Physica D 42 (1990) 257–264
North-Holland

AN ENERGY FUNCTION FOR SPECIALIZATION

W. BANZHAF[1] and H. HAKEN
Institut für Theoretische Physik und Synergetik, Universität Stuttgart, Pfaffenwaldring 57/IV, D-7000 Stuttgart 80, Fed. Rep. Germany

We present a model of unsupervised learning based on the minimization of an energy function. The minima of the energy function are related to the degree of specialization of a certain class of artificial neuronal cells – grandmother cells – in the neural network model proposed by Haken. The self-organizing properties of the system are demonstrated by feeding input into a network of such cells with originally randomized synaptic connections. The relation of this learning algorithm to other learning schemes, like e.g. Kohonen's feature maps, is outlined.

1. Introduction

In recent years dynamical systems have been used with considerable success for purposes of information processing. Especially in the field of pattern recognition they have proven to be useful. Processes modeled by differential or difference equations underlie all natural phenomena and are therefore candidates for successful implementations of natural information processing capabilities into computers of future generations.

In order to constrain the arbitrariness of dynamical processes, researchers are looking for dynamical laws which stand out in some sense, for instance those which are related to certain extremal or optimization principles. Thus, the derivation of a certain information processing dynamics from the maximization or (in physics) minimization of a particular scalar function increases its plausibility and gives a serious motivation to study this law in more detail. This may be one of the reasons for the recent success of the Hopfield model for associative memory [1].

Whereas the search for optimization principles has been successful in the case of a dissipative dynamics to recognize or classify patterns (see e.g. Haken [2]), the question of self-organized pattern learning has resisted such an approach for a long time. Again, a process observed in nature – adaptation of living organisms to their environment – provided a starting point to introduce different dynamical laws. Moreover, in the restricted case of "supervised learning" a particular scalar function, the error function, was a natural choice. In the more interesting case of unsupervised or self-organized learning, however, the formulation of an optimization principle is not so obvious. In the recent work of Linsker [3] we see one of the promising general approaches to this problem.

In the following, we shall propose another optimization principle for unsupervised learning, which may turn out to be equivalent if formulated sufficiently generally. For the moment, we shall restrict ourselves to a particular network architecture and study the consequences in that context in more detail. More specifically, we shall report here on recent progress made with the neural network architecture proposed in 1987 by Haken [2, 4]. In particular we use a local Hebb-like learning rule derived from what may be called a principle of cell specialization. This will be formulated in detail below.

To state the principle rather generally, a cell in the network competes with all the other cells to

[1]Now at: Central Research Laboratory, Mitsubishi Electric Corporation, 1-1, Tsukaguchi Honmachi 8-chome, Amagasaki, Hyogo, 661, Japan.

represent the patterns offered. The competition eventually settles when minimal overlap between patterns represented by different cells has been reached which accounts for a maximal specialization of the cells in the network. Thus the assumption is that a sort of "effectiveness" criterion is imposed on the cells due to the fact that it is costly to establish and supply any cell in a network. Though such a principle may not have a real justification in artificial systems, in natural systems at least it is reasonable. From this principle a dynamical law of connection modification is derived which will be demonstrated in the paragraphs to follow.

We only claim here that a principle of cell specialization can be applied to many neural network models and may lead to reasonable learning dynamics.

2. The network architecture

A few words are in order to give an overview of the system. The network consists of at least two layers of units (cf. fig. 1), the *input layer* (I) and the *processing layer* (II). Optional is a third layer (for output) or even more intermediate processing layers.

The input units, whose activity values q_i, $i = 1, \ldots, N$ vary continuously between -1 and $+1$,

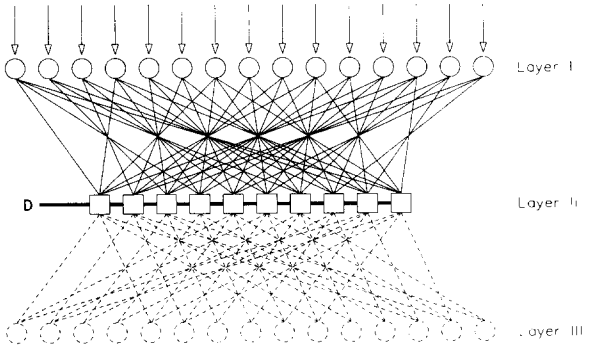

Fig. 1. Design of the overall system. Information flow is coming in through input cells q_i in layer I and is processed by layer II cells k with activity d_k. Output layer III is optional for generating patterns to implement associative recall.

communicate with the environment:

$$q_i \in [-1, +1].$$

The processing units (layer II) receive their inputs from these units via synaptic connections A_{ki}, the sum of which prepares the initial conditions of their internal activity dynamics $d_k(t)$, $k = 1, \ldots, K$:

$$d_k(0) = \sum_{i=1}^{N} A_{ki} q_i.$$

The internal units are connected so as to implement a winner-take-all network by coupling every unit to a global field D,

$$D(t) = \sum_{k=1}^{K} d_k^2(t),$$

and a competitive dynamics described by

$$\dot{d}_k(t) = d_k(t)\left[1 - 2D(t) + d_k^2(t)\right]. \tag{1}$$

After relaxing the dynamics, one cell (let us call it k') wins the competition ($d_{k'} = 1$). The dynamics is constructed such that this will be the cell that had the maximal absolute activity value from the beginning.

This particular dynamics can be found in other natural systems, e.g. lasers [5], and it is therefore a good candidate for an implementation of winner-take-all networks. Moreover, the dynamics is derivable from a scalar (energy) function $V(d_k)$ of the cell activities:

$$V(d_k) = -\tfrac{1}{2}D + \tfrac{1}{2}D^2 - \tfrac{1}{4}\sum_k d_k^4 \tag{2}$$

by applying the gradient

$$\dot{d}_k(t) = -\nabla_{d_k} V.$$

As was shown elsewhere [2, 4, 6], this network has pattern recognition abilities if the synaptic couplings are suitably determined. The network

and its dynamics were tested in detail in ref. [7] on a face recognition problem. A short review, however, will be given here to set the stage for the following discussion. For details, the interested reader may refer to refs. [2, 4, 6].

Suppose we have normalized prototype patterns v_l, $l = 1, \ldots, L$, that are to be recognized by the network. Then we use $K = L$ grandmother cells and either store pattern v_l in the respective synaptic filter A_k of the corresponding cell k (in the special case of orthogonal patterns) or we store the adjoint pattern v_l^+ in A_k (in the general case of non-orthogonal patterns). The adjoint vectors are defined as

$$v_l^+ = \sum_{l'} C_{l,l'} v_{l'} \tag{3}$$

with $C_{l,l'}$ being the inverse correlation matrix between patterns:

$$C_{l,l'} = \left(\sum_{i=1}^{N} v_{li} v_{l'i} \right)^{-1} . \tag{4}$$

The synaptic connections now act as filters decomposing arbitrary input patterns q which generally consist of superpositions of the known patterns v_l plus some noise n

$$q = \sum_l \alpha_l v_l + n .$$

They achieve this by translating the strength of any known pattern α_l into activities $d_k(0)$ of the corresponding grandmother cells k. Note that the noise n lies in a space orthogonal to all known patterns v_l. The highest activity $d_{k'}$, i.e. that of the cell k' responsible for the pattern with largest contribution α_l, is then amplified according to the network dynamics of eq. (1). If the network is equipped with the optional output layer connected to grandmother cells by another filter B_{ki} the patterns v_l can be stored in these filters. In this way, the network allows for associative recall observed at the output layer.

3. A specialization parameter

The purpose of this section is to identify a parameter which allows us to measure the degree of specialization of a certain cell k. Learning will then be derived from a maximization principle for such parameters.

Given M patterns to be learned by K cells with activities d_k, $k = 1, \ldots, K$, every cell may try to specialize on at least one of the patterns. The term specialization means – in the context of this network – the ability of a cell k to win a competition against the other $K - 1$ cells. On the basis of a pattern q this is provided for cell k by the following two criteria:

(a) an advantageous initial preparation $d_k(0) = A_k \cdot q$;

(b) a good position during the competition dynamics, as measured by

$$m_k = \langle d_k(t) \rangle_\tau \tag{5}$$

over transient times τ. Consequently, we claim that

$$s_k|_q = m_k d_k(0)|_q \tag{6}$$

measures the specialization of cell k on pattern q.

A general measure for the specialization state of a cell k (independent of the pattern q) is given by the ensemble average

$$\langle s_k \rangle_q ,$$

whereas

$$s = \left\langle \frac{1}{K} \sum_k s_k \right\rangle_q$$

averages the specialization over all the K cells.

Measuring the specialization by this method is only one way of doing it in the context of this network. Alternative measures can be found and we do not state that there exists a unique method. We want to emphasize, however, that a learning rule based on a specialization measure is very

effective and is probably realized even in natural systems.

4. An energy functional and the learning dynamics

We now consider how to maximize the specialization of cells under the constraint that the length of vectors A_k should be equal to 1 or at least tend to 1. This constraint is necessary in order to implement a competitive learning rule which gives any cell equal opportunities. The maximization may be achieved by defining a specialization functional and deriving the learning dynamics as the gradient descent from it. We propose the following scalar functional:

$$E(\boldsymbol{A},\boldsymbol{q}) = -\tfrac{1}{2}\sum_k s_k^2 l_k^2 = -\tfrac{1}{2}\sum_k m_k^2 d_k^2(0)\left(1-\tfrac{1}{2}\|\boldsymbol{A}_k\|^2\right), \qquad (7)$$

where l_k^2 was introduced for the dynamics to tend to normalized A_k vectors.

Note that the functional depends on $\boldsymbol{q}$ and thus results in different "landscapes" E for different $\boldsymbol{q}$. The idea behind this is that a changing environment is able to modify the learning dynamics, at least its detailed trajectory. The system becomes history-dependent and – at the same time – adaptive.

The learning dynamics derives from this specialization functional as the gradient

$$\dot{A}_{ki} = -\frac{\partial E(\boldsymbol{A},\boldsymbol{q})}{\partial A_{ki}} = m_k s_k\left[l_k^2 q_i - \tfrac{1}{2} d_k(0) A_{ki}\right]. \qquad (8)$$

Under a given input $\boldsymbol{q}$, the dynamics (8) for the synaptic connections A_{ki} tends to minimize E. One can easily verify that (8) tends to filters of equal length 1. Similar learning rules are studied in many models under the heading Hebbian rules, since the positive term proportional to q_i only needs local information about the cell's state as well as the forgetting (or stabilizing) term proportional to A_{ki}. The global factor $m_k s_k$ modulates the learning velocity of individual cells according to their specialization state.

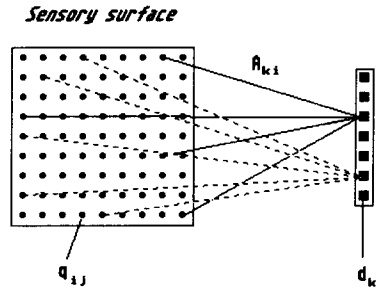

Fig. 2. The two-dimensional arrangement of cells in a sensory surface q_{ij} and a processing layer d_k.

As mentioned before, changing $\boldsymbol{q}$ will result in another dynamics $\dot{A}_{ki}$ due to changes in the energy landscape. Different cells will specialize on different patterns and the average energy $\langle E\rangle_q$ stabilizes only if a suitable adaption of cells to the probability density of inputs $P(\boldsymbol{q})$ is achieved. This feature is particularly useful if more patterns than cells are present, a case we shall study in our simulations below. In that case, the system organizes itself to classify the presented patterns into (best-match) classes.

5. Simulations

For simulation purposes we have chosen a typical classification situation (see fig. 2): An arrangement of two-dimensional sensory cells (input units) with $N = 100$ sensors q_{ij} are connected to the processing layer of $K = 20, 16, 8, 4, 2$ grandmother cells k by initially randomized connections. There was no local constraint and the filters of every cell k covered at the beginning the entire input space. Fig. 3a displays the initial state of the filters for $K = 20$ grandmother cells.

A pattern is provided by a high stimulation q_{ij} at site i, j, together with lower stimulations in its

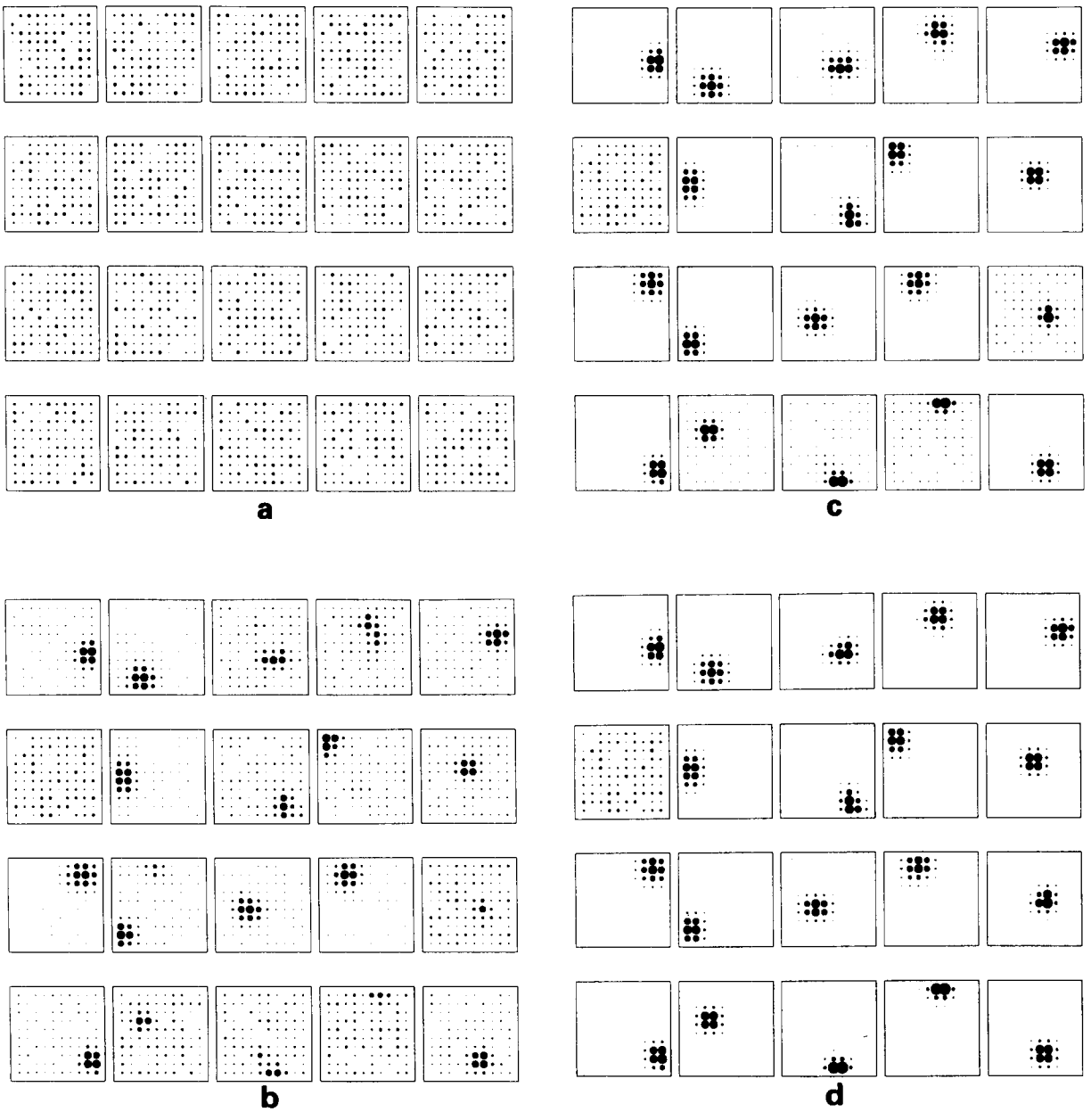

Fig. 3. Development of synaptic connections A_{ki} of grandmother cells $k = 1, \ldots, 20$. Synaptic strength proportional to the radius of black circles. (a) Before learning: All cells cover the whole surface. (b) After $r = 1000$ training steps. (c) After $r = 2000$ training steps. (d) After $r = 4000$ training steps. Cell 6 was not able to adapt to any pattern.

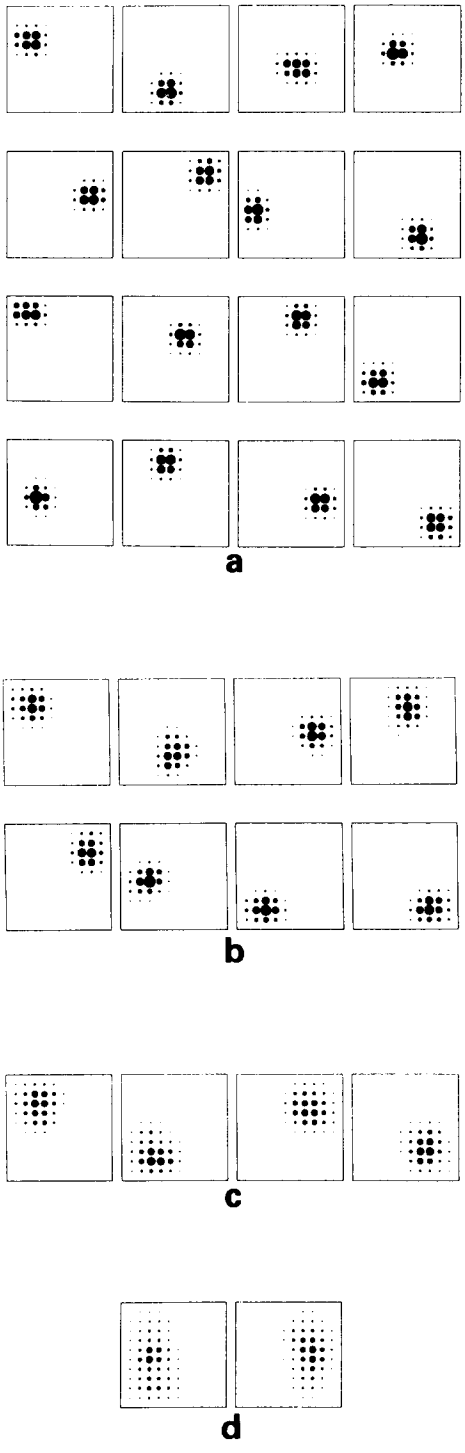

Fig. 4. Resulting connections in different runs with (a) $K = 16$, (b) $K = 8$, (c) $K = 4$, (d) $K = 2$ cells. The sensory surface is divided in nearly equal portions.

local neighborhood $q_{i'j'}$, $i' = i \pm l$, $j' = j \pm l$, $l = 1, 2, \ldots$. In other words, the patterns are chosen to be slightly correlated so as to allow a global ordering into neighborhoods. Here, the number of patterns, M, is 100 and we have many more patterns than cells for any of the cases studied.

Figs. 3b–3d show the synaptic filters (case $K = 20$) after $r = 1000, 2000, 4000$ training stimulations, respectively. The cells generally develop local sensitivities and respond after training to only a few patterns. One cell remained in its original state, a result which is not surprising since the gradient descent does not guarantee convergence to the global optimum. Quite evidently, after training the network, the cells are able to classify input patterns into different classes, the (maximal) number of which is given a priori by the number of grandmother cells participating in the competition.

Figs. 4a–4d show the results of runs with smaller numbers of grandmother cells. Clearly, the cells have to cover more and more sensory surface each.

A result of these simulations may be seen in the following.

(i) The system is able to classify patterns and thus to learn from noisy input data. Although no pattern of the kind seen in figs. 4a–4d was presented to the system, it nevertheless was able to develop a reasonable solution. The system reaches a stable state characterized by small fluctuations in the redistributed synaptic strength.

(ii) There is no built-in guarantee that the learning process will end up in the optimal solution, i.e. maximal specialization of all cells. Rather, the general result will be a nearly optimal solution.

(iii) A diffusion-like interaction in sensory cells is sufficient to generate local receptive fields, if a suitable competition between cells is implemented. The diffusion is able to correlate patterns which enables the cells in this case to generate neighborhood relations between patterns. The competition provides for learning of superpositions of presented patterns. After relaxation of the system, learning and competition may be turned off.

These sorts of patterns are by no means restricted to the simple point-like stimulations trained here. Those were merely chosen for purposes of demonstration. In another series of simulations we have shown what happens in cases $M = K$ and $M < K$ [8].

6. Discussion

This learning scheme has many similarities to Kohonen's learning algorithm and what he calls the formation of feature maps [9, 10]. Both learning algorithms could be termed non-equilibrium learning since the competitive systems are unrelaxed during learning. The adaptive abilities of both systems are comparable. We have hints on bound effects and on a shift in representation space towards regions with smaller probability density in our system, too.

The following differences from Kohonen's method are evident:

(i) The lateral inhibition between cells is uniform in our network. Neither time dependence of inhibition strength nor topological dependence of signals based on some notion of neighborhood in the network is introduced. The sharpening of signals during learning is due to an increase in specialization of cells; the topological mapping of signals is completely missing. Turning off competition results in a scattered map of input patterns, which is certainly useful for some applications.

(ii) The learning process automatically accelerates when specialization of cells proceeds. As a secondary effect of specialization and the forgetting term in eq. (8), the amount of redistributed synaptic strength decreases during training until it reaches minimal values if the optimal adaption of cells to the stationary probability distribution of inputs is reached.

(iii) The parameters which control the overall behavior are the three time constants: competitive activity dynamics, competitive learning dynamics and the training frequency.

In general, a multilayer system of cells is reasonable, as in the case of Linsker's network [11]. Accordingly, we should differentiate between a learning and a maturation state of the network, the latter without time-dependent connections and activity values in a layer. In this way, one layer after the other could process information and adapt to relevant signals in a self-organized manner. Our proposed learning dynamics leads to arbitrary synaptic strengths (with the constraint of being normalized for a cell), in contrast to Linker's connections, which saturate in extremal states. Carpenter and Grossberg [12] and Rumelhart's [13] competitive learning systems differ in the way they stabilize the network after learning.

Since the plasticity of the proposed algorithm is still present after the system has learned to classify input, and since it can readapt anew if the probability distribution of the input changes, experimental evidence [14, 15] concerning adaptive properties of living beings is at least not contradictory to the learning scheme presented here.

Acknowledgement

We wish to thank the "Stiftung Volkswagenwerk" for financial support.

References

[1] J.J. Hopfield, Proc. Natl. Acad. Sci. US 79 (1982) 2554.
[2] H. Haken, in: Computational Systems, Natural and Artificial, Proceedings of the Elmau International Symposion on Synergetics 1987, ed. H. Haken (Springer, Berlin, 1987).
[3] R. Linsker, IEEE Computer (March 1988) 105; presented at the Ninth Annual CNLS Conference on Emergent Computation, Los Alamos, 22–26 May 1989.
[4] H. Haken, Z. Phys. B 70 (1988) 121.
[5] H. Haken, Synergetics, An Introduction, 3rd Ed. (Springer, Berlin, 1983).
[6] H. Haken, in: Neural and Synergetic Computers, Proceedings of the Elmau International Symposion on Synergetics 1988, ed. H. Haken (Springer, Berlin, 1988).
[7] A. Fuchs and H. Haken, Biol. Cybern. 60 (1988) 17, 107.
[8] W. Banzhaf and H. Haken, Neural Networks, in press.

[9] T. Kohonen, Biol. Cybern. 43 (1982) 59.
[10] T. Kohonen, Selforganization and Associative Memory, 2nd. Ed. (Springer, Berlin, 1987).
[11] R. Linsker, Proc. Natl. Acad. Sci. US 83 (1986) 7508, 8390, 8779.
[12] G.A. Carpenter and S. Grossberg, Appl. Opt. 26 (1987) 4919.
[13] D.E. Rumelhart and D. Zipser, in: Parallel Distributed Processing, Vol. 1, eds. D.E. Rumelhart and J.L. McClelland (MIT Press, Cambridge, MA, 1986).
[14] M. Merzenich, presented at the Ninth Annual CNLS Conference on Emergent Computation, Los Alamos, 22–26 May 1989.
[15] W. Levy, Presented at the Ninth Annual CNLS Conference on Emergent Computation, Los Alamos, 22–26 May 1989.