The Emergence of Orientation Selectivity in Self-Organizing Neural Networks

Josh McDermott

Introduction

The past three decades have witnessed a host of impressive findings concerning the functional architecture of the brain. One trend that much recent research suggests is that there is a high degree of specialization within the brain: different parts of the brain do different, highly specific tasks. Furthermore, most estimates place the number of cortical neurons alone at 10^{10}, with roughly 10^{14} total synapses connecting them. Given this enormous number of neurons, and considering the extreme structural complexity embodied by a healthy and fully-developed mammalian brain, the project of elucidating this structure in some ways pales in comparison to that of explaining how it came to be. It is clear that the mammalian genome "cannot, in any naive sense, contain the full information necessary to describe the brain" (von der Malsburg, 1990). Yet somehow, in the process of development, the brain acquires its structure. The compelling nature of this quandary makes the notion of self-organization very appealing.

The term "self-organization" is generally used to describe the evolution of complex behavior in systems that consist solely of many very simple parts. The brain is an excellent example of such a system, for while neurons are by no means simple, many of their principle functional characteristics are believed to be relatively easy to approximate and model. In this review I will focus on research that suggests that the early stages of the mammalian visual system can be implemented solely through the self-organizing properties of neural networks.

Central to all such research is the assumption that synaptic modification occurs via some form of Hebbian learning, whereby the strength of a synapse between two neurons is increased if the activities of the neurons are correlated in time, that is, if one tends to be active at the same time that the other is. This learning rule was first suggested by Hebb in 1949 without an accompanying brain mechanism or body of evidence for its existence. Since then, the discovery of *long-term potentiation* (LTP) in the hippocampus has provided evidence that a form of Hebbian learning does occur in the brain, although how it is accomplished is still not clear. While skepticism about LTP remains, it is generally agreed that some form of local learning is bound to control connectivity in many cases, if only because the idea of completely central control seems inconceivable. Most efforts to show that self-organization can occur in structures resembling the brain thus make use of Hebbian learning.

Vision research has been among the most fruitful areas of neuroscience, and the basic structure of the early levels of the mammalian visual system are for the most part established. The research that I will be discussing concerns the visual system up to and including the primary visual cortex (V1). The visual pathway begins at the retina, where there are several layers of cells which project via the optic nerve to the lateral geniculate nucleus (LGN) in the thalamus, which in turn projects to layer 4C of V1. Cells in the inner layers of the retina and in the LGN are characterized by spatial-opponent receptive fields: Stimuli placed in a central, circular region of the receptive field of a cell in these regions tend to excite the cell, while stimuli placed outside the excitatory region tend to inhibit the cell. In addition, the retina and LGN are topographic: the spatial layout of their cells' receptive fields is topologically similar to the physical layout of the cells themselves. V1 cells differ from those in the LGN or retina in that many of them have orientation selective receptive fields; a cell in V1 will tend to respond strongest to a line of a particular orientation placed in its receptive field. Furthermore, as the Nobel Prize-winning research of David Hubel and Torsten Wiesel revealed, orientation selective cells are laid out in an orderly fashion. In addition to being retinotopically organized, a group of cells whose receptive fields correspond to a certain point in space will be arranged in a line such that their orientation preferences vary continuously along that dimension in cortex. Many have speculated that this layout has the favorable characteristic that each small portion of V1 has all the machinery necessary to perform the first steps in extracting information from the portion of the visual field which it represents. The research reviewed here is exciting because it suggests that many of the characteristics I have just described can be accomplished through usupervised self-organization.

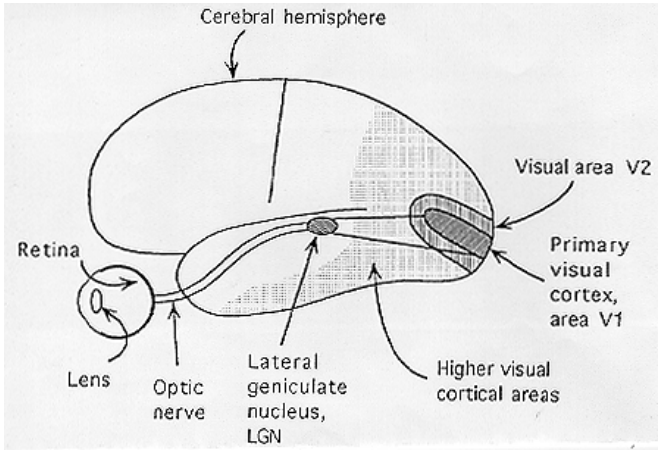

Figure 1. Schematic of the visual system. (Taken from Rose, D. A portrait of the brain. In *The Artful Eye* (1995), Gregory, Harris, Heard, and Rose, Eds.)

A host of experiments have shown the influence of the environment on the development of the nervous system. Although many results in this literature are controversial, it is

generally agreed that kittens raised in visually-altered environments developed abnormal patterns of orientation selective cells in V1 (see Blakemore and Cooper, 1970, Hirsh and Spinelli, 1970 for the first such experiments, and Movshon and Van Sluyters, 1981 for a review). Such results suggested to some that orientation selectivity was the result of learning processes in the visual system during postnatal development. This sparked several papers that described the development of orientation selectivity with Hebbian learning in neural networks that were exposed to structured "visual" input (see von der Malsberg, 1970, for one early attempt). However, with a few exceptions, animals in most of the classic visual plasticity studies do seem to develop *some* sort of orientation selectivity regardless of the visual environment (see Hirsh and Spinelli for a controversial exception). The environment seems only to be able to skew the distribution of orientation selectivity, rather than completely determine it, suggesting that the visual system is intrinsically biased towards developing cells with this property. Furthermore, Hubel and Wiesel originally found that some degree of orientation selectivity exists in kitten primary visual cortex immediately after birth, before exposure to any structured visual input (Hubel and Wiesel, 1963; Movshon and Van Sluyters, 1981). More recently it has been found that orientation selectivity is fully developed at birth in monkeys and sheep (Wiesel and Hubel, 1974; Ramachandran, Clarke, and Whitteridge, 1977). Based on this it was generally accepted that although the environment may be capable of affecting the properties of cells in the early levels of the visual system, it is not their source. This left the problem of how the visual system achieved orientation selectivity unsolved.

Self Organizing Neural Nets: Ralph Linsker's Work

A series of three papers by Ralph Linsker and the work by several others that followed them provide a possible explanation for the existence of prenatal orientation selectivity. In these papers Linsker demonstrated that spatial opponency and orientation selectivity could arise in an unsupervised feed-forward network trained with Hebbian learning on completely unstructured input. Linsker further showed that with the addition of lateral excitatory connections within the layer that developed orientation selective cells, the cells would develop their orientation preferences in a smoothly varying fashion in orientation columns similar to those in V1. These are surprisingly sophisticated properties for such a simple system to develop in the absence of structured input, and they will be the focus of this article.

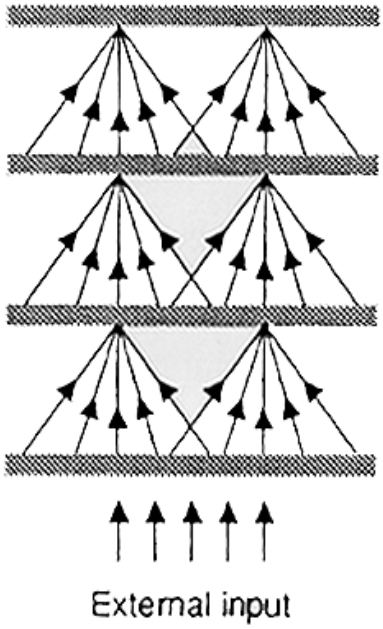

Figure 2. A feedforward neural network. The converging arrows from layer to layer represent the circular Gaussian distribution of input connections that each unit receives from cells in the previous layer. (Taken from Hertz, Krogh, and Palmer (1991) *Introduction to the Theory of Neural Computation*)

Linsker's network is a crude approximation of the visual system, with a two-dimensional input layer feeding to successive two-dimensional layers that can be interpreted as corresponding to different levels in the visual pathway. Each layer is composed of linear units and receives input from the previous layer. The inputs for a unit come from a Gaussian distribution over a local region of the previous layer. In the simulations he published, Linsker allowed the weights on connections from a given unit to take both positive and negative values. While this is biologically unrealistic, in that neurons are believed to be either excitatory or inhibitory, Linsker has reported that his results hold when the units are divided into classes which are constrained to have either only positive or only negative weights on their outgoing connections. Weights were also constrained to remain within certain positive and negative limits, which, based on physiological limitations, is biologically reasonable.

Linsker uses a version of the Hebbian learning rule:

[1] $\Delta w_{ij} = b + c(a_i - d)(a_j - e)$

where w_{ij} is the strength of the connection from neuron j to neuron i, a_i and a_j are the outputs of neurons i and j, respectively, and b, c, d, and e are constants. The importance of this equation is that the weight change is proportional to the product of a_i and a_j. Thus the weight change is most positive if a_i and a_j are correlated over time, and is most negative if they are anticorrelated. In order to prevent his simulations from being prohibitively long, Linsker averaged Equ. 1 over a number of presentations to the input layer, resulting in an equation for the time-rate-of-change of a given synaptic weight which could be solved for the mature weight values of a particular layer. His averaged equation can be understood as follows: Δw_{ij} is directly related to the correlation between the activities of the postsynaptic (labeled i) and the presynaptic (labeled j) neurons, and the activity of the postsynaptic neuron is a linear function of all its inputs. Thus the time-rate-of-change of w_{ij} is proportional to the degree to which the activity of neuron j is correlated with the other neurons that give input to neuron i. Explicitly,

[2] $\partial w_{ij}/\partial t \approx p + m\Sigma_k w_{ik} + \Sigma_k (Q_{jk} w_{ik})$

where p and m are constants, k indexes neurons that provide input to neuron i, and Q_{jk} is proportional to the correlation function of the activities at neurons j and k. This function is well-defined because each layer of cells receives input only from the preceding layer, and the layers are developed one at a time. Formulating the weight change this way allowed

Linsker to run his simulations more efficiently. Importantly, though, it explicitly reveals what turns out to be a crucial consequence of Hebbian learning in feedforward neural networks: change of the weights into units of a given layer is determined by the form of the correlation in the activities of the units of the previous layer.

The weights to layer B are developed based on the activity in the input layer A, and once the weights in layer B reach stable values, the activity in layer B, propagated from layer A, is used to develop the weights in layer C, and so on.

Given this enormous number of neurons, and considering the extreme structural complexity embodied by a healthy and fully-developed mammalian brain, the project of elucidating this structure in some ways pales in comparison to that of explaining how it came to be.

Because the network is developed in this manner, each layer's weights are a function only of the pattern of activity in the previous layer. Also important is the property of Equation 2 that either all or all but one of the inputs to any given cell will saturate to their limiting values. This property falls directly out of the equation; see Linsker 1986a for the proof.

Spatial Opponency

When the activity in layer A is uncorrelated, the correlation function is close to zero for most pairs of neurons. Thus $\partial w_{ij/\partial t}$ is close to constant for each connection from layer A to a cell in layer B. For appropriately chosen constants, this results in the saturation of all the weights from layer A to layer B at the upper positive limit. Because each cell in layer B receives input from a Gaussian distribution of cells in layer A, a given cell in layer B will, at maturity, function to compute a spatial average of a local region of activity in layer A. And because the receptive fields of nearby cells in layer B overlap, cells that are close together will include many of the same layer A neurons in their average, and will thus have correlated activities. Linsker shows that this correlation function is a Gaussian (whose peak is at the cell in question - obviously, the highest correlation is between a cell and itself - and falls off for cells whose receptive fields are far away).

Once the connections from layer A to layer B were mature, Linsker developed layer C using the Gaussian correlation function of layer B. He found that the morphology of layer C cells fell into a series of regimes depending of the values of p and m. Cells developed to have either i) all excitatory inputs, ii) all inhibitory inputs, iii) "ON-center" circularly symmetric opponent connections: a core of excitatory connections surrounded by a ring of inhibitory connections, iv) "OFF-center" circularly symmetric opponent connections: a core of inhibitory connections surrounded by a ring of excitatory connections, and v) spatially divided inputs such that approximately one side of the receptive field was composed of excitatory and the other half of inhibitory connections. I will focus on case (iii), where layer C develops center-surround receptive fields strikingly similar to those found in the retina and lateral geniculate nucleus, the two stages in the visual pathway that lead to primary visual cortex.

The spatial opponent receptive fields develop because of the presence of the Gaussian correlation function of layer B in Equation 2. For negative m and positive p, the following occurs during the maturation process: positive p values cause all weights to initially increase such that the w_{ik} contribution to $\partial w_{ij/\partial t}$ is positive. Then, since the sum in Equation 2 is over the synapses to the postsynaptic neuron i in layer C, which have a Gaussian distribution in layer B, $\partial w_{ij/\partial t}$ has a greater contribution from the correlations that the activity of neuron j has with neurons in the central region of neuron i's receptive field, and a smaller contribution from the correlations that it has with peripheral neurons, simply because there are fewer of them. But since the correlation function Q is a Gaussian for layer B, a given layer B neuron is most correlated to neurons nearby, and less with neurons farther away. Thus if neuron j is in the central region of neuron i's receptive field, its connection to neuron i will be increased more than will a peripheral neuron's. Although the correlation function between a given neuron and the neurons surrounding it are identical for all neurons, the Gaussian distribution of afferent inputs to postsynaptic cells causes the high portion of a peripheral neuron's correlation function to be sampled less frequently than is the high portion of a central neuron.

As a comparison, consider what would happen if Q were a constant function. Then each neuron in layer B would be equally correlated with all the other neurons in layer B. Were this the case, the Gaussian distribution of afferent inputs to the neuron i in layer C would not have the effect that it does, since the Q contribution to the sum in Equation 2 would be independent of where it was sampled.

Thus $\partial w_{ij/\partial t}$ is larger for neurons in the center of neuron j's input distribution than for neurons in the periphery. Furthermore, the absolute magnitude of $\partial w_{ij/\partial t}$ can be shifted by varying the constant m. For sufficiently large negative m, $\partial w_{ij/\partial t}$ is negative for peripheral neurons, and positive for central ones. Since Equation 2 causes all of the weights to a given neuron to saturate, this causes the connections from layer B neurons in the central region of a layer C neuron's receptive field to mature to the maximum excitatory value, while those in the periphery saturate to the inhibitory limit.

To summarize, layer B computes the local spatial average of the random activity in layer A, because of the Gaussian distribution of inputs. This makes the correlation function for layer B Gaussian, and combined with the Gaussian synaptic distribution, for appropriate values of the constants in Equation 2, spatial opponency results. Thus there are two crucial dependencies: the proper values of p and m, and the Gaussian synaptic distribution. These will be discussed later.

Orientation Selectivity

Development of orientation selective cells proceeds in similar fashion. Parameters are chosen such that layer C develops into "ON-center" cells. Then the correlation function is computed for layer C. Since layer C is, up to small deviations, uniform, the correlation function (which in general form is a function of two variables, corresponding to any two cells in a layer) is effectively a function of one variable. That is, because of the layers uniformity, the correlations of a given cell Z with another cell Y depends only on the distance between cell Z and cell Y, and not on the actual position of the two cells in the layer. Linsker thus idealizes the correlation function Q_{ij} to the function Q(s), where s is the distance between two cells.

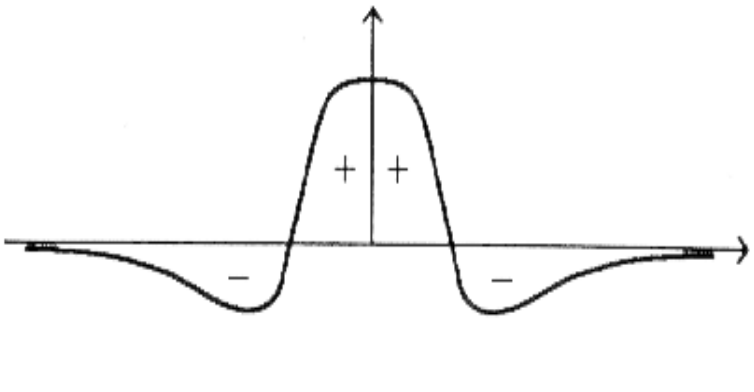

Figure 3. A "mexican-hat" function. For the correlation functions discussed in this article, the x-axis represents the distance between one unit and another, and the y-axis represents the correlation between the activities of the two units. (Taken from Hertz, Krogh, and Palmer (1991) *Introduction to the Theory of Neural Computation*)

For a layer of spatial-opponent cells, Q(s) has a "Mexican-hat form": Q is positive for small s, when cells are close to one another and their excitatory center cores overlap, negative for intermediate s, when the inhibitory surround of one cell overlaps the excitatory core of another, and zero for large s, when the cells are far apart and completely uncorrelated. This correlation function is then used to develop layer D. Linsker reported that there were a range of morphological options for layer D, including the formation of orientation selective cells, but that the regime of orientation selectivity was not very stable. One of the other regimes for layer D was spatial opponency, with the cells in layer D differing from those in layer C in that the "Mexican-hat" correlation function $Q_D(s)$ for layer D had deeper minima than did $Q_C(s)$. Linsker showed that "Mexican-hat" correlation functions with deeper minima produced in the following layer a more stable regime of orientation selectivity. He thus set the parameters of layer D such that it developed spatial opponent cells. He did this for layers D through F, with each successive layer having a Mexican-hat correlation function with deeper minima than the previous layer's function.

Linsker simply presents the results of his simulations, and does not discuss why the correlation functions develop deeper minima. However, some of the data he presents suggests that this deepening occurs because the cells develop receptive fields with larger inhibitory surrounds. Linsker gives the minimal values and the location of the zero-crossing for the Mexican-hat correlation functions of layers, and in addition to minimal values decreasing with successive layers, the zero-crossing appears to decrease. He also mentions that the minimal values of the correlation function vary inversely with the average sum $\sum_j w_{ij}$ of all the weights on the afferent connections to a unit i. In other words, there is a direct relationship between the relative proportion of inhibitory connections and the depth of the Mexican-hat minima. This is consistent with my observation that the zero-crossing decreases as the minima depth decreases, for as the zero-crossing decreases, the distance that two cells can be apart and still have positively correlated values decreases. But this happens for center-surround cells only if the center excitatory cores shrinks and the inhibitory surrounds grows. Furthermore, layers of cells with core and surround that are of approximately the same thickness will, logically, produce the deepest minima for their correlation function, because the core and surround can then completely overlap, causing the greatest anticorrelation. Since layer C cells have excitatory cores that are large compared to their surrounds, increasing the size of the inhibitory region in successive layers moves the two regions towards being of equal size, and acts to increase the maximum anticorrelation. Thus although Linsker doesn't actually discuss why the minima of the Mexican-hat function deepen with successive layers, it seems likely that the minima depth increases because the inhibitory surround of the cells' receptive fields grows in size.

Why do the inhibitory surrounds increase in area for progressive layers of cells? Again, Linsker doesn't discuss this, but the answer can be found by considering how the formation of center-surround cells from a Mexican-hat correlation function differs from the formation from the Gaussian correlation function of layer B. The main difference between the two functions is that the Mexican-hat has a smaller average value than does a Gaussian. Recall from my discussion of the center-surround cell formation process that the inhibitory regions form because the sum $\sum_k (Q_{jk} w_{ik})$ is smaller for peripheral connections than it is for central ones, and thus at some point in a cell's receptive field, $\partial w_{ij}/\partial t$ drops below zero, which sends w_{ij} to the inhibitory limit. If the average value of Q decreases, $\sum_k (Q_{jk} w_{ik})$ will decrease for all connections w_{ik}, and the threshold dividing excitatory and inhibitory connections will decrease, creating cells with smaller excitatory cores and larger inhibitory surrounds. Thus because layer C has a Mexican-hat correlation function, the cells in layer D develop smaller excitatory cores and larger inhibitory surrounds. But as discussed in the previous paragraph, increasing the size of the inhibitory surround in layer D causes Q_D to have deeper minima than Q_C. This in turn causes layer E to have larger

inhibitory surrounds (assuming that the parameters for layer E are set such that it develops center-surround cells), which causes layer E's Mexican-hat correlation function to be deeper than that of layer D, and so on.

Linsker reported that with a sufficiently deep correlation function for a layer of cells, the succeeding layer will develop orientation selective cells that are stable with respect to random changes in the initial weights. In the network he describes, he developed four layers of center-surround cells to obtain a suitably deep function.[1] He describes the results obtained by varying two parameters: the average sum $\sum_j w_{ij}$, which we will call g, and R_G, where R_G is the radius of the Gaussian distribution of afferent connections to a cell in layer G. Varying g is equivalent to varying the ratio of m and p, the two constants in Equation 2. Linsker found that g and R_G were the main parameters governing the development of orientation.

There were several broad classes of development characteristics, each corresponding to a range of R_G with respect to s_{min}, the location of the minimum value of the correlation function for layer F. I will discuss two of these classes. First, for R_G much less than s_{min}, layer G expressed morphology that was quite similar to that of layer C, which was discussed earlier.[2] Though Linsker does not discuss this, this class probably exists because when the radius inside which a cell draws its input is much smaller than the minimum of the Mexican-hat function, the cell can only "see" the central portion of the function. Equation 2, which is where the influence of the correlation function comes into play, sums over the correlation function only for s-values within the radius of the Gaussian input distribution, simply because the sum is over the cell's inputs. So if the radius of the Gaussian is much smaller than s_{min}, Equation 2 includes values of the correlation function for small s only, and over that restricted range, the Mexican-hat function resembles the Gaussian correlation function of layer C.

For R_G close to s_{min}, decreasing g results in a more interesting range of morphologies. For high g, the cells are obviously all-excitatory. As g is lowered, isolated regions of inhibitory connections appear. As g is decreased further, the G cells become bilobed: they have a central strip of excitatory connections, with parallel inhibitory bands of connections on either side. Such cells, clearly, are orientation selective. The orientation that a cell develops is random. As g is further decreased, the inhibitory side bands extend to enclose the excitatory region, thus forming a center-surround cell.

What causes orientation selectivity? Linsker answered this by referring to an energy function he created which has the property (common to other energy functions) of decreasing with every weight change. The weight development process is thus a process of gradient descent, and this sheds some light on the orientation formation. In order to avoid introducing new equations and variables, and in the interest of explaining the phenomenon in as biologically-grounded terms as is possible, I will attempt an explanation based on the weight-change equation instead of referring to Linsker's energy function. Recall the time-rate-of-change of a weight that was described by Equation 2:

[2] $\partial w_{ij}/\partial t \approx p + m\sum_k w_{ik} + \sum_k (Q_{jk}w_{ik})$.

The first two terms on the right hand side of Equation 2 are the same for all connections. The term $\sum_k (Q_{jk}w_{ik})$ is not, however. It will be similar for connections from cells that are nearby in layer F, because the correlation function will tend to multiply each weight by a similar number. For connections from cells that are further apart in layer F, $\sum_k (Q_{jk}w_{ik})$will tend to be different, because the correlation function for one connection will have a positive value when the correlation function of a second connection has a negative value. Thus there is an overall pressure within the development process for the connection from a cell in layer F to have a value equal to that of the connections from nearby cells, and different from those of connections from cells that are further away. However, from topological considerations this obviously cannot be realized in full. What happens instead is that contiguous regions of excitation or inhibition are formed which satisfy the pressure imposed by the correlation function as best as can be done. The stripes that produce orientation selectivity are one possible arrangement, and, as Linsker comments, they probably minimize the number of pairs of nearby cells which have different weight values, although no proof of this is given.

It is clear that the mammalian genome "cannot in any naive sense, contain the full information necessary to describe the brain." Yet somehow, in the process of development, the brain acquires its structure. The compelling nature of this quandary makes the notion of self-organization appealing.

Admittedly, this explanation of how orientation selectivity emerges is far from rigorous. However, Linsker's original papers sparked a great deal of interest, and several papers have been written that analyze his results. Linsker's results are nicely formalized by MacKay and Miller, who show that if the correlation function Q is turned into a matrix (a minor change in definition), the principle eigenvectors of this matrix resemble the weight ensembles in the various regimes of cell morphology that Linsker observed (MacKay and Miller, 1990) . The weight-change equation (Equation 2) can be viewed as multiplying a vector consisting of the weights of the connections to a cell by the correlation matrix of the preceding layer. The eigenvectors of a matrix M are the elements that are changed only by scalar multiplication when multiplied by M, and the principle eigenvector is the eigenvector that grows in length the most under multiplication by M. Thus it makes sense that the final weight configurations are the eigenvectors of the correlation matrix, because such configurations are the only ones that will maintain their form over development. MacKay

and Miller find that the principle eigenvectors of Gaussian covariance matrices tend to be of center-surround form, and that Mexican hat covariance matrices can have principle eigenvectors that are either center-surround or bilobed (having a positive stripe next to a negative stripe - i.e., morphologically similar to orientation selective receptive fields). Although MacKay and Miller do not state this explicitly, Miller's comments in another article (Miller 1990) indicate that the bilobed morphology (which would produce orientation selectivity) is the principle eigenvector when the Mexican hat covariance matrix has a narrow positive center (lower zero-crossings). This is in accord with Linsker's results, in that Linsker found that oriented bilobed cells were stable only in the higher levels of his simulations, when, as discussed above, the Mexican hat covariance function is of just this form.

The work of MacKay and Miller is additionally important because it reveals that cell morphologies in a layer of a feed-forward Hebbian networks are determined by the correlation matrix of activities in the previous layer. This provides an avenue for exploring self-organization, because once the activity patterns of a layer of cells is known, the eigenvectors of the correlation matrix can be computed, and compared to the receptive fields of cells in the next layer. If simple self-organization occurs, one might expect some match between the computed eigenvectors and the measured receptive fields.[3]

Orientation Columns

In the network described in the previous section, orientation preferences arose randomly from cell to cell, with no correlation between neighboring cells. It is a well-known feature of mammalian primary visual cortex that the orientation preference of cells varies in a continuous fashion over the cortical surface while remaining constant over displacements perpendicular to the surface (Hubel and Wiesel, 1963, 1974). The regularity of orientation distribution prompted Hubel and Wiesel to suggest that the basic unit of organization in V1 is the "hypercolumn," a group of columns of cells whose receptive fields represent the same point in visual space, with each column of cells having a different orientation and ocularity preference, thus providing the neural machinery to analyze a single point of the visual field (Mason and Kandel, 1991).

Linsker found that by adding excitatory lateral connections between the cells in layer G before the connections from layer F were developed, the orientation that a cell developed was no longer random, but that cells of a similar orientation organize into band-like regions. Due to computational constraints, he did not explore the full parameter space to see how closely the bands could come to resembling the patterns that are found in various species. Fortunately, his initial result has been pursued by other researchers, who have confirmed their robustness (Miller, 1992) . Furthermore, others have shown that another primary feature of V1, ocular dominance columns, can also arise from self-organization as a result of competition between the activity patterns of the two eyes (Miller, Keller, and Stryker, 1989). These results suggests that the principal features of V1 could result from very simple self-organization.

Regardless of whether orientation selectivity arises in the brain through mechanisms analogous to those described here, one of the most important insights to come out of the self-organization models is the importance of the covariance function of activities across a layer of cells whenever any type of Hebbian learning occurs in a system.

Discussion

There are several issues pertaining to Linsker's simulations that deserve mention. First, consider the assumptions implicit in Linsker's network. His model depends crucially on the afferent connections to each cell in a layer of his network having a Gaussian distribution about a point in the previous layer. The retinotopy of the early levels of the visual system is well established; the important question with regard to Linsker's model is whether retinotopy precedes the development of cell properties. When and how retinotopy arises in biological systems is not yet known, but there are several methods of developing retinotopy with unsupervised neural networks (see von der Malsburg 1990, and Hertz, Krogh, and Palmer 1991 for reviews). In any case, this assumption is by far the most reasonable of those implicit in Linsker's model.

A second element of the model that is less biologically reasonable is the uniformity of the cells - each cell can have both excitatory and inhibitory connections. In particular, each connection is treated identically by the development equations. Any given connection can assume the full range of excitatory and inhibitory values. Since it is usually agreed that neurons are either glutamatergic (excitatory) or GABAergic (inhibitory), this feature of the model is troubling. Linsker did this simply to speed the simulations. Since the development equation sends all the connections in an inhibitory region to the inhibitory limit and all the connections in an excitatory region to the excitatory limit, were there an even distribution of excitatory and inhibitory connections throughout a cell's transfer function the excitatory connections that ended up lying in the inhibitory region would be sent to zero, as would the inhibitory connections lying in the excitatory region. This basically wastes half the connections, which is why Linsker simplified the model by making each connection non-specific.

More significant is the fact that retino-geniculate and geniculate-cortical projections are believed to be purely excitatory, a fact not taken into account by Linsker's model. However, Ken Miller has conducted simulations with a far more biologically plausible model, where the layer corresponding to the LGN consists of two populations of center-surround cells,

ON and OFF, each of which make only excitatory connections to cells in the layer representing V1 (Miller 1989). Miller achieves results strikingly similar to those of Linsker. Miller's model achieves orientation selectivity via a mechanism inspired by his models of ocular dominance formation (Miller, 1992). Experiments depriving animals of the input from one eye suggest that in addition to there being a Hebbian component to synapse formation, there are also competitive influences: given two groups of correlated inputs, the strongest activated group may "win out" and develop connections of the greatest strength (Guillery, 1972). Making this assumption, Miller shows that when there are separate populations of ON- and OFF-center cells, and each cell in the next layer (representing V1) initially gets input from the same local retinotopic area within each population, bilobed receptive fields develop. This happens in his model because at small retinotopic distances, cells of the same type (ON or OFF) are correlated, while at larger distance, cells of opposite type are correlated. Combined with realistic interactions among his "cortical" cells, Miller finds that this model develops orientation selective cells whose arrangement within the layer is quite similar to cortical orientation maps. While this model seems vastly different from Linsker, the importance of correlation functions remains the crucial factor, and Miller's work thus suggests that in spite of the implausibility of Linsker's model, his results may reflect a property of neural networks that could also occur in the brain.

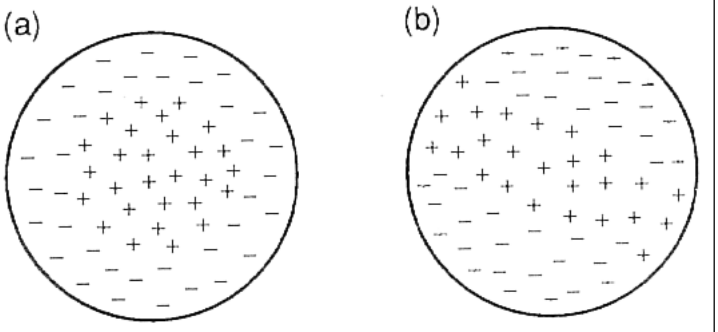

Figure 4. Idealized examples of weight structures for units in one of Linsker's networks. Each plus or minus-sign represents the sign of the weight of the connection *to* the unit in question *from* the unit in the preceding layer which occupies the position that the plus or minus-sign does in the diagram. A) A center surround weight structure. B) A bilobed, orientation-selective weight structure. (Taken from Hertz, Krogh, and Palmer (1991) *Introduction to the Theory of Neural Computation*)

Of key importance, then, to the biological relevance of these models, is the covariance functions of layer activity in the brain. Limited work has been done on this, but there is some indication that the appropriate correlations are present in the visual system (Miller, 1991)

A final feature of the model that is not explicitly grounded in biology is the layer-by-layer development process. Not enough is known about how multilayer neuronal systems develop in the brain to evaluate this facet of the model. Furthermore, similar weight patterns may result even if all layers are developed simultaneously

Ever since the discovery of orientation selectivity in V1, much research has been devoted to uncovering the mechanisms which underlie it. While there is not complete agreement as to these mechanisms, a host of studies are relevant to evaluating the extent to which Linsker's results actually explain orientation selectivity in V1. Here I briefly describe a few of them. Some of the most well-known studies have been conducted by Sillito, who showed that orientation selectivity was abolished if inhibition is blocked with the GABA antagonist bicuculline, suggesting that orientation selective cells receive both excitatory and inhibitory inputs, both of which are important for their function (Sillito, 1979). Because LGN cells are known to make only excitatory synapses with cells in V1, this was interpreted as showing that orientation selectivity is solely a product of cortical interactions. However, an alternative interpretation is that the application of bicuculline to an area of cortex leaves only excitatory connections between cortical cells. (thanks are due to Ken Miller for pointing this out to me) This could account for Sillito's result, in that the cells excited by lines of a certain orientation would then excite other nearby cortical cells, causing them to be active in spite of having orientation-tuned inputs from the LGN. The results of Nelson et al. support this view. They observed that orientation selectivity remains fully present in cat V1 cells when only the neuron being recorded from has its inhibitory inputs intracellularly blocked (Nelson et al. 1994). This suggests that excitatory inputs (which could presumably be coming from the LGN in agreement with the self-organizing models) are sufficient to generate orientation selectivity, and that intracortical inhibitory inputs mediate, if anything, an indirect effect through other neurons. David Ferster's research provides additional evidence that patterns of activity in LGN afferents do indeed play the crucial role in orientation selectivity that was originally proposed by Hubel and Wiesel. Ferster recorded intracellularly from cells in V1 while flashing oriented stimuli within a cell's receptive field, and by hyperpolarizing the cortical cell he was recording from to the IPSP reversal potential (done by injecting current through the recording electrode), Ferster enhanced the cell's EPSPs while suppressing its IPSPs. He found that the EPSPs of cortical cells are orientation tuned, in that rotating the stimulus away from the optimal orientation produced a decrease in the cell's EPSP (Ferster, 1986). While this demonstrates that excitatory connections are sufficient to produce orientation selectivity, the excitation could conceivably be cortical in origin, since there was no way for Ferster to determine its source. However, given the substantial excitatory input that V1 receives from the LGN, it seems likely that Ferster was observing LGN produced EPSPs, in which case the self-organizing models would be entirely consistent. While excitation is thus sufficient for orientation-selectivity, a role for inhibition is suggested by Hata *et al.*, who find through a cross-correlation analysis evidence

for inhibitory interactions between two simultaneously recorded neurons with slightly different orientation preferences, perhaps implicating inhibitory interneurons in the formation of orientation columns or in the fine-tuning of responses. (Hata *et al.*, 1988). In summary, while there is evidence that both inhibitory and excitatory inputs to cortex play a role in orientation selectivity, the existing evidence for the mechanisms of orientation selectivity is consistent with the assumptions that are, broadly speaking, made by the self-organizing models.

Implications of the Self-Organization Research

Linsker's work is intriguing because cells with spatial opponency and orientation selectivity, properties that have been thought to be of tremendous computational significance, seem to arise completely automatically in a situation where their functional utility has no bearing on their development. Given this observation, several points immediately come to mind. First, Linsker's work seems to suggest that orientation selective cells could arise in any network that has sufficiently deep correlation functions in a layer of spatial-opponent cells. Because he trained his network on completely unstructured input, there is nothing inherently "visual" about his result. Based on his results, one might expect to find cells with oriented receptive-fields in all sensory systems, because sensory systems have roughly similar initial structure to the networks Linsker used: they have a topographic input layer, project to the thalamus, and then project to a primary sensory area. Interestingly, orientation selective neurons analogous to those in V1 have been reported in the somatosensory cortex, supporting this (Hyvarinen and Poranen, 1978, cited in Sur,Garraghty, and Roe, 1988).

Relevant to this is a remarkable study by Sur, Garraghty, and Roe that "rerouted" the retinal projections of ferret visual systems at birth to the ferrets' medial geniculate nucleus, the auditory portion of the thalamus (Sur,Garraghty, and Roe, 1988). This was accomplished by ablating V1, V2, and the superior colliculus of the ferrets' visual systems as well as the inferior colliculus (the area that the MGN gets most of its projections from). Doing this causes the retina to project to the MGN and hence to auditory cortex. They found that the MGN developed visually responsive cells that were retinotopically organized, and that some of cells had spatially opponent receptive fields. The primary auditory cortex in such ferrets (whose connections from the MGN were not altered by the operation) also developed visually responsive cells, about twenty percent of which were orientation selective. It may have been the case that only those retinal cells that had not already established stable connections with the LGN were available to postoperatively project to the MGN, and thus there may have been comparatively few projections from the retina to the MGN, which might account for the low percentage of orientation selective cells. In any case, this study corroborates one point that Linsker's work suggests: namely, that orientation selectivity can arise in any layer of cells that receives input from a layer of spatial-opponent cells.

A final point that should be mentioned is that Linsker's network suggests that mammalian visual systems may have evolved several stages of center-surround cells in order to deepen the minima of the correlation functions of the layer preceding V1 such that it could develop orientation selective cells. The presence of center-surround cells in the LGN when there are spatial opponent cells in the retina does not currently have an explanation. Although the four layers of the network in his paper were contrived to be directly analogous to the levels of the early visual system, Linsker's analysis gives a reason why the redundant morphology may actually serve a functional role.

In conclusion, there are a variety of reasons to think that self-organization is one way that the brain's structure is achieved. We have considered the phenomenon of orientation selectivity in mammalian visual cortex. Linsker's self-organizing neural network develops orientation selective cells automatically and without any environmental cues. Although there are some inconsistencies with his results and experimental work on the brain, that his network develops a computationally useful property completely on its own is remarkable. More biologically realistic models that achieve similar results, such as those by Ken Miller, make a strong case for the presence of similar principles of organization in the brain, and there is reason to believe that these principles may be behind other properties of neurons in the visual system. Regardless of whether orientation selectivity arises in the brain through mechanisms analogous to those described here, one of the most important insights to come out of the self-organization models is the importance of the covariance function of activities across a layer of cells whenever any type of Hebbian learning occurs in a system. This principle alone has great potential for helping to unearth the origins of organizaiton in the brain. Indeed, if self-organization is a viable means of development, as it seems to be, then it is probably used with frequency in the brain, and self-organizing models will be of great use in identifying this.

REFERENCES

Blakemore, C. and G.F. Cooper. (1970). Development of the brain depends on the visual environment. *Nature*. **228**: 477-478.

Brown, T.H. and Chattarji, S. (1995). Hebbian synaptic plasticity. In *The Handbook of Brain Theory and Neural Networks*.. Arbib, ed. (Cambridge: MIT Press).

Cruetzfeldt, O.D., Kuhnt, U. and Benevento, L.A. (1974). An intracellular analysis of visual cortical neurons to moving stimuli. *Experimental Brain Research* **21**: 251-274.

Ferster, D. (1986). Orientation selectivity of synaptic potentials in neurons of cat primary visual cortex. *Journal of Neuroscience* **6**: 1284-1301.

Ferster, D. (1987). Origin of orientation-selective EPSPs in simple cells of cat visual cortex. *Journal of Neuroscience* **7**: 1780-1791.

Ferster, D. (1988). Spatially opponent excitation and inhibition in simple cells of the cat visual cortex. *Journal of Neuroscience* **8**: 1172-1180.

Ganz, L. and Felder, R. (1984). Mechanism of directional selectivity in simple neurons of the cat's visual cortex analyzed with stationary flash sequences. *Journal of Neurophysiology* **51**: 294-324.

Guillery, R.W. (1972). Binocular competition in the control of geniculate cell growth. *Journal of Comparative Neurology* **146**:407-420.

Hata, Y. et al. (1988). Inhibition contributes to orientation selectivity in visual cortex of cat. *Nature* **335**: 815-817.

Hertz, J., Krogh, A., and Palmer, R.G. (1991). *Introduction to the Theory of Neural Computation. (*Redwood City: Addison-Wesley Publishing Company).

Hirsh, H.V.B. and Spinelli, D.N. (1970). Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats. *Science* **168**: 869-871.

Hubel, D.H. and Wiesel, T.N. (1963). Shape and arrangement of columns in cat's striate cortex. *Journal of Physiology*. **165**: 559-568.

Hubel, D.H. and Wiesel, T.N. (1974). Sequence regularity and geometry or orientation columns in the monkey striate cortex. *J ournal of Comparative Neurology*. **195**: 267-294.

Kammen, D.M. and Yuille, A.L. (1988). Spontaneous symmetry-breaking energy functions and the emergence of orientation selective cortical cells. *Biological Cybernetics* **59**: 23-31.

Koch, C. and Poggio, T. (1985). The synaptic veto mechanism: does it underlie direction and orientation selectivity in the visual cortex? In *Models of the Visual Cortex*. Rose and Dobson, eds. (New York: John Wiley and Sons Ltd).

Linsker, Ralph. (1986). From basic network principles to neural architecture: emergence of spatial-opponent cells. *Proceedings of the National Academy of Science USA*., 7508-7512.

Linsker, Ralph. (1986). From basic network principles to neural architecture: emergence of orientation-selective cells. *Proceedings of the National Academy of Science USA*., 8390-8394.

Linsker, Ralph. (1986). From basic network principles to neural architecture: emergence of orientation columns. *Proceedings of the National Academy of Science USA*., 8779-8783.

MacKay, D.J.C. and Miller, K.D. (1990). Analysis of Linsker's simulations of Hebbian rules. *Neural Computation* **2**: 173-187.

MacKay, D.J.C. and Miller, K.D. (1990). Analysis of Linsker's application of Hebbian rules to linear networks. *Network* **1**: 257-297.

von der Malsberg, Christoph. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Biological Cybernetics* **14**: 85-100.

von der Malsberg, Christoph. (1979). Development of ocularity domains and growth behavior of axon terminals. *Biological Cybernetics* **32**: 85-100.

von der Malsberg, C. and Cowan, J.D. (1982). Outline of a theory for the ontogenesis of iso-orientation domains in visual cortex. *Biological Cybernetics* **45**: 49-56.

von der Malsberg, Christoph. (1990). Network self-organization. In *An Introduction to Neural and Electronic Networks* . Zornetzer, Davis, and Lau, eds. (San Diego: Academic Press Inc.), pp. 421-432.

Mason, C. and Kandel, E.R. (1991). Central visual pathways. In *Principles of Neural Science.* Kandel, Schwartz, and Jessell, eds. (Norwalk: Appleton and Lange), 420-439.

Miller, K.D. (1989). Orientation-selective cells can emerge from a hebbian mechanism through interactions between on- and off-center inputs. *Society for Neuroscience Abstracts* **15**: 794.

Miller, K.D. (1990). Correlation-Based Models of Neural Development. In *Neuroscience and Connectionist Theory*. Gluck and Rumelhart, eds. (Hillsdale, NJ: Lawrence Erlbaum Associates), pp. 267-353.

Miller, K.D. (1992). Models of Actvity-Dependent Neural Development. *The Neurosciences* **4**: 61-73.

Miller, K.D. (1989). Development of Orientation Columns Via Competition Between ON- and OFF-center inputs.*NeuroReport* **3**: 73-76.

Miller, K.D. (1994). A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity-dependent competition between ON- and OFF-center inputs. *Journal of Neuroscience* **14**: 409-441.

Movshon, J.A. and Van Sluyters, R.C. (1981). Visual neural development. *Annual Review of Psychology*. **32**: 477-522.

Nelson, S., Toth, L., Sheth, B., and Sur, M. (1994). Orientation selectivity of cortical neurons during intracellular blockade of inhibition. *Science* **265**: 774-777.

Roe, A.W., Pallas, S.L., Hahm, J., and Sur, M. (1990). A map of visual space induced in primary auditory cortex. *Science* **250**: 818-820.

Roe, A.W., Pallas, S.L., Kwon, Y.H., and Sur, M. (1992). Visual projections routed to the auditory pathway in ferrets: receptive fields of visual neurons in primary auditory cortex. *The Journal of Neuroscience* **12**: 3651-3664.

Rose, D. (1995). A portrait of the brain. In *The Artful Eye*. Gregory, Harris, Heard, and Rose, eds. (New York: Oxford University Press), pp.28-51.

Sillito, A.M. (1979). Inhibitory mechanisms influencing complex cell orientation selectivity and their modification at high resting discharge levels. *Journal of Physiology* **289**: 33-53.

Sillito, A.M. (1980). A re-evaluation of the mechanism underlying simple cell orientation selectivity. *Brain Research* **194**: 517-520.

Srinivasan, M.V., Zhang, S.W., and Rolfe, B. (1993). Is pattern vision in insects mediated by "cortical" processing?. *Nature* **362**: 539-540.

Sur, M., Garraghty, P.E., and Roe, A. (1988). Experimentally induced visual projections into auditory thalamus and cortex. *Science* **242**: 1437-1441.

Worgotter, F., Niebur, E., and Koch, C. (1992). Generation of direction selectivity by isotropic intracortical connections. *Neural Computation* **4**: 332-340.

Yuille, A.L., Kammen, D.M., and Cohen, D.S. (1989). Quadrature and the Development of Orientation Selective Cortical Cells by Hebb rules. *Biological Cybernetics* **61**: 183-194.

1 As I have mentioned and explained, Linsker notes that the Mexican-hat minima depth can also be increased by lowering the average sum $\sum_j w_{ij}$ of all the weights of the afferent connections to a unit i.

2 Recall that the development of layer C depended primarily on the constants m and p. These are contained in the parameter g which was varied in these simulations.

3 Clearly, strict feedforward networks do not exist in the brain. (Indeed, only 10 percent of the inputs to the LGN come from the retina!) Linsker's work suggests, however, that many of the receptive field properties of cells may arise from feedforward interactions.

Josh McDermott '98 (jmcderm@fas.harvard.edu) is a Special Concentrator in Cognitive Science living in Leverett House. He has broad interests in vision, computational and cognitive neuroscience, and the neural basis of and philosophical issues surrounding consciousness.