O

One of the most intriguing problems encountered in neurophysiology has been the elucidation of the origin of feature-sensitive neurons and, in particular, their spatial order, as observed in numerous parts of the CNS (early theoretical works on feature-sensitive cells are reported in Refs 1–4). The emergence of elementary spatial order in sets of simulated feature-sensitive cells was first demonstrated by von der Maalburg in 1973 (Ref. 5, see also Refs 6–9). On the basis of the approaches of von der Maalburg and others, Kohonen10–12 was able to define some of the most fundamental conditions of self-organization and to express them in the form of an effective algorithmic scheme. This algorithm has emerged as being very powerful in producing many types of, even abstract, ordered ‘maps’ based on various sets of input data. More recent work has looked for the physiological counterparts that are linked to this principle13–14.

This algorithm has emerged as being very powerful in producing many types of, even abstract, ordered ‘maps’ based on various sets of input data. More recent work has looked for the physiological counterparts that are linked to this principle13–14.

Three types of neuronal organization can be called ‘brain maps’: sets of feature-sensitive cells, ordered projections between neuronal layers and ordered maps of abstract features. The latter are most intriguing as they reflect the central properties of an organism’s experiences and environment. It is proposed that such feature maps are learned in a process that involves parallel input to neurons in a brain area and adaption of neurons in the neighborhood of the cells that respond most strongly to this input. This article presents a new mathematical formulation for such adaptation and relates it to physiological functions.

Trends Neurosci. (1999) 22, 135–139

Where the abstract feature maps of the brain might come from

Teuvo Kohonen and Riitta Hari

Three types of neuronal organization can be called ‘brain maps’: sets of feature-sensitive cells, ordered projections between neuronal layers and ordered maps of abstract features. The latter are most intriguing as they reflect the central properties of an organism’s experiences and environment. It is proposed that such feature maps are learned in a process that involves parallel input to neurons in a brain area and adaption of neurons in the neighborhood of the cells that respond most strongly to this input. This article presents a new mathematical formulation for such adaptation and relates it to physiological functions.
Although the SOM principle is well known among brain theorists, many experimental neuroscientists seem to be ignorant of the specific requirements of this self-organizing process and the implications of the theoretical results to their work. This article, therefore, aims to provide a simple introduction to SOMs, which are useful for any scientist wondering whether a certain brain map could have emerged according to the SOM principle. The basic requirements for self-organization in an abstract set of elements and in a mathematical model will be described and two SOM simulations will be presented together with a final discussion of the possible physiological implementations of the SOM principle in real brain structures.

Requirements for self-organization

The self-organizing process can be realized in any set of elements, illustrated schematically in Fig. 1, where only a few basic operational conditions are assumed. For simplicity, let the elements (for example, single neurons or groups of closely cooperating neurons) form a regular planar array and let each element represent a set of numerical values, M (a model). These values can correspond to some parameters of the neuronal system and in contemporary computational neuroscience it is customary to relate these to synaptic efficacies. It is also assumed that each model is modified by the messages the element receives.

Let us assume that some mechanism exists by which an incoming message, X (a set of parallel signal values), can be compared with all models, M. In brain theory it is customary to refer to ‘competition’ between the elements when they are stimulated by common input, and the element whose parameters are closest to this input is activated the most. This element is called the ‘winner’ if it succeeds, while remaining active itself in suppressing the activity in the neighboring neurons by, for example, lateral inhibition. The winner model is denoted by M. Another requirement for self-organization is that the models will be modified only in the local vicinity of the winner(s) and that all the modified models will then resemble the prevailing message more accurately than before.

When the models in the neighborhood of the winner start simultaneously to resemble the prevailing message (X) more accurately, they also tend to become more similar mutually, that is, the differences between all the models in the neighborhood of M are smoothed. Different messages at different times affect separate parts of the set of models, and thus, after many learning steps, the models (M) start to acquire values that relate to each other smoothly over the whole array, in the same way as the original messages (X) in the ‘signal space’ do in such a situation, maps related topologically to the sensory occurrences start to emerge. These three subprocesses – the broadcasting of the input, the winner selection, and the adaptation of the models in the spatial neighborhood of the winner – seem to be sufficient, in general, to define a self-organization process that then results in the emergence of the topographically organized maps.

Simulations with idealized networks

Self-organizing processes can be implemented using many mathematical rules. The particular rules selected for this article constitute a compromise between the effectiveness of self-organization and biological realization. The following equations, which are different from the old SOM algorithms, have been introduced in order to relate the adaptation processes to physiological facts.

A simple representation of a message (X) is the data vector, \( X = (X_1, X_2, \ldots, X_i) \), that is, a list of numbers that represent the signal values. The model \( M \) is similarly represented by the ‘model vector’ \( m = (m_1, m_2, \ldots, m_j) \) of the same dimensionality as \( X \). Let the \( m \) be normalized in length such that the sum of squares of the elements is the same for each vector. Then the abbreviated notation, \( M \times X \) for vectors \( M \) and \( X \), can also change the models in the neighborhood of the winner so that the correlation between them and the input message increases, while the normalized lengths of the \( m \) are preserved approximately.

Let the neighborhood of the winner be described by the neighborhood function \( h_j \) which reaches its maximum for the winner (when \( i = c \)). The value of \( h_j \) decreases as the distance of neuron \( i \) from the winner \( c \) in the array of neurons increases. \( h_j(0) \) can also change with time. In simulations, \( h_j(0) \) can be given different mathematical forms.

Equation 2 shows how the modified values of the ‘winning’ models, \( m(t+1) \), are assumed to depend on the messages they receive: the larger the difference between the message at time \( t \) and the (weighted) model, \( m(t) \), the larger is the

$$
M_{c}(t + 1) = \max \{ |M_{i}(t) - x(t)| \} \quad \text{(1)}
$$

A message \( x(t) \) at time \( t \) will modify the values of models \( m(t) \) to new values \( m(t + 1) \) one time step later. The SOM algorithm applied in this article modifies the models in the neighborhood of the winner so that the correlation between them and the input message increases, while the normalized lengths of the \( m \) are preserved approximately.

Equation 2 shows how the modified values of the ‘winning’ models, \( m(t+1) \), are assumed to depend on the messages they receive: the larger the difference between the message at time \( t \) and the (weighted) model, \( m(t) \), the larger is the
change towards values of the message; however, owing to \( h \), the changes occur in the vicinity of the winner, namely:

\[
m(t + 1) = m(t) + h(t)(x(t) - m(t)) - x(t)(m(t) - x(t))
\]

(2)

Owing to the factor \( m(t)x(t) \), this particular updating law keeps the lengths of the \( m \) at about a constant value\(^1\).

This modeling approach also applies to the case where the signals \( x \) and the synaptic efficacies \( m \) attain all-or-none values. In a large population of neurons their correlation \( m \times x \) then represents the coincidence of active \( x \) values with the existing \( m \) connections.

The SOM algorithms have been used successfully to simulate various brain-like feature maps, such as tonotopic maps\(^2\), maps of the perceptual space\(^3\), color maps\(^4\), and many others. The following two computational examples illustrate what other types of maps may emerge in idealized neural networks. For details of these simulations see Refs 12,20,21.

Simulation 1: phonemotopic maps
To produce phonemotopic or phoneme maps, Kohonen\(^5\) analyzed continuous natural Finnish speech and formed the input vectors \( x(t) \) for Eqs 1 and 2 from the signal power at 15 frequency bands sampled regularly (every 20 ms). The original simulations were based on slightly different algorithms but the new experiment, illustrated in Fig. 2, has been carried out using the new Eqs 1 and 2. After a few tens of thousands of spectral samples had been presented to the network, the values \( m \) began to represent models of different phonemes in an orderly fashion. The circles depict neuronal units that form a regular grid and the symbols denote the phoneme (or, more accurately, the short-term phonemic spectrum) to which each unit is best tuned.

So far, pure phoneme maps have not been found in the human brain. However, indirect evidence suggests that such ordered representations exist. For example, phonemic representations of human children are modified strongly during the first two years by the native language\(^6\), which would agree with the result obtained from the simulation that the map is organized according to the sensory input it receives. Moreover, magnetoencephalographic evoked-response studies\(^7\) indicate that the auditory cortex reacts differentially to the coincidence of active \( x \) values with the existing \( m \) connections to the network, the values \( m \) began to represent models of different phonemes in an orderly fashion. The circles depict neuronal units that form a regular grid and the symbols denote the phoneme (or, more accurately, the short-term phonemic spectrum) to which each unit is best tuned.

Simulation 2: word-category maps
The purpose of this example, adapted from Ref. 20, is to demonstrate that the structure of a language on a low level, already in very narrow word contexts, reflects the semantic content of the words. Therefore, it becomes possible for even a simple artificial system, such as the SOM, to learn to identify semantic aspects of words from texts in an unsupervised way.

The words in written text are symbolic items that were converted into numerical input vectors \( x \) in the following way: each word in the text was replaced by a random vector \( x \) that consisted of seven random-number components; \( x \) is the word position in the text. These random vectors were different, but unique for each word in the vocabulary. The purpose of this randomness is to eliminate the effect of the word appearance on the results. Later simulations, with very large vocabularies, have used even 90-component vectors\(^8\). To take into account the co-occurrences of neighboring words, the input to the SOM consisted of a pair of code vectors of adjacent words of the type:

\[
x = (x_{i-1}, t)
\]

(3)

The SOM was then computed using the \( x \) vectors picked up from the whole text. When the model-vector values became stationary, the SOM models were labeled by those words whose \( x \) vectors best resembled the respective computed models. In this way the 'semantic map' of Fig. 3 was formed by using three-word phrases constructed of 30 words\(^9\). One should note the automatic segregation of the words into word classes and the fine structure that distinguishes, for example, the animate objects from the inanimate ones.

Later, Honkela et al.\(^{10}\) constructed similar but more detailed maps for large collections of natural text, Grimm brothers’ fairytales (English translation), by using triplets of words in the context. This same principle...
has also materialized in large document maps\textsuperscript{31} in which a second-level SOM organizes documents on the basis of their word-category histograms. The largest SOM so far realized has consisted of over 100,000 models (‘neurons’) and organized over one million documents.

Conclusions drawn from category-specific recognition deficits in patients\textsuperscript{26} and from recent functional imaging studies\textsuperscript{30,31} suggest that the brain contains selective and local representations of concrete and abstract concepts, categories of objects (animals, tools, etc.) and even contains distinct representations for animate versus inanimate objects, respectively. The word-category maps imply that the context of entities can be the driving force for the organization of their representations. This context need not be stated verbally as above, but can be formed directly from sensory experiences.

**Implementation of SOMs in the real brain**

It is, of course, unrealistic to expect that maps as clearly organized as those observed in various ideal simulation studies would emerge in the cerebral cortex. Preprocessing in neural realms is already much more complex than that used in simulations. However, the mammalian brain is known to support several feature maps, either orderly representations of the receptor surfaces on skin, retina or cochlea, or purely abstract maps, which are discussed in this article. The acoustic maps are assumed to be associated with sounds to which the organism is most frequently exposed to. Tonotopic maps, which are known to exist both in auditory pathways and in several cortical auditory fields\textsuperscript{32–33}, are often thought to result from ordered projections that originate from the basilar membrane, but they have also been produced successfully by the SOM algorithm directly\textsuperscript{34}.

Essentially every level of the nervous system exhibits plasticity under certain circumstances\textsuperscript{35} and, thus, feature maps can also be expected at all levels. Further- more, even the ‘hard-wired’ maps are known to depend on the sensory experience and they evidently are, at least to some extent, the result of self-organization.

Materialization of the SOM principle in the nervous system would first require a mechanism that distributes essentially the same, or strongly correlated, information to the neurons or neuron groups of a certain region. Spatial diameters of the maps are then determined by the longest distances reachable by common (or highly correlated) input in the neuronal layer. The SOM principle might work well in the thalamocortical system, where the diameter of the map would be determined primarily by the size of the thalamocortical axonal arbor, which reach 1–2 mm in length (maximally\textsuperscript{36–37}), and by the ramifications of the apical dendritic trees of the receiving cortical cells, which have diameters of a few millimeters\textsuperscript{38}. Thus, a typical cortical map that describes a single type of feature would be (maximally) 5 mm wide. Second, one needs a mechanism that ‘selects’ the ‘winner’ neurons, that is, the center around which adaptation will take place. Lateral interconnectivity with excitatory and inhibitory connections\textsuperscript{39–41} could play a central role in this selection, which results in enhanced discharge rates at a place where the original excitation was high, and suppression of activity elsewhere. Third, the restriction of learning to the neighborhood of the winner(s) might simply follow the clustering of triggering activity, but one can also assume that some type of chemical ‘learning factor’ that controls modification locally, but does not activate the neurons, is emanating from the active neurons\textsuperscript{42–43}.

**Hebbian learning versus SOM**

As explained above, the formation of SOM requires modification of synaptic strengths in the winners’ neighborhood. In computational neuroscience such changes are usually assumed to obey the law of Hebb\textsuperscript{44}, which suggests that the synaptic efficacy increases in proportion to the conjunction of presynaptic and postsynaptic activities of the neurons. Kohonen\textsuperscript{11–12} modified Hebb’s law because chemical retrograde messengers, such as nitric oxide, have been discovered that make it possible for a neuron to modify not only its own synapses but also those of nearby neurons\textsuperscript{45}.

Consider neuron $i$ that has output activity $y_i$ and presynaptic activity at a synapse $j$ equal to $x_j$. Hebb’s law for the change of the efficacy of synapse $j$ is traditionally written as:

$$m_{ij}(t+1) = m_{ij}(t) + a y_i x_j$$

(4)

where $a$ is a proportionality factor.

In theoretical examinations, the models of neuronal networks are often simplified by assuming that one processing unit represents, for example, one pyramidal neuron. However, for the physiological correlates of the SOM maps it is more expedient to select a neuronal group consisting of say a few hundred neurons of different types for the processing unit\textsuperscript{43}. The dense intracortical interconnections, which are spread to nearby neurons, and the learning rate in Hebb’s law can be regarded as being proportional to the mean activity level of the whole neuronal group.

Although some general evidence for the Hebb-type synaptic modifiability exists\textsuperscript{46–47}, the inaccuracy of available measurements does not justify the use of a more detailed mathematical law. The following expression must therefore only be regarded as a hypothesis of what effects might account for the adaptation law expressed by Eqn 2.

First, as $y_i$ and $x_j$ are non-negative, $m_{ij}$ according to Eqn 4, would grow without limits. This is unrealistic, and another necessary modification is, therefore, ‘active forgetting’, which is the increase of the synaptic efficacy by the presynaptic activation ($x_j$) that must be opposed by some sort of active corruption of the efficacy of the synapse.

If any active neuronal group denoted by the subscript $c$ controls the modifiability or metaplasticity\textsuperscript{48} of synapses in its neighborhood ($i$) by any means, $y_{ci}$ can be replaced in Eqn 4 by the ‘neighborhood function’, $h_c$.

Equation 4 can be replaced with Eqn 5 provided that (1) the net depolarization of the cell, owing to all presynaptic activations, is proportional to the correlation of inputs with the synaptic efficacies, that is, to $m \cdot x_j$; (2) the corruption of $m_{ij}$ is proportional to the net depolarization and further to the value of $m_i$ itself, and (3) the neighborhood function is taken into account:

$$m_{ij}(t+1) = m_{ij}(t) + h_c y_{ci} x_j - \beta [m_i x_j f(t)]$$

(5)

where $\beta$ is another proportionality factor. If the ‘active’ component also needs the inhibitory action from the neighborhood, then $\beta = h_c x_j$. Eqn 5, written in vector form, then becomes exactly the same as Eqn 2. These hypotheses, therefore, lead to the formulation...
of the SOM process described above and are thus consistent with the basic functional principle of SOM. This particular equation (Eqs 5) is new in computational neuroscience.

Predictions from the simulated feature maps

Theoretically, some sort of competition on common input between spatially neighboring cells is always necessary for the emergence of feature-specific cells. Similarly, the condition for the production of ordered feature maps is that the ‘winner’ cells force their spatial neighbors to learn the same input. Such conditions might be ubiquitous in the CNS.

Abstract maps

In the simulated feature maps, the area occupied by a feature value is proportional to some power of the probability of co-occurrences of the feature. Theoretically, the exponent can assume values between one third and one. The simulated phoneme map of Fig. 2 illustrates this ‘magnification’ phenomenon: for example, the most common Finnish phoneme /a/ is mapped into 1/8 of the total of 96 neurons. Cortical sensory representations are also dynamic and have a capacity for reorganization after lesion or altered experience15,16.

Imperfect maps

The simulated SOMs sometimes contain typical unintended features. For example, if self-organization starts from different random initial states, the orientation of the resulting maps may vary or form a mirror image. The ‘initial values’ of brain maps are most probably determined genetically, but reverse ordering is still possible. If the self-organization process is not perfect, artificial networks can be fragmented into two or several parts with opposite order.

Overlapping maps

The neurons of the model network might be organized to receive afferents from different modalities. If only one modality is used at a time, several independent maps can be superimposed on the same network. Similarly, if, for example, two types of input are used simultaneously, a common map is formed, with representations for both modalities15. The map is oriented according to the strongest features in either input and the representation is redundant: cells respond similarly to inputs from either modality15.

Concluding remarks

The abstract maps might be more common in the brain than is thought. One difficulty encountered in finding them, especially in humans by means of non-invasive functional brain imaging tools17,18, is their very small extent. Therefore, simulated SOMs and real brain maps have not been compared frequently. However, their similarities suggest that the same simple principle might underlie, at least in an abstract form, the emergence of feature maps in the living brain at scales of less than a few millimeters. It is not yet known whether larger-scale maps could be formed according to the same principle or perhaps by a multi-level organization. Much work, therefore, remains to be undertaken in order to find new feature maps and to understand the detailed mechanisms that underlie reorganization of cortical representations. Experimentalists who discover new feature-sensitive neurons are urged to probe the surrounding area in detail to find related detectors. It is hoped that predictions based on SOM simulations would stimulate neurophysiological research to explore further the neuronal implementation of the models, the synaptic plasticity laws, the spatial scales of the maps and, in general, factors that control map formation.

Selected references

5. von der Malsburg, C. (1973) Kybernetik 14, 83–100
34. Mishkin, D. (1949) Organization of Deliberate, John Wilier and Sons
38. Santhirathinai, J. (1975) Brain Res. 93, 475–496

Acknowledgments

The authors’ research is supported by the Academy of Finland. The authors thank their colleagues, especially Academician Olli V. Lounasmaa, for valuable comments on the manuscript.