Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 1695–1698, March 1986
Biophysics

Spin glass model of learning by selection

(Darwinism/categorization/Hebb synapse/ultrametricity/frustration)

GÉRARD TOULOUSE, STANISLAS DEHAENE, AND JEAN-PIERRE CHANGEUX

Unité de Neurobiologie Moléculaire and Laboratoire Associé au Centre National de la Recherche Scientifique, no. 270, Interactions Moléculaires et Cellulaires, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cédex 15, France

Contributed by Jean-Pierre Changeux, October 31, 1985

ABSTRACT **A model of learning by selection is described at the level of neuronal networks. It is formally related to statistical mechanics with the aim to describe memory storage during development and in the adult. Networks with symmetric interactions have been shown to function as content-addressable memories, but the present approach differs from previous instructive models. Four biologically relevant aspects are treated—initial state before learning, synaptic sign changes, hierarchical categorization of stored patterns, and synaptic learning rule. Several of the hypotheses are tested numerically. Starting from the limit case of random connections (spin glass), selection is viewed as pruning of a complex tree of states generated with maximal parsimony of genetic information.**

Aside from the inneist, or preformist, point of view, according to which experience does not cause any significant increase of order in an already highly structured brain organization, two main classes of learning theories have been proposed and discussed (for review see ref. 1). On the empiricist side, the initial state is considered as a *tabula rasa*, and the whole internal organization results from direct instructive prints by the environment. Alternatively, selectionist theories postulate that the increase of internal order associated with experience is indirect (2–8). The organism generates, spontaneously, variable patterns of connections (3) at the sensitive period of development, referred to as "transient redundancy" (6), or variable patterns of activity named prerepresentations (7, 8) in the adult. Interaction with the environment merely selects or selectively stabilizes the preexisting patterns of connections and/or firings that fit with the external input, a step named "resonance" (7, 8). As a correlate of learning, connections between neurons are eliminated and/or the number of accessible firing patterns is reduced.

Several attempts to model learning at the level of large ensembles or "assemblies" of interconnected neurons have been made in quantitative terms mostly with the help of statistical mechanics (9, 10). Their revival is largely due to the introduction by Hopfield (10) of the conceptual simplification that (*i*) if one restricts the interactions between neurons only to symmetric ones, this allows for the introduction of an energy function and, as a consequence, the dynamics of neuronal networks can be viewed as a downhill motion in an energy landscape and (*ii*) then, the reallowance for dissymmetric interactions does not discontinuously upset the picture.

On the other hand, such models still belonged to the empiricist mode of learning with the initial state taken as a flat energy landscape (*tabula rasa*) that becomes progressively structured and complex by direct instructions from the environment.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

The aim of this communication is to propose a model of learning by selection based on an advance in the statistical mechanics of disordered systems—namely, the theory of spin glasses (11–13). In contrast to the empiricist approach, the initial state is viewed as a complex energy landscape with an abundance of valleys typical of spin glasses with learning consisting of the progressive smoothening and gardening of this landscape. The paper also contains a biological critique of the standard instructive version of the Hopfield model, referred to here in short as the instructive model. The main proposals for a selectionist model of learning are outlined and preliminary numerical results are reported and discussed.

The Activity of Neuronal Networks Described by Statistical Mechanics

The all-or-none firing of a neuron is represented by a spin that can take two values: $S = +1$ (firing), $S = -1$ (rest). A pattern of activity, α, of a network of N neurons is represented by a spin configuration (S_i^α), $i = 1, \ldots, N$, that lies at one of the corners of a hypercube in N-dimensional configuration space. Two patterns of activity, α and β, may then be compared through their overlap, which is an index of proximity or matching in configuration space:

$$q^{\alpha\beta} = \frac{1}{N} \sum_{i=1}^{N} S_i^\alpha S_i^\beta. \qquad [1]$$

The neurons interact via binary synapses of synaptic strength, T_{ij}. With the assumption (5) of symmetric interactions $T_{ji} = T_{ij}$, an energy function can be written as follows:

$$E = -\frac{1}{2} \sum_{i \neq j} T_{ij} S_i S_j - \sum_i h_i S_i, \qquad [2]$$

where h_i is a local field acting on spin S_i and is often used to represent an external input (yielding an apparent shift of the firing threshold of a neuron). The neuron dynamics is such that, in the absence of probabilistic effects leading to random spontaneous activity, each spin tends to decrease its energy. A stable configuration is, therefore, a local minimum of the energy E. On the other hand, probabilistic effects can be described by introducing a finite temperature (14).

Synaptic modifications have been hitherto often expressed by the learning rule:

$$\Delta T_{ij} \sim \langle S_i S_j \rangle, \qquad [3]$$

where the brackets mean some time average. This expression, referred to as the "generalized Hebb rule," differs from the original Hebb rule (15), which may be written as

$$\Delta T_{ij} \sim \left\langle \left(\frac{S_i + 1}{2} \right) \left(\frac{S_j + 1}{2} \right) \right\rangle, \qquad [4]$$

and exclusively takes into account reinforcements of excitatory synapses. Rule 3 has attractive features—it is local and formally natural—but it also has undesirable ones—for instance, when neuron i makes inhibitory synapses with neuron

j, rule 3 would predict a modification of synaptic strength T_{ij} and eventually a reversal from inhibitory to excitatory, if none of the neurons is firing and the synapses are silent.

Instructive Models of Learning

Instructive models of learning (10) postulate that, in the initial state, the interactions between neurons are vanishingly small and the energy landscape is flat (*tabula rasa*). Storage into memory of an activity pattern α, where $S_i = \mu_i^\alpha$, results from the following synaptic modification,

$$\Delta T_{ij} = \frac{1}{N} \mu_i^\alpha \mu_j^\alpha, \qquad [5]$$

and the network is said to have learned M patterns, $\alpha = 1, \ldots, M$, when the interactions have been set to

$$T_{ij} = \frac{1}{N} \sum_{\alpha=1}^{M} \mu_i^\alpha \mu_j^\alpha, \qquad [6]$$

as a consequence of the successive prints of the M input patterns. With such interactions **6**, the network functions as a distributed, fault-tolerant, content-addressable memory. Starting from any input data, the network configuration rapidly converges toward a local minimum and recognizes the closest stored memory pattern (provided M is not too large and no confusion takes place) (10).

Assuming that the learned patterns are random and uncorrelated, Hopfield (10) has suggested that the maximal storage capacity is $M_c = \gamma N$ (with $\gamma < 1/2$, since each pattern corresponds to N bits of information, and the information is stored in the interactions, with $N^2/2$ of them) and further has shown that loss of recall occurred around $\gamma = 0.15$, an estimate that has been confirmed by subsequent analytical calculations (16, 17).

Such instructive mode of learning, if legitimate and useful for artificial intelligence, does not hold for the brain for the following reasons.

(*i*) As more and more patterns are stored, according to formula **6**, the synaptic patterns keep changing sign. As was already stressed by Hopfield (10), it is the signs of the interactions T_{ij} together with their absolute values that are responsible for the proper shaping of the energy landscape. Thus, storing a new memory amounts largely to reversing the signs of a particular set of synaptic strengths. Yet, no physiological evidence exists of synaptic sign reversal, such as a shift of postsynaptic response from excitatory to inhibitory, as a cellular correlate of learning (1).

(*ii*) Up to now, the ultimate organization of stored patterns in memory space has been viewed as a configuration-space-filling *jardin à la française*, with a regular distribution of the basins of attraction corresponding to the various stored patterns (18). A more hierarchical distribution less prone to confusions, with categorization properties and correlations between stored patterns, appears more appropriate for higher brain functions, even if it is wasteful of configuration space.

(*iii*) The hypothesis of an initial state with vanishing interactions does not take into account the existence of an already connected and functional neuronal network at the moment learning occurs.

Spin Glasses

Spin glasses, by definition, consist of networks of spins with symmetric random (positive and negative) interactions. The energy is simply given by

$$E = -\frac{1}{2} \sum_{i \neq j} T_{ij} S_i S_j. \qquad [7]$$

The mean field theory of spin glasses (valid for a fully connected network and a number of spins N large) is intricate but yields a simple physical picture for the energy landscape. The total number of local minima in configuration space is exponential in N. However, the dominant valleys (their importance is weighed by the Boltzmann factor, which favors low-energy valleys) have positive mutual overlaps. More precisely, to any spin state, time-reversal symmetry associates another state with all spins flipped and the same energy, so that the previous statement holds for each half of the valleys separately. In geometric terms, the dominant valleys of a spin glass lie within a cone, centered at the origin and of right angle in configuration space (one-half of the valleys within one sector of the cone, the other half within the opposite sector). Such a right-angle cone spans a very small fraction of configuration space, which is another way of stating that the dominant valleys are strongly intercorrelated.

Furthermore, the distribution of these valleys possesses an ultrametric structure (11)—i.e., a hierarchical organization of clusters within clusters—in configuration space. A similar ultrametric distribution occurs in taxonomy when species are classified, for instance, according to protein sequence homologies (19).

The spin glass energy landscape thus exhibits, spontaneously, a categorized organization. The appearance of ultrametricity for large heterogeneous assemblies is a remarkable feature, which may be partly understood by realizing that there are fewer bonds ($N^2/2$) than possible spin configurations (2^N), and thus that the energy states have to exhibit some form of correlation. Indeed, if ever random multiple-spin (ternary, quaternary, etc.) interactions are introduced, since they occur in larger combinatorial number than ordinary binary interactions, the energy landscape tends to become more and more rough, and the notion of hills, passes, and valleys eventually disappears (20).

Spin Glass Model of Learning by Selection

The proposal we make here is that the theory and formalism of spin glasses appear particularly adequate to model learning by selection. As discussed (7), selection has been postulated to operate during development on a variable connective organization (3, 6) or, in the adult, on variable patterns of activity named prerepresentations (7, 8). In both cases, a significant (though limited) randomness characterizes the initial state. This legitimizes the modeling of this "fringe" state by a network of N neurons with randomly connected excitatory and inhibitory synapses that would behave as a spin glass.

In brief, a spin glass has an energy landscape with: (*i*) an abundance of valleys and (*ii*) dominant valleys strongly intercorrelated (positive mutual overlaps) in a tree-like fashion.

Item (*i*) gives to this network the property of a "generator of internal diversity" (7, 21)—that is, each valley corresponds to a particular set of active neurons and plays the role of a prerepresentation. Item (*ii*) further indicates a spontaneous categorization of the prerepresentations.

Learning with very small synaptic changes is both advantageous and possible. It is advantageous because it tends to preserve ultrametricity—i.e., the spontaneous hierarchical categorization of the prerepresentations. It is possible because learning by selection involves the stabilization of preexisting valleys instead of creation of new ones. The foremost constraint is interdiction of synaptic sign reversals. Other proposals for the learning rule fall into two categories. The rule should remain local but avoid the inconsistencies mentioned above. In addition, a weighted factor is introduced and contributes to the selection of the input patterns that match the prerepresentations (resonance). This selective factor enters naturally as a time average, if one assumes that the synaptic changes occur during a relaxation time of the

configuration initiated by the input pattern. More coherent synaptic modifications will favor input patterns that match with preexisting valleys.

In summary, learning by selection may occur as follows: An input pattern sets an initial configuration that converges toward an attractor of the dynamics (bottom of a valley, i.e., a prerepresentation). The energy of this selected valley is lowered by synaptic modifications (particularly if the learning time is longer than the relaxation time), and its basin of attraction is shifted and enlarged at the expense of other valleys. Starting from a hierarchical distribution of valleys, the learning process can be viewed as pruning of a tree, analogous to that occurring in the course of phylogenesis. As a consequence the whole energy landscape evolves during the learning process, the already stored information influencing the prerepresentations available for the next learning event. Moreover, the constraints on the synaptic modifications give internal rigidity to the system. Not every external stimulus can equally be stored. Selection by the external stimuli among internal prerepresentations has its counterpart in selection by the internal network among external inputs. In a parallel assembly of such networks, one may further speculate that an input pattern will select its memory location, the place where it fits, if any.

A prerequisite of learning by selection is the existence of a nontrivial valley structure prior to the interaction with the outside world. There are only two ways whereby a neuronal network with symmetric binary connections can exhibit such a structure.

One way is via frustration (22). The frustration function $\Phi(c)$ of a closed loop (c) of interacting spins is the product of the interactions around the loop:

$$\Phi(c) = \prod_{(c)} T_{ij}. \qquad [8]$$

If $\Phi(c) > 0$, it is possible to find a spin configuration around the loop such that each bond is satisfied. If $\Phi(c) < 0$, this is not possible, and the spin configurations can be at best partially satisfactory. In this latter case, the loop is said to be frustrated. Frustration is a source of metastability and degeneracy. By definition, an unfrustrated network, where all loops are unfrustrated, has only two minima, related by time-reversal symmetry.

The other way to get multiplicity of valleys is via disconnection. A network, broken into p disconnected unfrustrated clusters, has 2^p minima.

The rich valley structure of a long-range, fully connected spin glass stems from frustration. It differs sharply from the valley structure of a set of disconnected clusters, although intermediate cases are conceivable. Indeed, if all the neurons are decoupled, the storage capacity is in some sense maximal, but all the useful properties of a distributed, associative memory are lost.

Any realistic neuronal network model for biological learning by selection should include both frustration and disconnection. Not only is the initial connectivity in central nervous systems far from maximal (see for instance, the anatomical evidence for columns) but the occurrence of synaptic elimination during development is well documented (6, 23). The constrained learning rules, introduced above, preserve frustration because they forbid sign reversals and allow for synaptic elimination.

Numerical Implementations and Results

Our model contains a set of hypotheses that are precise enough to be tested numerically, and we have begun a systematic investigation of their consequences. Some salient results, for small network sizes ranging from $N = 30$ to $N = 200$, are reported. A more elaborate discussion will be presented elsewhere. For the sake of clarity, we have studied separately the effects of each of our basic hypotheses and compared them with the instructive model.

The *Tabula Rasa* Withdrawn. In the initial state, the synapses are set with random signs and an average strength S. It is known from spin glass theory that partial learning of an arbitrary pattern can be obtained with synaptic increments of order $S/\sqrt{N}$ for a complete graph of N neurons. Keeping unchanged the form of the generalized Hebb rule, for the sake of comparison,

$$\Delta T_{ij} = \frac{\varepsilon S}{\sqrt{N}} \mu_i^\alpha \mu_j^\alpha, \qquad [9]$$

we have checked that retrieval quality [more precisely, in notations defined below, a normalized index $R = (q_a - q_b)/(1 - q_b)$] is a function of ε, independent of network size N. Furthermore, for $\varepsilon \gtrsim 2.5$, retrieval quality was found to be practically perfect.

As more and more patterns are stored, the strength of a given synapse undergoes a random walk, with steps of length $\varepsilon S/\sqrt{N}$ starting from the initial values $+S$ or $-S$. Whenever the strength of a synapse hits the value zero (an occurrence possible after learning $p \sim \sqrt{N}/\varepsilon$ patterns) it is prevented from changing sign. Two subsequent rules are conceivable, and both have been examined. Either the synapse is altogether eliminated, which is a strong form of the constraint, or its strength is temporarily blocked at zero until it eventually receives an increment of the correct sign, which obviously constitutes a weaker constraint.

In the case of the strong constraint, ruin theory (24) predicts that the fraction of surviving synapses will decay as $1/\sqrt{p}$, for p large (where p is the number of memorized patterns). With the weaker constraint, the fraction of nonvanishing synapses tends toward a constant.

We have defined a global learning index G and studied its variation as a function of p. This learning index is the difference between retrieval overlaps (a retrieval overlap is the overlap between an input pattern and its attractor) measured after learning and before learning, summed over all p patterns. Note that learning an additional pattern modifies the retrieval of previously stored patterns. Thus, the global index has to be completely recalculated after each learning event.

For p small, $G(p)$ is linear in p; for $\varepsilon \gtrsim 2.5$, the slope is the same for the *tabula rasa* condition or the non*tabula rasa* condition. Both curves are also asymptotically linear for p large (with smaller slope) and superimposed, showing a regime where the influence of the initial state has been lost. In the intermediate regime, the two curves differ. In addition, there is a difference between the cases with sign constraints (under weak or strong form) and the case without, which is clearly observed even on the smaller samples ($N = 30$).

Learning Strength and Selectivity. For comparison with previous studies, the values of ε chosen above were so large as to "burn a hole" in the energy landscape, for any input pattern. Such storage is clearly unselective. We have plotted the statistics of retrieval-overlap-after-learning q_a versus retrieval-overlap-before-learning q_b for various values of ε. Starting from $\varepsilon = 0$, for which the curve is obviously along the diagonal $q_a = q_b$, there is a range of values of ε for which marked fluctuations in retrieval quality are observed, before the hole burning regime sets in, with $q_a = 1$.

These results prove the existence of a diversity and an incipient selectivity. Note that, in these simulations as in earlier studies (10, 14, 16, 17), the learned configurations are the input patterns, because no relaxation effects are taken into account. The selectivity in the learning process, resulting from the existence of an initial structured energy landscape, will be enhanced by averaging over time. A learned configuration will then be intermediate between an input pattern

and its attractor, and the total amount of synaptic modification will be larger for a matching pattern than for a nonmatching one.

Alternatives to the "Generalized Hebb Rule"

Consistent with current models of regulation of synapse efficacy inspired from the allosteric properties of the acetylcholine receptor (25), one may express the change in the efficacy of a synapse between neurons i and j as a function of the activities of the other neurons k afferent on j, as

$$\Delta T_{ij} \sim \sum_k C_{ji}^k \langle S_i S_k \rangle, \qquad [10]$$

the coefficient C_{ji}^k being determined by chemical and geometrical factors, such as the relative positions of the synapses (i, j) and (k, j) on the dendrites of neuron j. Such a general expression points to the possibility that the printing process does not stabilize with exact precision a given imposed pattern but rather introduces a shift between an input and its trace. However, at this stage, we limit ourselves to a modification of the generalized Hebb rule **3**, which eliminates its most obvious flaws while keeping symmetric interactions. A simple way consists in replacing rule **5** by

$$\Delta T_{ij} = \frac{1}{4N} [3\ \mu_i^\alpha\ \mu_j^\alpha + (\mu_i^\alpha + \mu_j^\alpha) - 1]. \qquad [11]$$

Then, no synaptic modification occurs if $\mu_i = \mu_j = -1$, as desired. Consequently, every neuron will not be equally stabilized after the storage of one pattern and any stored pattern will have some labile spots.

As a first step, we have looked at the consequences of learning rule **11** in comparison with the generalized Hebb rule **5**, within the instructive model. The new rule has been found to affect the retrieval quality of the Hopfield model significantly. The reduction of the performances is comparable in magnitude to the effect of withdrawing the *tabula rasa* hypothesis (with generalized Hebb rule and without synaptic sign constraints) as described above.

Conclusions

Learning by selection is a generalization to the development of neuronal networks (3, 6) and to higher brain functions (4, 5, 8) of the selectionist (or Darwinist) mechanisms that have already been successfully applied to the evolution of species and antibody biosynthesis (2, 19). The spin glass model described here creates an additional bridge between statistical mechanics and theoretical biology and may offer original theoretical "tools" to quantitatively treat the neuronal bases of highly integrated brain processes. At this stage the model contains a severe restriction in scope due to its limitation to static memory patterns (time sequences and synchronicity effects are beyond present investigation).

One major neurobiological outcome of our model is the description of a memory with a hierarchical, ultrametric, structure which offers possibilities of "categorization" (11) on a rather simple basis—an initial "fringe" state of random synapses yielding a spin glass-like energy landscape and strong learning constraints at the storage level. This does not preclude, but rather complements, a hierarchical categorization at the encoding level (26) that originates, for instance, from a more innate organization of the sensory analyzers at the cortical level with multiple entries of the inputs into a layered architecture. In this framework, our study considers the less genetically determined layers that would then receive partially precategorized inputs.

In conclusion, this learning process can be epitomized as pruning (by selection) instead of packing (by instruction). It is too early yet to predict what will be the most fruitful implementation of this model, but two ideas appear profound and worth stressing. The first idea for the physicist is that selection, par excellence, is pruning of a tree and that the spin glass supplies the tree with parsimony of genetic information. The second idea for the biologist is that random synapses in a neuronal network cannot be equated with a *tabula rasa*.

We acknowledge valuable discussions with D. Amit, E. Bienenstock, J. J. Hopfield, H. Sompolinsky, and M. Virasoro. G.T. thanks the Aspen Center for Physics where part of this work was carried out. Computations were performed on the IBM 4341 of the Centre de Calcul de l'Ecole Normale Supérieure.

1. Marler, P. & Terrace, H., eds. (1984) *The Biology of Learning* (Springer, Berlin).
2. Jerne, N. (1967) in *The Neurosciences: A Study Program*, eds. Quarton, G., Melnechuk, T. & Schmitt, F. O. (The Rockefeller Univ. Press, New York), pp. 200–208.
3. Changeux, J. P., Courrège, P. & Danchin, A. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 2974–2978.
4. Edelman, G. (1978) *The Mindful Brain* (MIT Press, Cambridge, MA).
5. Finkel, L. & Edelman, G. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1291–1295.
6. Changeux, J. P. & Danchin, A. (1976) *Nature (London)* **264**, 705–712.
7. Changeux, J. P., Heidmann, T. & Patte, P. (1984) in *The Biology of Learning*, eds. Marler, P. & Terrace, H. (Springer, Berlin), pp. 115–133.
8. Heidmann, A., Heidmann, T. & Changeux, J. P. (1984) *C.R. Acad. Sci. Ser. 2*, **299**, 839–844.
9. Little, W. & Shaw, G. (1978) *Math. Biosci.* **39**, 281–290.
10. Hopfield, J. J. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 2554–2558.
11. Mézard, M., Parisi, G., Sourlas, N., Toulouse, G. & Virasoro, M. (1984) *Phys. Rev. Lett.* **52**, 1156–1159.
12. Toulouse, G. (1984) *Helv. Phys. Acta* **57**, 459–469.
13. Mézard, M. & Virasoro, M. (1985) *J. Phys. (Les Ulis, Fr.)* **46**, 1293–1307.
14. Peretto, P. (1984) *Biol. Cybern.* **50**, 51–62.
15. Hebb, D. (1949) *The Organization of Behavior* (Wiley, New York).
16. Amit, D. J., Gutfreund, H. & Sompolinsky, H. (1985) *Phys. Rev. A* **32**, 1007–1018.
17. Amit, D. J., Gutfreund, H. & Sompolinsky, H. (1985) *Phys. Rev. Lett.* **55**, 1530–1533.
18. Hopfield, J. J., Feinstein, D. I. & Palmer, R. G. (1983) *Nature (London)* **304**, 158–159.
19. Ninio, J. (1983) *Molecular Approaches to Evolution* (Princeton Univ. Press, Princeton, NJ).
20. Derrida, B. (1980) *Phys. Rev. Lett.* **45**, 79–82.
21. Stein, D. L. & Anderson, P. W. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 1751–1753.
22. Toulouse, G. (1977) *Commun. Phys.* **2**, 115–119.
23. Cowan, W., Fawcett, J., O'Leary, D. & Stanfield, B. (1984) *Science* **225**, 1258–1265.
24. Feller, W. (1957) *An Introduction to Probability Theory and Its Applications* (Wiley, New York).
25. Heidmann, T. & Changeux, J. P. (1982) *C.R. Acad. Sci. Ser. 2* **295**, 665–670.
26. Virasoro, M. (1985) in *Disordered Systems and Biological Organization*, eds. Bienenstock, E., Fogelman, F. & Weisbuch, G. (Springer, Berlin), in press.