PERGAMON

Neural Networks 11 (1998) 1253–1258

Neural Networks

1998 Special Issue

Continuous attractors and oculomotor control

H. Sebastian Seung*

Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974, USA

Received and accepted 30 April 1998

Abstract

A recurrent neural network can possess multiple stable states, a property that many brain theories have implicated in learning and memory. There is good evidence for such multistability in the brainstem neural network that controls eye position. Because the stable states are arranged in a continuous dynamical attractor, the network can store a memory of eye position with analog neural encoding. Continuous attractors in model networks depend on precisely tuned positive feedback, and their robust maintenance requires mechanisms of synaptic plasticity. These ideas may have wider scope than just the oculomotor system. More generally, the internal models postulated by theories of biological motor control may be recurrent networks with continuous attractors. © 1998 Published by Elsevier Science Ltd. All rights reserved.

Keywords: Recurrent networks; Continuous attractors; Learning and memory; Motor control; Internal models; Multistability; Positive feedback; Reverberating activity

Recurrent feedback loops pervade the synaptic connectivity of the brain. One possible role of these feedback loops is to endow neural networks with multiple stable states, or dynamical attractors (Amit, 1995). Previously, I have marshalled the evidence for such multistability in the *oculomotor integrator*, the brainstem neural network that controls eye position (Seung, 1996). Here I would like to present similar arguments to readers who are familiar with neural network theory but may not have a background in oculomotor control. A secondary goal is to discuss the wider relevance of dynamical attractors to motor control in general.

The multistability observed in the oculomotor integrator has chiefly been modeled with linear networks in which tuned positive feedback sustains reverberating activity patterns (Cannon et al., 1983; Galiana and Outerbridge, 1984; Seung, 1996). Multistability in these linear networks is of a special sort: the stable fixed points are arranged in a *continuous attractor*. This continuity is required for consistency with the analog, graded encoding of eye position in neural activity.

Multistability is not a generic property of linear networks; it requires precise tuning of synaptic strengths and other parameters. Could nonlinearity obviate the need for precise tuning? Recent research has shown that network models of the integrator can be constructed from spiking, conductance-based model neurons (Lee et al., 1997). But the resulting nonlinear network is no less dependent than linear networks on precise tuning, because the need for tuning is deeply rooted in the structural instability of a continuous attractor. Although robust multistability without precise tuning could easily be implemented in nonlinear networks with discrete point attractors separated by high dynamical barriers (Hopfield and Tank, 1986), this scenario is not compatible with the analog coding of eye position observed in the integrator.

The dependence on precise tuning has two major implications. The first is that any biological system can be expected to fall short of the ideal continuous attractor. Instead it will have an approximation to the ideal: a continuous manifold on which drift is very slow, but not exactly zero. The second is that robust maintenance of a continuous attractor requires mechanisms of synaptic plasticity. Adaptive changes in synaptic strengths could maintain the tuning of positive feedback, as proposed in learning network models (Arnold and Robinson, 1997; Seung, 1997). Further progress in modeling would be aided by experimental characterization of the types and time scales of synaptic plasticity in the integrator.

According to modern computational theories, biological motor control is performed by an internal model of the motor plant embedded inside a negative feedback loop

* Corresponding author. E-mail:seung@bell-labs.com

PII: S0893-6080(98)00064-1

(Jordan, 1995). Unfortunately, there is little agreement as to the precise nature of the internal models, or even where in the brain they might reside. With the lack of direct physiological evidence for internal models, it is difficult to resolve these questions.

The state of affairs is much better in the oculomotor system, owing to three decades of single unit recording in awake, behaving animals. A wealth of experimental data indicates that the internal model used for maintaining eye position is the integrator, which has been localized to specific brainstem nuclei. The nature of the internal model is also known: it appears to be a recurrent network with a continuous attractor.

I would like to suggest that the internal models of other motor systems are also recurrent networks with continuous attractors. To further develop this theory, it will be necessary to understand how proprioceptive input interacts with the recurrent network dynamics.

In summary, continuous attractors seem computationally suited for representing the continuous variability in motor control. Similarly, I have argued elsewhere that perceptual variability can be represented by continuous attractors in recurrent networks (Seung, 1998; Ben-Yishai et al., 1995).

1. The memory of eye position

This paper is restricted to the neural basis of the simplest oculomotor behavior, which is observed in the dark with the head fixed. Under these conditions, normal humans are able to hold their eyes stationary at arbitrary positions for up to tens of seconds (Becker and Klein, 1973; Hess et al., 1985). To perform this task, the brain maintains a constant level of activation of the extraocular muscles in the absence of sensory feedback. In other words, the brain maintains a memory of the current eye position.

The brain area that stores this memory is called the oculomotor integrator (Robinson, 1989). In cats and monkeys, the integrator for horizontal eye position has been localized to two brainstem nuclei, the nucleus prepositus hypoglossi (NPH) and the medial vestibular nucleus (MVN) (Moschovakis, 1997). An analogous brainstem nucleus has recently been identified in goldfish (Pastor et al., 1994). There are four types of evidence for localization. First, anatomical studies show that integrator neurons project to extraocular motor neurons in the abducens nucleus that control horizontal eye position. Second, neural activity in the integrator encodes horizontal eye position. Third, lesion or inactivation of the integrator impairs or destroys the ability to hold the eyes still. Fourth, microstimulation of the integrator causes persistent changes of eye position.

In the dark and with the head fixed, the memory of eye position is independent of sensory input. The primary vestibular afferents from the semicircular canals carry only their tonic activity when the head is fixed. In the dark there is no visual motion signal from the retina. While proprioceptive afferents may carry some information about eye position, this information does not appear to be used by the brain for the control of eye position, as there is no stretch reflex in the extraocular muscles (Keller and Robinson, 1971). In short, the memory of eye position is maintained centrally, and does not depend on a sensorimotor feedback loop.

Neural activity x_i in the integrator takes the form

$$x_i = k_i(E - E_i), \tag{1}$$

where i denotes the index of the neuron. The slope k_i is called the position sensitivity. The threshold position for firing is E_i. Above threshold, the rate is linear in eye position, and below threshold it is zero. It is incorrect to regard the activity patterns of Eq. (1) as being caused or determined by eye position. Rather, they are stable states that are intrinsic to the integrator and do not depend on sensory information. The integrator is able to maintain a memory of eye position because it has multiple stable states of Eq. (1), one for each eye position E.

2. Multistability in linear networks

The multistability embodied in the rate–position relationship (Eq. (1)) may be a collective property that emerges from the interactions between neurons in a network. The simplest network model embodying this possibility is completely linear (Cannon et al., 1983; Galiana and Outerbridge, 1984; Seung, 1996),

$$\tau\dot{x}_i + x_i = \sum_j W_{ij}x_j + b_i. \tag{2}$$

The activity of each neuron is described by a continuous variable x_i, which is the instantaneous rate of action potentials from neuron i. The synaptic time constant τ reflects the fact that neurons communicate through synapses, which temporally filter the trains of discrete action potentials.

Neuron i is driven by recurrent feedback from other neurons in the network, where the strength of the connections is given by W_{ij}. Without recurrent feedback, the rate of neuron i would take on the value of the bias term b_i, which includes the effect of intrinsic pacemaker currents or tonic drive from extrinsic sources.

What conditions must W_{ij} and b_i satisfy in order for the network to be multistable? The fixed points of a linear network satisfy N linear equations in N unknowns, written in matrix form as

$$(\mathbf{I} - \mathbf{W})\mathbf{x} = \mathbf{b}. \tag{3}$$

Generically, $\mathbf{I} - \mathbf{W}$ is nonsingular, and there is a unique fixed point given by $\mathbf{x} = (\mathbf{I} - \mathbf{W})^{-1}\mathbf{b}$. In other words, multistability is not a generic property of a linear network.

However, in special cases the linear system (Eq. (3)) can be underdetermined, i.e. the linear network can have more than one fixed point. This occurs when $\mathbf{I} - \mathbf{W}$ is singular,

and **b** is in its range space. Since **b** is in the range space of $\mathbf{I} - \mathbf{W}$, there exists at least one solution of Eq. (3). Given this single solution, a whole family of solutions can be generated by adding vectors from the null space of $\mathbf{I} - \mathbf{W}$, which is defined as the set of all solutions of the homogeneous equation $(\mathbf{I} - \mathbf{W})\mathbf{x} = 0$.

In other words, the set of fixed points of the network is the affine space obtained by displacing the null space of $\mathbf{I} - \mathbf{W}$ from the origin. The stability of this set of fixed points depends on the eigenvalues of **W**, as explained in detail elsewhere (Seung, 1996). Here we note only that the null space is spanned by the right eigenvectors of **W** with eigenvalues that are precisely unity. Stability is guaranteed when the rest of the eigenvalues have real parts that are less than unity.[2]

Then the set of fixed points is a continuous attractor of the dynamics (Eq. (2)). It is an attractor, because any trajectory of the dynamics converges to some point on the set. It is also continuous, with the same dimensionality as the null space of $\mathbf{I} - \mathbf{W}$.

3. Continuous attractors and neural coding

The preceding general discussion of continuous attractors has prepared us for the task of explaining the particular form (Eq. (1)) of multistability observed in the oculomotor integrator. We assume that the null space of $\mathbf{I} - \mathbf{W}$ is one-dimensional, so that the network dynamics has a one-dimensional line of fixed points. Let $\boldsymbol{\xi}$ be a solution of the homogeneous equation,

$$(\mathbf{I} - \mathbf{W})\boldsymbol{\xi} = 0 \tag{4}$$

Then all vectors in the null space of $\mathbf{I} - \mathbf{W}$ are multiples $E\boldsymbol{\xi}$, where E is a scalar. The bias vector is assumed to be in the range space of $\mathbf{I} - \mathbf{W}$, so that some solution $\mathbf{x}^0$ of Eq. (3) exists. Then all solutions of Eq. (3) can be written in the form

$$x_i = x_i^0 + E\xi_i \tag{5}$$

This is identical in form to Eq. (1), provided we identify ξ_i with the position sensitivity k_i, and $x_i^0 = -k_i E_i$.

Not only do linear network models offer an explanation of the persistence of activity patterns in the integrator, they also explain the relationship between the neural encoding of eye position and the synaptic weights of the network. Namely, the vector of position sensitivities is proportional to the vector $\boldsymbol{\xi}$ that solves the equation $(\mathbf{I} - \mathbf{W})\xi = 0$.

4. Descent on an energy landscape

For every stable linear network, there exists a Lyapunov function, a quadratic function of the x_i that is bounded below and decreases everywhere except at fixed points (Slotine and Li, 1991). Then the network dynamics Eq. (2) can be visualized as descent on the energy landscape defined by the Lyapunov function.[3] The energy landscapes sketched in Fig. 1 are graphs in which the vertical coordinate is the value of the Lyapunov function, while the horizontal coordinates are two of the many activities x_i of the neurons in the network.

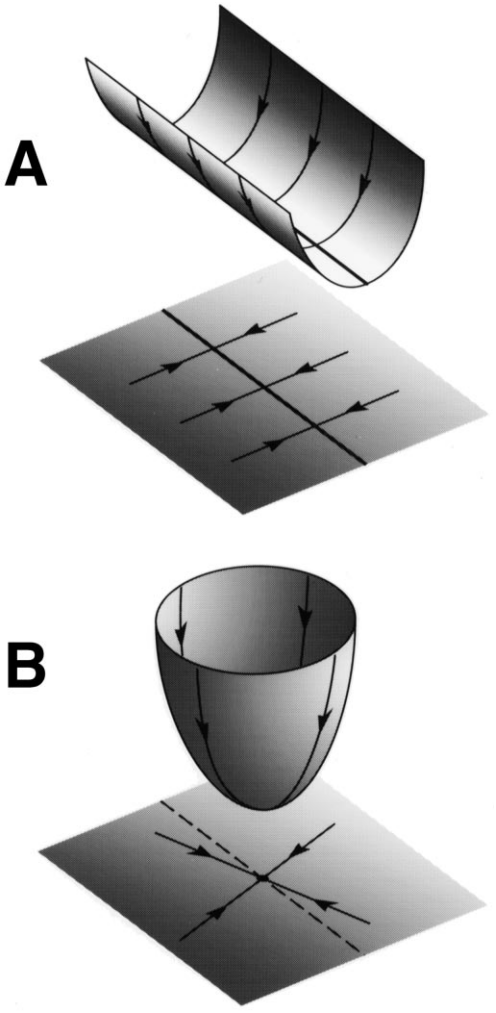

Fig. 1. Graphs of energy functions of stable linear networks. (a) Line attractor — in this special case, all dynamical trajectories converge to a line of fixed points at the bottom of a trough-shaped energy landscape. (b) Point attractor — this is the generic case in which all trajectories converge to a unique fixed point. Reproduced from (Seung, 1996). Copyright (1996) National Academy of Sciences, USA.

For the special case of a line attractor, the energy landscape is trough-shaped, as shown in Fig. 1(a). Every trajectory of the dynamics converges to some point on the continuous line of minima. Because the network may remain at any location along the bottom of the trough, it is able to store a memory of a single scalar variable. The

[2] Throughout this paper, the term 'stability' is used in the sense due to Lyapunov (Slotine and Li, 1991).

[3] In general the dynamics performs descent, but not steepest descent unless the network connections are symmetric.

direction vector of the line at the bottom of the trough is in the null space of $\mathbf{I} - \mathbf{W}$, or equivalently an eigenvector of $\mathbf{W}$ with unity eigenvalue.

This multistability contrasts with the generic case, in which there is a single stable fixed point, or point attractor. The energy landscape is shaped like a bowl, as shown in Fig. 1(b). Every trajectory of the dynamics converges to the unique minimum of the energy landscape.

5. Robustness

The existence of a line attractor in a linear network depends on two types of precise tuning: the null space of $\mathbf{I} - \mathbf{W}$ must be one-dimensional, and its range space must contain the bias vector $\mathbf{b}$. Generic small changes in $\mathbf{b}$ and $\mathbf{W}$ cause these conditions to be violated, destroying the line of fixed points. The effects of mistuning can be visualized using the energy landscape. In a precisely tuned network, the bottom of the trough in Fig. 1(a) is perfectly flat and level. Mistuning the $\mathbf{b}$ and $\mathbf{W}$ parameters causes the bottom of the trough to deviate from this ideal, resulting in drift along the bottom.

Because of this dependence on tuning, it is unrealistic to expect an ideal line attractor in a biological neural network. Rather, one should expect an approximation to a line attractor, a line of points at which the drift velocity is slow. In accord with this expectation, the memory of eye position in human subjects is gradually corrupted over time. During gaze-holding in the dark, there is generally a slow systematic drift, usually less than one degree per second in normal subjects (Becker and Klein, 1973; Hess et al., 1985).

Furthermore, we should expect a graded loss of memory function in a biological neural network with a continuous attractor that is subjected to lesion or inactivation. The severity of dysfunction should correlate roughly with the extent of the lesion or inactivation. Such graded loss is seen in experiments (Moschovakis, 1997).

These experimental observations lend qualitative support to the idea that the integrator is a tuned system. It is important to make quantitative measurements of sensitivity to mistuning, and compare them with the predictions of network models. The theoretical prediction is that the precision of tuning required to produce a given memory time constant is set by its ratio to the synaptic time constant τ. To achieve a memory time constant of 10 s, one percent accuracy in tuning is required, assuming that $\tau = 100$ ms. It is still not clear whether this level of precision is biologically plausible.

6. Nonlinear networks

Although linear network models are very useful, it is important to construct more biophysically realistic models. Simulations of the integrator based on spiking, conductance-based model neurons have been carried out (Lee et al., 1997). Two important theoretical questions about nonlinearity arise in these simulations. First, is it possible to implement continuous attractors in nonlinear networks? Second, can nonlinearity make continuous attractors more robust?

The synaptic connections in the spiking, conductance-based model were tuned by first reducing it to a rate-based model of the form

$$\tau \dot{x}_i + x_i = f\left(\sum_j W_{ij} x_j + b_i\right), \tag{6}$$

which is a simple modification of the linear dynamics (Eq. (2)). The function f is the relationship between firing rate and current, which has a threshold nonlinearity, and a slight sublinearity above threshold (Serafin et al., 1991; du Lac and Lisberger, 1995). A low rank form $W_{ij} = \xi_i \eta_j$ was assumed for the synaptic connection matrix, and an optimization algorithm was used to reduce the drift velocity $\dot{x}_i$ along a line in state space. It was not possible to tune W_{ij} and b_i to realize a perfect line attractor (Seung, 1996; Lee et al., 1997). However, it was possible to realize some approximation, a line in state space along which drift is slow.

The parameters b_i and W_{ij} obtained by minimizing drift in the rate-based model were then inserted into the original network of conductance-based model neurons. Simulations of these spiking neurons showed that the network exhibited activity of the form (Eq. (1)) that persisted for long periods of time.

The rate–position relationship in the nonlinear network had a threshold nonlinearity, and was therefore more realistic than the purely linear rate–position relationship (Eq. (5)) of linear network models. However, the nonlinear network still required precise tuning of parameters, much like a linear network.

7. Learning networks

A solution to the problem of robustness might be found in synaptic plasticity, which could establish and maintain the precise tuning required by a continuous attractor. Several types of synaptic learning rules have been theorized, though there is still no direct physiological evidence of synaptic plasticity in the integrator.

Visual feedback from retinal slip is a candidate error signal for driving synaptic plasticity in the integrator. A synaptic learning rule based on presynaptic activity and retinal slip has been used to self-organize a network (Arnold and Robinson, 1997). Some behavioral evidence for the effect of visual feedback on the integrator has been obtained using visual–vestibular conflict stimuli (Tiliket et al., 1994).

However, experiments with dark-reared rabbits (Collewijn, 1977) and cats (Harris and Cynader, 1981)

have shown that the integrator develops much of its functionality without any visual experience at all. Therefore, a learning rule that does not depend on retinal slip has been proposed (Seung, 1997). The synaptic changes are Hebbian, in the sense that they depend on both pre- and postsynaptic activity. This type of learning can be regarded as driven by an error signal derived from an efference copy of eye position.

A third possibility, so far unexplored in modeling, is that proprioceptive feedback is used as an error signal to drive synaptic plasticity. Some behavioral evidence for a role for proprioceptive feedback in oculomotor adaptation has been reported (Lewis et al., 1994).

8. Internal models as recurrent networks

I have argued that continuous attractors are relevant for oculomotor control. Can this concept be more generally applied to other motor systems? Current computational theories of motor control postulate an internal model of the motor plant embedded in a negative feedback loop (Jordan, 1995). The negative feedback aspect of these theories has some physiological basis in studies of the role of proprioceptive afferents in spinal stretch reflexes.

In contrast, there is little direct physiological evidence for internal models. Their existence is inferred indirectly from the fact that animals deprived of sensory feedback from both proprioceptive and visual modalities do not suffer complete loss of motor abilities. No one knows where the internal models are located in the brain, although the cerebellum is a popular guess. The nature of the internal model is also a matter of great debate. Popular hypotheses include forward and inverse models, which are representations, often as layered feedforward neural networks, of the relationship between position commands and motoneuron activation.

In the oculomotor system, the integrator can be regarded as an internal model. The location of this internal model is known, unlike in other motor systems. As a steady eye position can be maintained without proprioceptive or visual feedback, the quality of the internal model is very good. And physiological studies of this internal model indicate that it is a recurrent neural network with a continuous attractor.

Continuous attractors might also be found in the internal models of other motor systems. For example, an internal model for limb control could contain a continuous attractor consisting of the stable activity patterns required to hold the limb steady at various positions. The dimensionality of the attractor would be equal to the number of degrees of freedom of the limb. The major challenge for such a theory is to understand how proprioceptive input interacts with the dynamical attractor.

9. Conclusion

The stable patterns of activity observed in the oculomotor integrator during gaze-holding can be interpreted as fixed points on a continuous dynamical attractor. Precise tuning of positive feedback produces a continuous attractor in a linear network. The continuity of the set of fixed points reflects the continuous, analog nature of the neural coding of eye position (1). Current theoretical work focuses on more complex models that include the nonlinearities of excitable membranes and synaptic conductances, and also on synaptic learning mechanisms that could robustly maintain tuning.

Unlike other motor systems, the oculomotor system lacks a stretch reflex. Because of this special feature, the oculomotor integrator in the dark can be regarded as an internal model stripped of its surrounding negative feedback loop. Consequently oculomotor researchers have been able to focus on the nature of the internal model, which has led to the realization that it is based on a continuous attractor. This discovery could have relevance for motor control in general. The internal models of other motor systems may also turn out to be recurrent networks with continuous attractors.

Acknowledgements

I have benefited from helpful discussions with David Tank, Haim Sompolinsky, Dan Lee, Emre Aksay, and Sam Wang. This work was supported by Lucent Technologies.

References

Amit, D.J. (1995). The Hebbian paradigm reintegrated: local reverberations as internal representations. *Behavioral and Brain Sciences*, *18*, 617–626.

Arnold D.B., & Robinson D.A. (1997). The oculomotor integrator: testing of a neural network model. *Experimental Brain Research*, *113*, 57–74.

Ben-Yishai, R., Bar-Or, R.L., & Sompolinsky, H. ((1995). Theory of orientation tuning in visual cortex. *Proceedings of the National Academy of Sciences USA*, *19*, 3844–3848.

Becker W., & Klein H.-M. (1973). Accuracy of saccadic eye movements and maintenance of eccentric eye positions in the dark. *Vision Research*, *13*, 1021–1034.

Cannon S.C., Robinson D.A., & Shamma S. (1983). A proposed neural network for the integrator of the oculomotor system. *Biological Cybernetics*, *49*, 127–136.

Collewijn H. (1977). Optokinetic and vestibulo-ocular reflexes in dark-reared rabbits. *Experimental Brain Research*, *27*, 287–300.

du Lac S., & Lisberger S.G. (1995). Membrane and firing properties of avian medial vestibular nucleus neurons in vitro. *Journal of Comparative Physiology*, *A176*, 641–651.

Galiana H.L., & Outerbridge J.S. (1984). A bilateral model for central neural pathways in vestibuloocular reflex. *Journal of Neurophysiology*, *51*, 210–241.

Harris L.R., & Cynader M. (1981). The eye movements of the dark-reared cat. *Experimental Brain Research*, *44*, 41–56.

Hess K., Reisine H., & Dursteler M. (1985). Normal eye drift and saccadic drift correction in darkness. *Neuro-ophthalmology*, *5*, 247–252.

Hopfield J.J., & Tank D.W. (1986). Computing with neural circuits: a model. *Science*, *233*, 625–633.

Jordan, M.I. (1995). Computational motor control. In M.S. Gazzaniga (Ed.), *The cognitive neurosciences*. Cambridge, MA: MIT Press.

Keller E.L., & Robinson D.A. (1971). Absence of a stretch reflex in extraocular muscles of the monkey. *Journal of Neurophysiology*, *34*, 908–919.

Lee, D.D., Reis, B.Y., Seung, H.S., & Tank, D.W. (1997). Nonlinear network models of the oculomotor integrator. In J.M. Bower (Ed.), *Computational neuroscience: Trends in research*. New York: Plenum Press.

Lewis R.F., Zee D.S., Gaymard B.M., & Guthrie B.L. (1994). Extraocular muscle proprioception functions in the control of ocular alignment and eye movement conjugacy. *Journal of Neurophysiology*, *72*, 1028–1031.

Moschovakis A.K. (1997). The neural integrators of the mammalian saccadic system. *Frontiers in Bioscience*, *2*, d552–d577.

Pastor A.M., de La Cruz R.R., & Baker R. (1994). Eye position and eye velocity integrators reside in separate brainstem nuclei. *Proceedings of the National Academy of Sciences USA*, *91*, 807–811.

Robinson D.A. (1989). Integrating with neurons. *Annual Review of Neuroscience*, *12*, 33–45.

Serafin M., de Waele C., Khateb A., Vidal P.P., & Mulethaler M. (1991). Medial vestibular nucleus in the guinea-pig. I. Intrinsic membrane properties in brainstem slices. *Experimental Brain Research*, *84*, 417–425.

Seung H.S. (1996). How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences USA*, *93*, 13339–13344.

Seung H.S. (1997). Learning to integrate without visual feedback. *Society for Neuroscience Abstracts*, *23 (1)*, 8.

Seung H.S. (1998). Learning continuous attractors in recurrent networks. *Advances in Neural Information Processing Systems*, *10*, 654–660.

Slotine, J.-J. & Li, W. (1991). *Applied nonlinear control*. Englewood Cliffs, NJ: Prentice Hall.

Tiliket C., Shelhamer M., Roberts D., & Zee D.S. (1994). Short-term vestibulo-ocular reflex adaptation in humans. I. Effect on the ocular velocity-to-position neural integrator. *Experimental Brain Research*, *100*, 316–327