interactions (according to Granger causality) from non-causal interactions in the neural activity leading up to the activity of a particular neuronal unit at a particular time (a neural reference, NR). According to our analysis, activity within neuronal populations may involve the selection of causal interactions from diverse and dynamic repertoires, and learning may involve synaptic changes that modulate selection at the population level.

An illustrative application of our analysis to a brain-based device provided an opportunity to explore patterns of statistically causal interactions in large neuronal repertoires during behavior. Causal interactions in Darwin X arose from network dynamics generated by interactions among anatomy, the specific embodiment of Darwin X, and behavior in a structured environment, and therefore could not be inferred directly from design specifications. Our analysis of Darwin X revealed several novel aspects of such
interactions, notably the small size of causal cores as compared to context networks (see Figures 4–7) and the reduction in size (refinement) of causal cores during learning (see Figures 10 and 11). We verified that causal cores did not comprise only the strongest synaptic weights or synaptic interactions and could not in general be identified on the basis of these features (see Figure 9).

Because brain-based devices such as Darwin X reflect vertebrate neuroanatomy and neurophysiology, their analysis can provide potentially useful heuristics for the interpretation of neurobiological data (Krichmar & Edelman, 2002; Seth, Sporns, & Krichmar, 2005). Below, we discuss some implications of these results for the interpretation and analysis of population neural activity, for the relation between synaptic plasticity and behavioral learning, and for the dynamic mechanisms underlying hippocampal function. We also discuss limitations, constraints, and possible extensions of the approach.

4.1 Population Activity and Neuronal Variability. A common view of neural population activity is that activity patterns encode information about sensory, motor, or internal state variables (Pouget, Dayan, & Zemel, 2000; Georgopoulos, Schwartz, & Kettner, 1986; Lee, Rohrer, & Sparks, 1988), and that population activity patterns constitute representations that provide substrates for computational operations (Averbeck, Latham, & Pouget, 2006; Harris, 2005). According to this view, mechanisms are required for encoding afferent signals into activity patterns as well as for decoding these activity patterns in order to generate output (Pouget et al., 2000; Seung & Sompolinsky, 1993). An alternative perspective, which accords with the theory that the brain is a selectional system (Edelman, 1978, 1987, 1993), is that population activity provides a dynamic repertoire from which specific causal pathways are selected during behavior and learning. The analysis here provides a means of identifying these pathways in the form of causal cores. According to this analysis, causal cores are not information-bearing representations and do not require explicit encoding and decoding. Instead, they are dynamic, causally effective processes that give rise to specific outputs.

These views offer contrasting interpretations of neuronal variability. Experimental observations of the brain during behavior indicate large variations in neural activity even for performance of the same task, or perception of the same stimulus (Edelman, 1987; Softky & Koch, 1993; Schwartz, 1994; Grobstein, 1994). Often such variability has been treated as an inconvenience and minimized or removed using averaging in order to reveal the underlying activity patterns. From a population coding perspective, attempts have been made to identify whether variation is unavoidable neural noise or, alternatively, is part of a signal transmitted to other neurons (Stein, Gossen, & Jones, 2005; Knoblauch & Palm, 2005; Averbeck et al., 2006) via a neural code (deCharms & Zador, 2000).
contrast, according to the view presented here, variability in the neural activity leading to a given event is expected and indicates the existence of diverse and dynamic repertoires underlying selection.

4.2 Synaptic Plasticity and Behavioral Learning. It is often assumed that learned behavior depends on the cumulative effects of long-term potentiation (LTP) or depression (LTD) of synapses. However, synaptic plasticity and behavioral learning operate over vastly different temporal (Drew & Abbott, 2006) and spatial scales. Although experimental evidence suggests that behavioral learning can induce LTP and LTD in certain cases (Malenka & Bear, 2004; Dityatev & Bolshakov, 2005), the precise functional contributions of synaptic plasticity to learned behavior have remained unclear. The results here raise the possibility that synaptic plasticity may underlie learning via modulations of causal interactions at the population level, specifically, by giving rise to the refinement of causal cores. Accordingly, the comparatively large size of causal cores early in learning may reflect a situation where selection had not yet acted to consolidate a causal pathway within a large space of possible pathways.

Recent experimental evidence is consistent with this selectionist view. For example, Barnes and colleagues sampled populations of sensorimotor striatal projection neurons, while rats learned a conditional T-maze task (Barnes, Kubota, Hu, & Graybiel, 2005). They found that variability in task-related spiking activity (as measured by entropy) diminished during training and overtraining. Interestingly, by including extinction and reacquisition phases in the learning paradigm, they showed that this reduction in variability can be reversed.

4.3 Causal Interactions in the Hippocampus. The close correspondence of Darwin X with the neuroanatomy and neurophysiology of the mammalian hippocampus suggests one comparatively specific hypothesis regarding hippocampal function. Intra-entorhinal interactions, which dominated most context networks, were comparatively rare in Granger networks, and especially in causal cores. This raises the possibility that mammalian entorhinal cortex may play a modulatory role, as opposed to a causally driving role, with respect to neuronal responses in CA1 during learning. By the same token, the prevalence of trisynaptic and perforant pathways within causal cores suggests that these pathways may causally drive CA1 responses. Although these possibilities are difficult explicitly to test in vivo, the notion of causally efficacious trisynaptic and perforant pathways is supported by both anatomy (see Figure 3) and recent experimental and modeling results that suggest a functionally significant perforant pathway (Brun, Øtnass, & Molden, 2002; Hasselmo, Bodelon, & Wyble, 2002). In addition, evidence is accumulating that entorhinal cortex can modulate behavioral expression, for example by driving the formation of associations
for long-term memory (Frank & Brown, 2003) and by influencing the expression of conditioned fear in rats (Burwell, Bucci, Sanborn, & Jutras, 2004; Hebert & Dash, 2004). Future work will address in detail further implications of this analysis for theories of hippocampal function.

4.4 Role of Neuronal Inactivity. The operation of any neural system is likely to involve neurons for which momentary inactivity is functionally significant. Our analysis accounts for this aspect of neural dynamics in the identification of causal cores from context networks, but not in the identification of context networks themselves. The reason is that momentary inactivity can be functionally significant only as part of an extended time series. As described in section 2.2, context networks comprise those synaptic interactions that could potentially have influenced a neural event at a particular time. Accordingly, context networks incorporate neural interactions corresponding to active synapses only. By contrast, the assessment of causal significance by Granger causality is based on relative predictive ability of time series, which requires each time series to exhibit periods of inactivity as well as periods of activity. Therefore, the identification of a causal core from a context network naturally accommodates momentary neuronal inactivity because such inactivity contributes to predictive ability.

4.5 Validating Causal Hypotheses Using Lesions. A useful means of testing causal relations in general is to lesion or perturb elements of the studied system (Keinan, Sandbank, Hilgetag, Meilijson, & Ruppin, 2004). We have shown in previous work that a Granger causality analysis can successfully predict the behavioral consequences of single-neuron lesions within a highly simplified neural network model and is superior in this regard to predictions based on synaptic weights or synaptic interaction strengths (Seth, 2005). In the analysis here, lesion studies of particular causal cores would be difficult to interpret for the reason that the ongoing behavior of Darwin X depends only minimally on the activity of any single neuronal unit. Moreover, Darwin X is a degenerate system (Edelman & Gally, 2001) in which many different causal circuits are likely to give rise to the same behavior. Having said this, it may be informative for future studies to compare lesions to entire brain regions, including, for example, entorhinal areas and trisynaptic/perforant areas.

It bears emphasizing that the present analysis provides causal descriptions in the absence of lesions or exogenous perturbations. The resulting causal cores therefore correspond to the dynamics of the intact system during natural behavior. By contrast, the interpretation of causal inferences based on external interventions is complicated by the fact that the studied system is either no longer intact (for lesions) or may display different behavior (for exogenous perturbations).
4.6 Covariance Stationarity. A requirement for the present analysis is that time series are covariance stationary; i.e., means and variances do not change over time. However, biological time series often exhibit strong autocorrelations that are inconsistent with covariance stationarity. Absence of covariance stationarity can be addressed in several ways. This study employed first-order differencing, which can be effective in removing autocorrelative structure but implies that the resulting time series reflect changes in activity rather than absolute activity levels. An alternative approach is prewhitening (Box, Jenkins, & Reinsel, 1994; Langheim, Leuthold, & Georgopoulos, 2006), which involves fitting a separate univariate regression model to each time series and performing any subsequent multivariate analysis on the residuals only. Finally, since short time series are more likely to be covariance stationary than long time series, a third approach would be to break each time series into several comparatively short segments. Although this approach has the advantage of allowing analysis of time-varying causal influences (Hesse, Möller, Arnold, & Schack, 2003), the reduced number of samples in each time series risks that statistical inferences of causality may be weaker.

The present analysis is highly conservative with respect to covariance stationarity for the reason that we included in each time series the maximum number of data points possible. We assessed the robustness of our results to this factor by repeating our analysis after splitting each time series into two equal parts. We found that the causal cores corresponding to each time series segment were nearly identical to the causal cores corresponding to the full series, suggesting that our approach was not excessively conservative.

4.7 Application to Other Neural Simulations. Although our results were generated by analyzing a subset of CA1 neuronal units in a particular brain-based device (Darwin X), it should be emphasized that our analysis can be applied to a broad range of neural simulations. Such applications could usefully assess the generality of the results described here. Some of these results—for example, the particular constitution of context networks and causal cores—may be expected to vary according to the details of the analyzed system. Other results may be more general—for example, the correspondence of causal core refinement with behavioral learning and the small size of causal cores as compared to context networks.

Darwin X was based on a mean-firing rate neuronal model, in which the activity of each neuronal unit represented the combined activity of approximately 100 real neurons. Identifying causal cores in real neural systems will require extending our analysis to inferences of causality based on spiking neuronal dynamics, for example by combining point-process models with maximum likelihood models (Okatan, Wilson, & Brown, 2005). Such an extension could usefully be tested on future brain-based devices having spiking neuronal dynamics.
4.8 Application to Empirical Data. More generally, application of our analysis requires the identification of all (or a large fraction) of the connected and coactive precursors to a given NR. At present, this is fully achievable only in a computational model, and only in a brain-based device can an NR be meaningfully associated with real-world behavior. However, recent methodological advances suggest that investigators soon may be able to characterize both anatomical and functional architectures in biological systems at microscopic scales. For example, retrograde transneuronal transport of rabies virus has been used to characterize the microanatomy of connectivity between the basal ganglia and cortex (Kelly & Strick, 2004). Functional microarchitectures have been mapped using metabolic markers for neuronal activity (Konkle & Bielajew, 2004), high-resolution optical imaging (Raffi & Siegel, 2005), the genetic manipulation of neurons to express fluorescent voltage sensors (Ahrens et al., 2004) and the bulk-loading of calcium-sensitive indicators (Cossart et al., 2005; Ohki et al., 2005). Importantly, several of these techniques can be applied to behaving primates (Raffi & Siegel, 2005).

These advances reflect a changing experimental emphasis from the blind recording of neurons toward the detailed characterization of specific neurons as their dynamics evolve in time in behaving animals. New analysis methods will be required in order to make sense of the data generated by these new techniques. We suggest that our analysis provides one such method and that the insights that potentially could follow may best be exposed by prior application of the analysis to brain-based devices. Application of the novel experimental methods described above will also permit testing of several predictions of the model, in particular, the small size of causal cores as compared to context networks, and the correspondence between learning and causal core refinement.

Acknowledgments

We are grateful to Jeffrey Krichmar, Joseph Gally, and Eugene Izhikevich for many constructive discussions. We also benefited from the comments of two anonymous referees. We thank Dr. Krichmar in particular for his contribution to the design and implementation of Darwin X. This research was supported by grant N00014-03-1-0980 from the Office of Naval Research, by DARPA, and by the Neuroscience Research Foundation.

References


Received December 8, 2005; accepted July 12, 2006.