

SHAWN R. LOCKERY AND TERRENCE J. SEJNOWSKI

*Computational Neurobiology Laboratory, Salk Institute for Biological Studies,
La Jolla, California,
and The Howard Hughes Medical Institute***I. Introduction**

Whether the interest is in discovering the neural basis of behavior or in reverse engineering the nervous system to reveal its secrets of computation and control, modeling and simulation play a central role in the process of discovery. Many interesting behaviors are subserved by large, nonlinear, and highly interconnected neural networks that are too complicated to grasp intuitively. Modeling of complex networks could be used to gain insight into their biological counterparts. However, such models typically contain many free parameters that cannot be set by the available physiological or anatomical data. One approach to these difficulties is to choose values for these parameters using an optimization algorithm constrained by biological data. This chapter illustrates several different applications of one such algorithm called backpropagation (Rumelhart *et al.*, 1986), a widely used gradient descent technique, to the well-defined neural circuit of the local bending reflex of the leech. After introductory remarks on optimization in network modeling, we review our use of optimized network models to demonstrate the plausibility of distributed processing in the local reflex. We next show how varying the assumptions of the model led to unexpected local bending networks involving dedicated rather than distributed processing mechanisms. A final section demonstrates the use of optimization to study how the memory for nonassociative conditioning can be stored in distributed

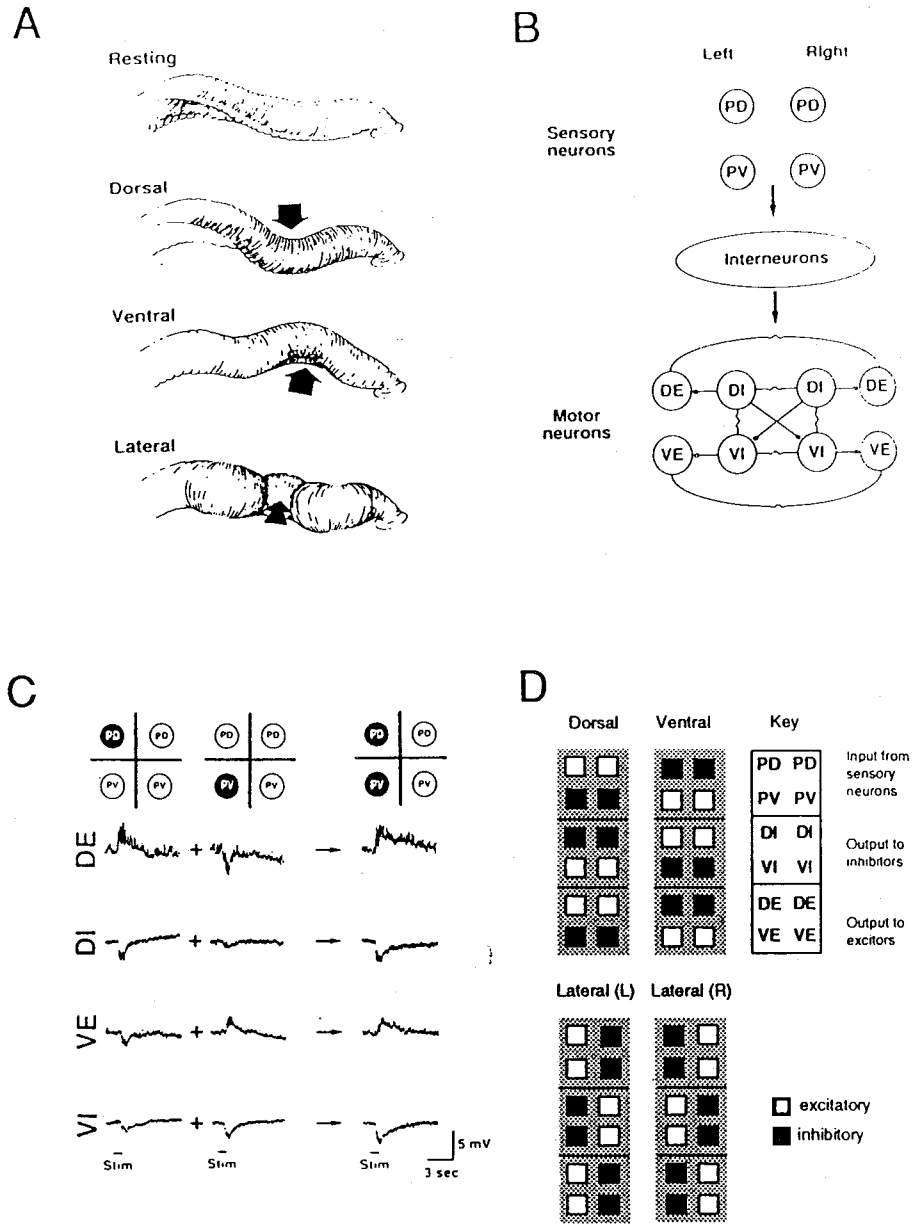


FIGURE 2. Network model of the local bending reflex of the leech. (A) Behavior: dorsal, ventral, and lateral stimuli produce local U-shaped bends. (B) Simplified neural circuit: the main input to the reflex is provided by the dorsal and ventral P cells (PD and PV). Control

fit to the biological data is obtained. More efficient random search strategies include simulated annealing (Kirkpatrick *et al.*, 1983) and genetic algorithms (Goldberg, 1989; Beer and Chiel, Chapter XII, this volume). In gradient descent methods, the effect of each parameter on the fit of the model to the data is determined and parameters are changed in the direction that improves the fit, i.e., reduces the error of the model. In numerical differentiation, a simple gradient descent method, parameters in the model are increased one at a time by a small fraction. If this reduces the error, the change is retained; if not, the opposite change is made. As in random methods, this procedure is repeated until a satisfactory fit is obtained. Backpropagation is a more efficient procedure for computing the derivatives in which the changes required for all the parameters are calculated simultaneously. Once the derivatives have been calculated, they can be used to update the parameters as in numerical differentiation. Many variations of gradient descent are available, such as conjugate gradient (Battiti, 1992) and methods using second derivatives (Parker, 1987).

III. The Local Bending Reflex

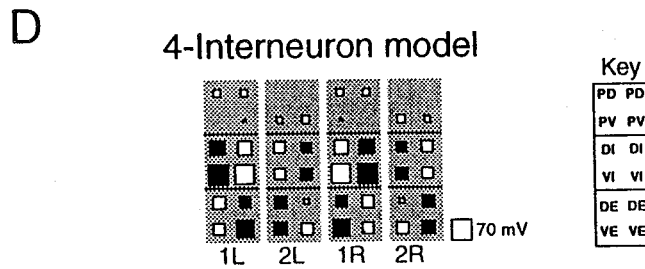
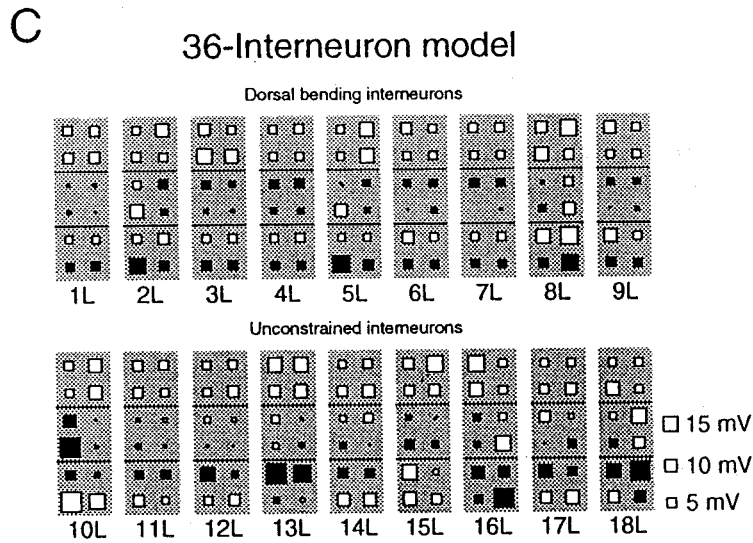
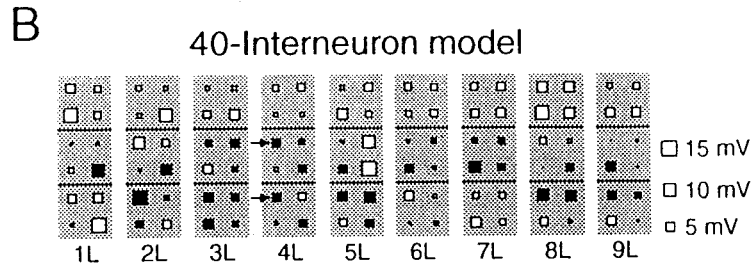
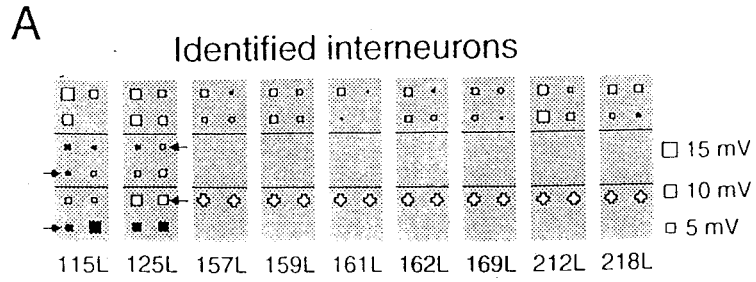
We use backpropagation as means of searching the parameter space associated with a model of the local bending reflex in the leech. In response to a moderate mechanical stimulus, the leech withdraws from the site of contact (Fig. 2A). This is accomplished by contracting longitudinal muscles beneath the stimulus and relaxing longitudinal muscles on the opposite side of the body, resulting in a local U-shaped bend (Kristan, 1982). Major input to the local bending reflex is provided by dorsal or ventral pressure-sensitive mechanoreceptors or P cells (Fig. 2B, PD and PV; Nicholls and Baylor, 1968). Contraction and relaxation of longitudinal muscles are controlled by a total of eight types of motor neurons, an excitatory

FIGURE 2. (cont.) of local bending movements is largely provided by motor neurons whose projective field is restricted to one quadrant (left or right, dorsal or ventral) of the body. Dorsal and ventral quadrants are innervated by both excitatory (DE and VE) and inhibitory (DI and VI) motor neurons. Inhibitors inhibit excitors of the same body quadrant, and dorsal inhibitors inhibit contralateral ventral inhibitors (filled terminals). (C) Physiological input-output function: intracellular recordings from the four motor neurons in response to stimulation of one or two P cells (filled circles). The motor neurons shown have projective fields ipsilateral to the stimulated P cell(s). Similar recordings were obtained with other patterns of P cell stimulation and from contralateral motor neurons (not shown). (D) Hypothetical local bending interneurons dedicated to the detection of dorsal, ventral, and left (L) and right (R) lateral stimulus locations. Each interneuron has effects on motor output that are consistent with withdrawal from the stimulated site. White boxes represent excitatory connections; black boxes represent inhibitory connections. The presynaptic or postsynaptic neuron for each connection is given in the key.

(DE or VE) and inhibitory (DI or VI) type for the dorsal and ventral quadrants on the left and right side of each body segment (Fig. 2B; Stuart, 1970; Ort *et al.*, 1974). The task of the interneurons in the local bending reflex is to compute a behavioral input–output function: the mapping relation between patterns of P cell activation and patterns of motor neuron excitation and inhibition sufficient for the animal to withdraw from the stimulus. The input–output function has been studied experimentally by making intracellular recordings from each of the eight types of motor neuron in response to P cells stimulated singly or in dorsal, ventral, or lateral pairs (Fig. 2C) (Lockery and Kristan, 1990a).

In a simple model of the local bending reflex, dorsal, ventral, and lateral bends are produced by types of interneurons specific for each form of the response (Fig. 2D). To determine how the interneurons in the reflex computed the local bending input–output function, a subpopulation of local bending interneurons contributing to dorsal local bending was identified using physiological and morphological criteria. Nine types of dorsal bending interneuron, which have excitatory connections to DE and receive excitatory connections from PD (Lockery and Kristan, 1990b), were found. Interestingly, several aspects of the other connections made by this subpopulation (Fig. 3A) are inconsistent with a commitment to only dorsal local bending and thus with the simple model (Fig. 2D). First, all but one type of dorsal

FIGURE 3. (opposite) (A) Average connection strengths of identified local bending interneurons (Lockery and Kristan, 1990b). Interneurons are numbered according to their location on the standard map of the leech midbody ganglion (Muller *et al.*, 1981). The left (L) member of each left–right pair of interneurons is shown, except for cell 218, which is unpaired. Symbols as in Fig. 2D, except that box area is proportional to synaptic strength determined from pairwise intracellular recordings. White plus signs indicate excitatory connections of unknown strength determined from extracellular recordings of DE. Blank spaces indicate connections that have not been determined because the presynaptic neuron lies on the ventral surface of the ganglion while the postsynaptic neuron lies on the dorsal surface. The connections are not consistent with the dedicated interneuron model (Fig. 2D). (B) Connection strengths of interneurons 1L to 9L in a 40-interneuron model after optimization. Like all other interneurons in the model, these are excited by ventral as well as dorsal stimuli and have connections to most motor neurons. Thus the connections of model interneurons are qualitatively similar to the connections of identified local bending interneurons. (C) The 36-interneuron model. Interneurons 1L–9L (dorsal bending interneurons) were constrained to receive four excitatory P cell inputs and have outputs to excitatory motor neurons consistent with dorsal bending. Interneurons 10L–18L (unconstrained interneurons) were constrained only to receive 4 excitatory P cell inputs; no constraints were placed on the sign or amplitude of their output connections. After optimization to the local bending data set, most of the unconstrained interneurons had developed connections to the excitatory motor neurons that were consistent with ventral bending. (D) The 4-interneuron model. Both members of each left–right pair are shown. Networks with fewer interneurons could not be optimized to produce local bending motor output.



bending interneuron receive substantial excitatory input from PV, indicating that these neurons are also active in ventral and lateral local bends. Second, the connections from an interneuron to the inhibitory motor neurons are not always opposite in sign to its outputs to the excitatory motor neurons controlling the same body quadrant (Fig. 3A, interneuron 125, arrows). Thus, the connections of the subpopulation of local bending interneurons suggest a distributed processing strategy in which each interneuron is active in some or all forms of local bending; motor neuron excitation and inhibition would thus result from balanced combinations of appropriate and inappropriate inputs from many interneurons acting in concert.

IV. The Distributed Model of Local Bending

Modeling the reflex was prompted by the need to demonstrate that the distributed processing hypothesis is consistent with the responses of the interneurons and with the physiological details of the input-output function of the reflex. The possibility remained that another type of interneuron, as yet undiscovered, is required to produce accurately the known set of local bending input-output relations. The model has 4 sensory neurons, 8 motor neurons, and 40 interneurons and thus 480 connections, representing the actual local bending circuit (Fig. 2B) (Lockery and Sejnowski, 1992). This is referred to as the 40-interneuron model. The number of interneurons was based on an estimate of the number of local bending interneurons that remain to be identified in the biological network. Each neuron in the model is represented as a single electrical compartment (Segev *et al.*, 1989) with a physiologically determined input resistance and a time constant. The membrane potential is updated as a function of time and depends only on the synaptic current injected by chemical and electrical synapses.

The large number of connections in the model entails a parameter space too large to search by hand. Therefore, the backpropagation algorithm was used to adjust the connections. To make the model more realistic, backpropagation was forced to operate within additional physiological constraints. First, only excitatory connections were allowed from sensory neurons to interneurons in the model, because only excitatory connections have so far been found between sensory neurons and interneurons in the biological network (Lockery and Kristan, 1990b). Second, the sigmoidal function for interneurons and motor neurons was shifted so that the output of a unit that receives no net input is zero, as in leech neurons (Granzow *et al.*, 1985). Third, each interneuron on the left was paired with one on the right to maintain homologous input and output connections, in accordance with the overall bilateral symmetry of connections in the leech nervous system. Fourth, no connections between interneurons were allowed. Fifth, the model

included all the chemical and electrical connections between the motor neurons. Synaptic weights from sensory neurons to interneurons and from interneurons to motor neurons were adjusted by the algorithm until the amplitudes and time courses of synaptic potentials recorded in the model motor neurons in response to each pattern of sensory input matched the smoothed and scaled replicas of the physiological synaptic potentials (Fig. 2C).

After training, the input and output connections of hidden units (Fig. 3B) in the model network qualitatively resemble the connections of identified local bending interneurons (Fig. 3A). In particular, interneurons receive inputs from ventral as well as dorsal input units, most have connections to all motor neurons, and the connections to the inhibitory motor neurons are not always opposite in sign to those onto the excitatory motor neurons controlling the same body quadrant (Fig. 3B, interneuron 4L, arrows). The similarity between model hidden units and interneurons in the biological network shows that the local bending input-output function can be achieved with interneurons similar to those identified physiologically. Additional interneurons with receptive and projective fields (output targets) that differ radically from the subpopulation of identified interneurons are not required. In hundreds of optimization runs from different initial positions in weight space, a different final point in weight space was reached each time. Thus, there are many different points in weight space that produce a physiologically accurate local bending input-output function utilizing a distributed processing strategy for computing the input-output relations.

V. Distributed Models of Local Bending with Functionally Specific Interneuronal Subpopulations

The correspondence between the identified interneurons and the interneurons in the 40-interneuron network establishes the possibility of using a distributed processing model to account for the input-output function of the reflex. However, several aspects of the output connections of the identified interneurons suggest that the actual network may use a strategy intermediate between the fully distributed solution of the 40-interneuron network and the dedicated interneuron solution of Fig. 2D. Consistent with the dedicated solution, all nine types of interneuron excite the DE motor neurons, and interneurons 115 and 125 also inhibit the VE motor neurons, effects that quite possibly are shared by the identified interneurons whose output connections have not yet been measured. On the other hand, both the inputs and the outputs to the inhibitory interneurons are consistent with a distributed solution, as noted above. Thus the identified interneurons are likely to constitute a subpopula-

tion that operates in a partly dedicated and partly distributed fashion. This suggests a new version of the model having two subpopulations of interneurons. The first, like the identified interneurons, is partially dedicated to dorsal bending. The second subpopulation might be partially dedicated to ventral or lateral bending.

To determine whether such a model can account for the input-output function of the reflex, a population of 36 model interneurons was divided into two separate subpopulations of 18 interneurons (Lockery and Sejnowski, 1992). We used 18 interneurons in this subpopulation to reflect the fact that nine types of interneurons have been identified that are partly dedicated to dorsal bending and that all but one of these comprises a pair of bilaterally symmetrical interneurons. During optimization, interneurons in the first subpopulation, referred to as dorsal bending interneurons, were constrained to excite DE and inhibit VE. Interneurons in the second subpopulation had no such constraints and were referred to as unconstrained interneurons. In both subpopulations, the constraints on the input connections and symmetry were the same as in the 40-interneuron model. After optimization, the performance of the 36-interneuron network was identical to that of the 40-interneuron model. Inspection of the connections of the subpopulation of model dorsal bending interneurons showed that they were like the identified dorsal bending interneurons, in accordance with the additional constraints placed on this subpopulation (Fig. 3C). Inspection of the unconstrained interneurons showed that most (61%) had output connections to DE and VE that were consistent with ventral bending and had mixed effects on the inhibitory motor neurons. The other major type of interneuron either excited or inhibited all four excitatory motor neurons. These results show that the local bending input-output function can be computed by networks with a separate subpopulation that corresponds closely to the identified interneurons. The unconstrained interneurons complement the effects of the dorsal bending interneurons and suggest possible connectivities of as-yet-unidentified interneurons in the biological network. In repeating this simulation many times, none of the networks contained interneurons with outputs consistent with lateral bending. Thus, this type of interneuron is not necessary for computing the input-output function.

VI. Minimal Local Bending Networks

To determine whether 36 interneurons are required for local bending or whether a smaller number would suffice, we used optimization to seek solutions having fewer interneurons (Lockery and Sejnowski, 1992). To increase the likelihood that a solution would be found, the requirement that each interneuron have an input

from all four sensory neurons was removed. The symmetry constraint was retained, but no constraints were placed on the output connections. Networks with fewer than four interneurons could not be optimized to produce recognizable local bending motor output patterns. Therefore, the minimum number of interneurons appeared to be four (Fig. 3D). However, we cannot rigorously exclude the possibility that in the networks with fewer than four interneurons the optimization procedure became trapped in a local minimum. That the local bending input-output function can be produced by as few as four interneurons indicates a high degree of redundancy may be present in the biological network.

The four-interneuron networks require two pairs of interneurons to accommodate three basic types of local bending: dorsal, ventral, and lateral. Enforcing this requirement led to hitherto unexpected mechanisms for computing the input-output function. While some of the four-interneuron networks used variations of the distributed processing network, some of the networks had interneurons that were specific for particular patterns of sensory input and motor output. The interneurons in the network shown in Fig. 4 were specific for dorsal or ventral inputs. Surprisingly, the same interneurons had exactly the motor outputs expected of lateral bending interneurons. Thus, the local bending input-output function can be produced by a novel solution involving dedicated interneurons in which there is a dissociation between the sensory and motor specificities of the two types of interneurons.

VII. Possible Engrams in Nonassociative Conditioning of the Local Bending Reflex

At the level of individual reflexes, learning can be defined as a change produced by experience in the input-output function of the underlying neural network. Learning in many systems is thought to be the result of changes in synaptic strength. Thus, when a reflex is conditioned, the network is moved to a point in weight space associated with a new input-output function. A major objective in the cellular analysis of learning and memory is to identify the sites of synaptic plasticity underlying the change in input-output function, a task that has been referred to as a search for the engram (Squire, 1987). Ideally, from an experimental point of view, the new point in weight space will be far from the original point so that the engram will comprise many large, hence easily detectable, changes. However, one might imagine that the same change in input-output function could be achieved by moving a much shorter distance in weight space. If so, then learn-

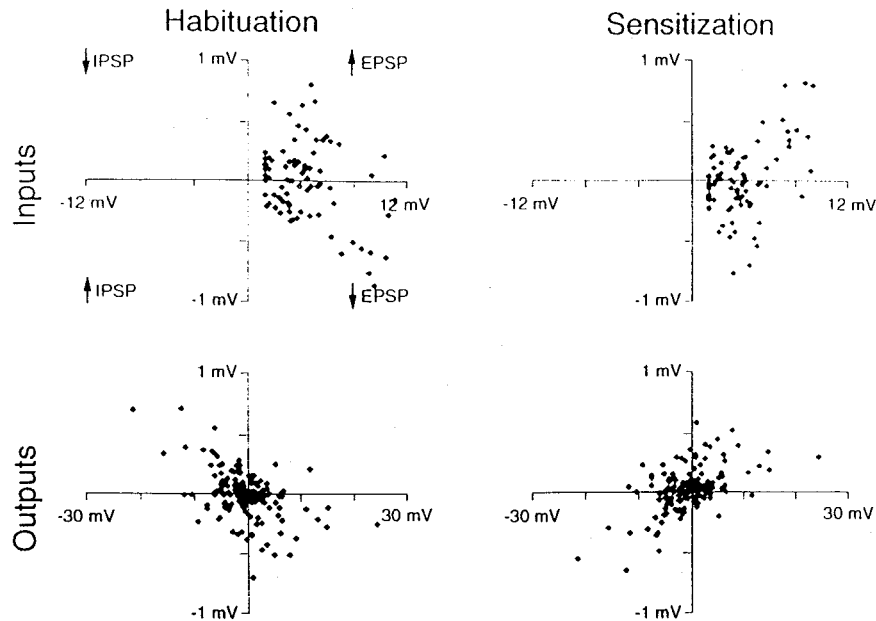


FIGURE 4. Scatter plots of the changes in synaptic strength produced by reoptimization. In each panel, the strength of a connection before reoptimization is plotted on the abscissa. The ordinate gives the change in that connection (before reoptimization – after reoptimization). Thus, connections whose strength increased fall in the upper right and lower left quadrants, while connections whose strength decreased fall in the lower right and upper left. Input connections were constrained to be positive in the model, hence there are no points on the left in the two upper panels. These plots show that habituation and sensitization were produced by the combined effect of increases and decreases in synaptic strength.

ing would be the result of some number of small changes in synaptic strength and the engram could thus be difficult to detect.

The local bending reflex exhibits several forms of nonassociative learning, including sensitization, warm-up, and habituation (Lockery and Kristan, 1991, and unpublished results). Little is known about the engrams for nonassociative learning in distributed processing systems. We therefore sought to examine the characteristics of engrams produced by using backpropagation to reoptimize the connections in a normal local bending network to produce habituated or sensitized local bending responses. Because backpropagation makes small changes in weights at each iteration, this approach was expected to yield solutions in which the final differences in synaptic strengths were small, if such solutions exist for the local bending network.

Backpropagation was thus used as a means of searching weight space for habituated or sensitized networks that were close to the original network, not as a model for the underlying mechanisms of synaptic plasticity whereby the learning is induced or retained. This approach could provide a worst-case scenario: if learning is distributed as widely as possible among the interneurons, how small are the changes in synaptic strength likely to be?

In these simulations we assumed that habituation entails a 50% reduction in the peak amplitude of the motor neuron synaptic potentials in each output pattern in the training set and that sensitization entails a 50% increase. Starting with the 40-interneuron network optimized for normal local bending, we reoptimized this network to the habituated or sensitized state. The change in synaptic strength at each synapse was determined by measuring the difference in the the peak of the simulated synaptic potential in response to a standard stimulus in the presynaptic neuron before and after reoptimization.

In reoptimizing six different 40-interneuron networks for habituation, the average change in synaptic strength (absolute value) was 0.22 mV; for sensitization it was 0.20 mV. The changes in synaptic strength were visualized in scatter plots where the change in synaptic strength was plotted against the strength of the connection before reoptimization (Fig. 4). In such a plot, the connections that increased in strength fall into the upper right and lower left quadrants, those that decreased in strength fall into the upper left and lower right, and unchanged connections lie along the abscissa. The scatter plots show that the engram produced by backpropagation was widely distributed, since almost every input and output connection in the network changed. A simple model of nonassociative learning predicts that habituation is due to decreases in synaptic strength and sensitization to increases in synaptic strength. For habituation, the scatter plots revealed that while most of the changes were consistent with the simple model, many increases in synaptic strength also occurred, in both the input and output connections of the interneurons. A similar effect was noted in sensitization, where many decreases in synaptic strength occurred. Taken together, these results show that, for each normal local bending network model, there exist nearby positions in weight space associated with habituated or sensitized motor output. Moreover, the nearby solutions involve a mixture of increases and decreases in synaptic strength, regardless of whether motor output increases or decreases in the learning.

The existence of habituated and sensitized networks involving many small changes raises the question of whether such changes would be detectable in practical physiological experiments. This was addressed by asking how much of the change in motor output could be accounted for by all the changes that were larger

than a given sensitivity threshold. In a quantal analysis of a central synapse in the leech, Nicholls and Wallace (1978) were able to resolve differences in synaptic potentials as small as 0.25 mV. At this level of resolution, approximately 40% of the learning encoded by the distributed engrams produce by backpropagation would be detectable. This sets an approximate lower bound on the detectability of nonassociative learning in the local bending reflex.

VIII. Conclusion

We have used backpropagation as an optimization algorithm to explore the weight space associated with a model of the distributed processing of sensory information in the local bending reflex of the leech. In optimizing the 40-interneuron model we found, as in other networks, that there are many different points in weight space that produce a physiologically accurate input-output function. In restricting the algorithm to smaller regions of weight space by limiting the value of interneuron output weights to observed ranges, as in the 36-interneuron networks, we found that qualitatively different networks with populations of dorsal ventral bending interneurons are also possible. A further restriction to the still smaller region of weight space defined by a network with only four interneurons showed that this was the minimum number of interneurons necessary and revealed unexpected types of dedicated interneurons. Finally, in reoptimizing networks to produce habituated or sensitized local bending responses, we found that the memory for nonassociative learning in distributed processing networks can involve many small changes at almost every weight in the network, a situation that could be hard to uncover in practical physiological experiments.

Whether the local bending reflex operates as any of these models suggests will require identification of the as-yet-undiscovered local bending interneurons and measurement of their input and output connections strengths. Whether memory is encoded as reoptimization suggests will require identifying the sites of synaptic plasticity underlying nonassociative learning in the reflex. Whatever the results, these prior explorations of the local bending weight space provide a framework in which to place the actual biological solutions and thus deepen our understanding of the solutions nature has chosen. Far from being limited to well-defined invertebrate networks, this approach is a general one that can be applied to any neural system for which the input-output function is known or can reasonably be assumed. It should therefore be useful in a great variety of modeling studies (Zipser and Andersen, 1988; Lehky and Sejnowski, 1988; Anastasio and Robinson, 1990; Fetz

et al., 1990; Servan-Schreiber *et al.*, 1990; Tsung *et al.*, 1990; Zipser, 1991; Krauzlis and Lisberger, 1991; Pouget *et al.*, 1992).

Acknowledgments

Supported by NIH and HHMI postdoctoral fellowships.

References

- ANASTASIO, T. J., and ROBINSON, D. A. (1990). *Biol. Cybern.* **63**, 161–167.
- BATTITI, R. *Neural Comp.* **4**, 141–166.
- FETZ, E. E., SHUPE, L. E., and VENKATESH, N. M. (1990). *International Joint Conference on Neural Networks*. San Diego, CA, 675–679.
- GOLDBERG, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.
- GRANZOW, B., FRIESEN, W. O., and KRISTAN, W. B., JR. (1985). *J. Neurosci.* **5**, 2035–2050.
- HODGKIN, A. L., and HUXLEY, A. F. (1952). *J. Physiol.* **117**, 500–544.
- KIRKPATRICK, S. C., GELATT, D. J., and VECCHI, M. P. (1983). *Science* **220**, 671–680.
- KRAUZLIS, R. J., and LISBERGER, S. G. (1991). *Science* **253**, 568–571.
- KRISTAN, W. B., JR. (1982). *J. Exp. Biol.* **96**, 161–180.
- LEHKY, S. R., and SEJNOWSKI, T. J. (1988). *Nature* **333**, 452–454.
- LOCKERY, S. R., and KRISTAN, W. B., JR. (1990A). *J. Neurosci.* **10**, 1811–1815.
- LOCKERY, S. R., and KRISTAN, W. B., JR. (1990B). *J. Neurosci.* **10**, 1816–1829.
- LOCKERY, S. R., and KRISTAN, W. B., JR. (1991). *J. Comp. Physiol. A* **168**, 165–177.
- LOCKERY, S. R., and SEJNOWSKI, T. J. (1992). *J. Neurosci.* (to appear).
- MIALL, C., and WOLPERT, D. (1990). In Selverston, A. I. (ed.), *Neural Computation* (pp. 77–88). Society for Neuroscience, Washington, DC.
- MULLER, K. J., NICHOLLS, J. G., and STENT, G. S. (1981). *Neurobiology of the Leech*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
- NICHOLLS, J., and WALLACE, B. G. (1978). *J. Physiol. (Lond.)* **201**, 157–170.
- NICHOLLS, J. G., and BAYLOR, D. A. (1968). *J. Neurophysiol.* **31**, 740–756.
- ORT, C. A., KRISTAN, W. B., JR., and STENT, G. S. (1974). *J. Comp. Physiol. A* **94**, 121–154.
- PARKER, D. B. (1987). *First International Conference on Neural Networks II*, San Diego, CA, pp. 593–600.
- POUGET, A., FISHER, S. A., and SEJNOWSKI, T. J. (1992). In Moody, J. E., Hanson, S. J., and Lippmann, R. P. (eds.), *Advances in Neural Information Processing Systems*, Vol. 4, Morgan Kaufmann Publishers, San Mateo, CA, (pp. 412–419).
- RUMELHART, D. E., HINTON, G. E., and WILLIAMS, R. J. (1986). *Nature* **323**, 533–536.
- SEGEV, I., FLESHMAN, J. W., and BURKE, R. E. (1989). In Koch, C., and Seger, I. (eds.), *Compartmental Models of Complex Neurons* (pp. 63–96). MIT Press, Cambridge, MA.
- SEJNOWSKI, T. J., and ROSENBERG, C. R. (1987) *Complex Syst.* **1**, 145–168.
- SERVAN-SCHREIBER, D., PRINTZ, H., and COHEN, J. D. (1990). *Science* **249**, 892–895.

- SQUIRE, L. R. (1987). *Memory and Brain*. Oxford University Press, New York.
- STUART, A. E. (1970). *J. Physiol.* **209**, 627-646.
- TSUNG, F. S., COTTRELL, G.W., and SELVERSTON, A. I. (1990). *International Joint Conference on Neural Networks I*, San Diego, CA, pp. 169-174.
- ZIPSER, D. (1991). *Neural Comp.* **3**, 179-193.
- ZIPSER, D. and ANDERSEN, R. A. (1988). *Nature* **331**, 679-684.