BNL-61443

March 10, 1994

Self-Organization in a Simple Brain Model

Dimitris Stassinopoulos and Per Bak

Brookhaven National Laboratory

Department of Physics

Upton, NY 11973, USA

Preben Alstrøm

The Niels Bohr Institute

Department of Physics

Copenhagen 2100, Denmark

Abstract

Simulations on a simple model of the brain are presented. The model consists of a set of randomly connected neurons. Inputs and outputs are also connected randomly to a subset of neurons. For each input there is a set of output neurons which must fire in order to achieve success. A signal giving information as to whether or not the action was successful is fed back to the brain from the environment. The connections between firing neurons are strengthened or weakened according to whether or not the action was successful. The system learns, through a self-organization process, to react intelligently to input signals, i. e. it learns to quickly select the correct output for each input. If part of the network is damaged, the system relearns the correct response after a training period.

How does the brain work?

Two points of view as to where to look for the "secret" are often expressed:

1) The truth is in the detail. The brain consists of neurons. Once we understand the mechanism of the single neuron, we understand in principle everything. Thus, we must put emphasis on measuring the properties including the flow of chemicals, electrical potentials and pulses etc. at the synapses, axons etc. This traditional view has been very successful in science, most noticeably in particle physics where all matter has been reduced to a few quarks and gluons.

2) The truth is in the complexity. The brain has billions of neurons, each connected to thousands of other neurons. Once you have enough neurons, properly connected, intelligent behavior emerges by some magic. It has even been said that the brain must necessarily be so complicated that it can not possibly be understood by the brain. How then can we possibly generate a theory which deals with all these elements? Even to write down the map of the brain would require libraries of books.

Let us look into these two point of views. Let us compare with the way we would "understand" a man-made object, namely a computer.

First, following the strategy of looking into the details, we would take the computer apart and study its smallest parts. We would measure the characteristics of the transistors, that is, how the various currents and potentials depend on each other. We would have to understand the quantum mechanical properties of the materials, silicon etc, on which the transistor is based. Clearly, this will lead nowhere. Without any idea about the function that the transistors perform, no insight emerges. The computer engineer couldn't care less about how the transistor works - it is irrelevant for his purposes.

Second, although it is a popular view that a computer works because of its vast number of circuits, it is not so. The world's largest computers work the same way as the smallest pocket calculator. It simply has more storage, more processors, more input-output devices etc.

Thus, neither of the two points of views are correct for the computer, and most certainly they are not correct for the brain either. In order to understand the computer, one has to understand the principles by which the elements are put together. Whether the

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

elements are of one type or another, whether they are of electrical, optical, or mechanical nature is irrelevant as long as they perform the correct function, that is for instance to carry out a simple logical operation such as an "AND" or "OR" logical operation on two bits. One does not have to explain the complete system with its myriads of connection to understand the computer. The truth is not in the intricacies. A computer is basically a simple device, sending numbers or bits from one location to another, and performing trivial operations with pairs of those numbers.

The same, we argue, goes for the brain. The goal must be to understand the principles by which the neurons interact. This doesn't mean that the study of the hardware, such as the flow of Ca^{++} and Na^{+} ions at the synapses and axons, is irrelevant, in the same sense that the feasibility of constructing transistors is not irrelevant for the computer, but simply that this study can be decoupled from a general study of the mechanisms of the brain.

There is, however, one major conceptual difference between understanding the computer and understanding the brain. The computer was built by design. An engineer put together all the circuits etc. and made it work. In other words, with no engineer, we have no computer. However, there is no engineer around to connect all the synapses of the brain. One might imagine that the brain is ready and hard-wired at birth, with its connections formed by biological evolution and coded into the DNA. This does not make any sense. Evolution is efficient, but not that efficient. The amount of information contained in the DNA is vastly insufficient to specify all neural connections. The structure has to be *self-organized* rather than by design.

Thus, in order to understand the brain, we must understand the principles by which it organizes itself, presumably through its interaction with the environment. In order to be biologically feasible, those principles have to be simple and robust. In analogy with the computer, once those principles are understood there might be little qualitative difference between the smallest lobster brain and the human brain. If we are lucky, the difference is quantitative rather than qualitative. This "evolutionary" conjecture has not gained much acceptance; not because of lack of plausibility but because it failed to meet the immediate challenge it raises: to prove by demonstration the existence of such a simple and plausible

model.

Conventional attractor neural network models (For reviews see Amit[1] and Hertz et al[2]) work in two modes: a learning mode where the strengths of the neural connections are computed and a retrieving mode where the network recognizes input signals, i. e. provides the same pattern for several similar input patterns. More advanced models use complicated back-propagation algorithms which continuously update the connections by a computation not performed by the neural network itself. These models have been important in constructing technologies for pattern recognition, and emphasis has been on maximizing their capacity for learning, without regards to questions raised in realistic modelling of brain function. From its birth, a real brain is "on its own" in an environment that constantly changes with no outside agent to turn switches between learning mode and retrieval mode.

Recently, Alstrøm and Stassinopoulos[3] addressed some of these points in a new class of neural networks, denoted *adaptive performance networks*. The central idea is the introduction of a global evaluative feedback signal, a dynamic threshold, and a reinforcement rule with no need of further computation. Here, we address the question of how can we get intelligent behaviour not through engineering but through self-organization. We shall demonstrate that this type of network can be trained to react "intelligently" to external sensory signals.[4] In a fashion analogous to the behaviorist techniques used in the training of animals we introduce our system with a set of external signals each of which rewards a specific action. The system learns to recognize all signals and choose the corresponding rewarding action. "Learning" and "retrieving" are two aspects of the same dynamical process. It must be. Individual neurons don't know what is globally going on; they perform their thing automatically, without concern to whether they contribute to a learning or a retrieving task. Only an outside observer is able to identify what is going out as learning or retrieving by inspecting its behavior.

The goal of any scientific theory or model is to capture the essential elements of experiments or observations in nature. Here we wish to model intelligent behavior at its simplest. To be concrete, consider the situation in which a system provides food to a "monkey" if the correct button is pressed. Which button is correct depends on whether a red or a green light is on. This signal, which is shown to the monkey, is all the information

the monkey has in order to figure out which is the right button at every instant. The monkey learns the correct reaction after a "learning" period of trial and error. If the outside world changes, i.e., the "correct" buttons are switched, the monkey should be able to modify its behavior. The monkey is able to learn progressively more complicated patterns. The ability of a model to mimic this process of learning "intelligent" responses (leading to satisfaction) to outside signals is denoted "artificial intelligence."

We start by visualizing our model-brain in its embryonic state: a network of neurons with random connections. Little genetic information is needed to construct such random networks. Sensory signals are fed into the brain randomly. The neural output, such as stimulation of muscle fibers, is also sent randomly. The environment responds to the action directed by the brain's output by rewarding (or not rewarding) it. The result is fed back to the brain through a global signal, which could be a change in the level of a hormone or an increase in the blood-sugar content. There is no mechanism by which the information can be fed back selectively to the individual neurons.

In our picture the interplay with the environment is essential in organizing the brain's ability to explore and become more experienced; allowing it to react intelligently. In order to represent this, our model interacts with the "outer world" in three different ways (Fig. 1). There is i) an input signal giving information about the state of the outer world; ii) a resulting action by the system toward the environment; iii) a global feedback signal indicating whether this action was successful or not in accomplishing the goal. Our model is necessarily grossly oversimplified; its sole purpose is to demonstrate certain simple general principles.

We have studied two network topologies: a layered one and a random one. In the latter model, both *inputs*, *outputs*, and *internal connections* are completely random. N neurons are each connected randomly to C other neurons. The neurons can be either in a firing state, $n_i = 1$, or a non-firing state, $n_i = 0$. The input to the i'th neuron from other neurons is $h_i = \sum J_{i,j} n_j$, where the summation is over the C interacting neighbors. Initially, the j's are randomly chosen in the interval $0 < J < 1$. The neuron fires if the input exceeds a threshold T. The interactions with the environment are implemented as follows.

i) The sensory signal is represented by an additional contribution, h' to the input signal of a number of random neurons. These various branches can be thought of as different features of the input signal such as sound, shape, color, smell, position, size, etc. Different inputs are represented by different sets of random input neurons (see Fig. 1).[6] ii) The output signal is the firing state of a set of randomly selected output neurons. For each input signal, the action is considered successful if one or more specific but randomly selected neurons, belonging to the set of output neurons, are all firing. iii) If the action is successful, a positive reinforcing signal $r << 1$ is fed back to all firing neurons. If the action is unsuccessful a negative signal is fed back. The reinforcement modifies all connections between firing neurons $J_{i,j} \rightarrow J_{i,j} + [rJ_{i,j}(1 - J_{i,j}) + h]n'_i n_j$, where n'_i denotes the state of the i'th neuron at the next time step and h is a random noise between $-h_0$ and h_0. The inputs are normalized, $J_{i,j} \rightarrow J_{i,j} + J_{i,j}/\sum_j J_{i,j}$.

Thus, if the action is successfull, all connections between firing neurons are reinforced, whether or not they participated in delivering the correct output; if the action is unsuccessful, the connections between firing neurons is weakened.

In addition to the above input-output functions, the model has a global control mechanism for the activity (Alstrøm and Stassinopoulos) for the total number of firing output neurons, A. It is important that this be kept to a minimum. If A exceeds a value A_0 the threshold T is reduced, while if A is smaller than A_0 the threshold is increased, $T \rightarrow T + \delta \mathrm{sgn}(A - A_0)$. Thus, if there is no output, or the output is too low, the system is "thinking", that is its sensitivity is increased until an appropriate output is achieved. If the system is "confused", i. e. there is too much output, the sensitivity is lowered. Modulatory chemicals released into the brain help performing this function for the real brain, in addition to participating in the formation of the synapses, the J's discussed above.

At each time step the system is updated in parallel following the algorithm above. The performance P of the network is defined as the average success rate over 250 successive time steps. Figs. 2-6 show the results for a number of different tasks.

In the layered version, the neurons are arranged in rows, with each neuron firing to the three nearest neighbors in the next row. Inputs are random, but output neurons are those in the bottom row. At each time step the system is updated in parallel following the

algorithm above.

First, the "monkey" experiment defined above was simulated. A layered network with 256 neurons was studied, with $C = 3$ ($\eta_0 = 0.01, r = 0.1$). Two input signals, each with 16 random input neurons, were chosen. For each input, a pair of output cells was defined in the bottom row. The input signals were switched every 2000 time steps (or when complete success, meaning that the selected output neurons were active while all other neurons were not, has been achieved over 250 consecutive steps). Figure 2a shows the performance versus time. First, there is a period which we can identify as a learning period in which the success rate is low and oscillating. Eventually the networks locks into a state where success is obtained very quickly in response to the switching of inputs. In this phase, the system reacts intelligently to the input signal. It switches quickly back and forth between the two correct outputs. The transition from the learning phase to the retrieval phase is quite abrupt. We emphasize that no outside switch was activated at this point. Figure 2b shows a similar curve for the random-topology case. Again a sharp, self-organized transition from a learning mode to a retrieval mode is observed.

What happens inside the network during the learning phase? Through a complicated self-organization process, the system creates internal contacts or connections between selected parts of the input signal and the correct output cell(s). The process can be thought of as the formation of a river network connecting output with input. When the output is incorrect, the river flow is reduced at existing connections. When the flow is correct, the flow is reinforced. When there is too little output, the river beds are widened.

The state of the system after completion of the learning phase cannot be calculated by means of a simple algorithm. (The synapses are formed by self-organization rather than design). The "fast" dynamical switching between one connection pattern and another under switching of the outside signal in the "retrieval" phase following the long learning phase is quite complicated. Figure 3a,b shows the firing patterns for the red and green responses, respectively. Figure 4 shows a movie of the switching process from the "red" response to the "green" response. The switching from "red" to "green" and back takes place through eleven intermediate steps. We doubt that any engineer would come up with such a solution. If we were free to construct the network "by design" we could obviously

come up with a much simpler and efficient solution. The memory lies in the conservation of parts of the river beds from previous correct connections.[7]

The system has self-organized into a state where the change of "water supply" at random positions causes a fast conversion to the correct output. In the learning phase the system is very sensitive to the relative small changes in input- in that sense it is chaotic. No such dynamical switching takes place in conventional neural networks where connections are essentially hard-wired in the retrieval mode.

Figure 5 shows the response to "damage" of the network. After ~150000 time steps, a block of 30 neurons was removed from the network. After a transient period the network has relearned the correct response, carving new connections in the network. In other words, instead of using some features of the input signal the system learns to use other features. Think of this as replacing "vision" with "smell." The memory is distributed and robust,[8] as it should be in order to represent real brain function. The new firing patterns are shown in Figure 3c,d.

Figure 6 shows the situation where a third input (and corresponding pair of output cells) was added after the first two responses had been learned. After a transient period where the system is confused and the success rate is low, the network eventually learns the three appropriate responses. Figure 3e, f, and g shows the firing patterns after the three inputs have been learned.

A brain working according to the principles illustrated here requires a minimum of biological complexity – it is a relatively simple organ without much structure. Little information is needed to construct the simple network with essentially arbitrary connections. The correlations that control the switching behavior of the system hint to the fact that 'it is not only the well developed "riverbeds," and where most of the activity takes place, that are important for the function of the network but also the relatively silent regions in between.' Evidence of this can be seen in the rather complicated landscape of the $J_{i,j}$s (Fig. 3h,i). The landscape is strikingly rugged. This is somewhat counterintuitive. One might have expected well-carved riverbeds and isolated switches. Seemingly, this configuration which from an engineering point of view is more efficient, it is not compatible with the self-organization process.

In conclusion, we have constructed a simple model simulating aspects of brain function. The build-up of the $J_{i,j}$ landscape is due to a self-organization process. We suggest that simple robots performing "intelligent" tasks can be constructed following the principles outlined here.

The work done at Brookhaven National Laboratory was supported by the U.S. Department of Energy, Division of Materials Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76CH00016.

References

1. Amit, D. J. *Modelling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press, Cambridge, 1989).
2. Hertz, J., Krogh A., & Palmer, R. G. *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood, 1991).
3. Alstrøm, P. & Stassinopoulos, D., submitted to Phys. Rev. Lett. (1994).
4. Stassinopoulos, D. & Bak, P., submitted to Nature (1994).
5. McCulloch, W. S. & Pitts, W. Bull. Math. Biophys. 5, 115 (1943).
6. The original AS-model[3] is "blind" in the sense that it operates with a fixed input signal at the upper layer.
7. A more detailed study of the dynamics of this learning mechanism is in progress.
8. To check for robustness we tested the performance of the system when signals were presented randomly and for an arbitrary duration of time.

Figure Captions

Figure 1. Block diagram of brain model. Each signal is represented by random inputs to a number of neurons. For each signal, here red or green, there is a combination of one or more output neurons (shaded circles) which must fire in order to achieve success. The environment feeds back a signal indicating whether or not success was achieved. a) Layered network; b) Random network.

Figure 2. a) Performance vs. time for layered system with two input signals which are switched every 2000 time units or when the system is consistently successfull. After a training period during which the network self-organizes, the system enters an intelligent state with fast switching between the correct outputs. b) Same for random network; the two input signals are presented for 5000 time units unless consistent success has occured.

Figure 3. Firing patterns. a,b) The two sets of input neurons are colored red and green, respectively. For the red input, output cells 10 and 15 of the bottom row must be triggered simultaneously to achieve success; for the green input the output cells 7 and 12 must be triggered. The yellow squares indicate neurons which are firing for the two inputs in the fast switching mode. c,d) The same as above but in the case where the system has relearned the correct response after removal of a block of 30 neurons (shaded area). Note the difference from the original response. e, f, g) Same as above, but with three inputs. The response of the two original inputs (e,f) is different from the original one (a, b). h, i) The configurations of $J_{i,j}$s pointing to the right, for the two cases discussed above (a-b, c-d). The different values are depicted with a rainbow-color map ranging from black and dark blue for the lowest values to red for the highest.

Figure 4. Movie showing the "fast" switching between the "green" response and the "red" response. The transition from "green" to "red" takes place through five complicated steps and back to "green" through an additional six steps.

Figure 5. Performance for the layered system, but with 30 neurons damaged after 150000 time steps. The system has relearned the correct response after 210000 time steps.

Figure 6. Same system as shown in Figure 2a, with a third input added after 150000 steps. After a confused learning period, the correct output for all three inputs is learned after 450000 time steps.

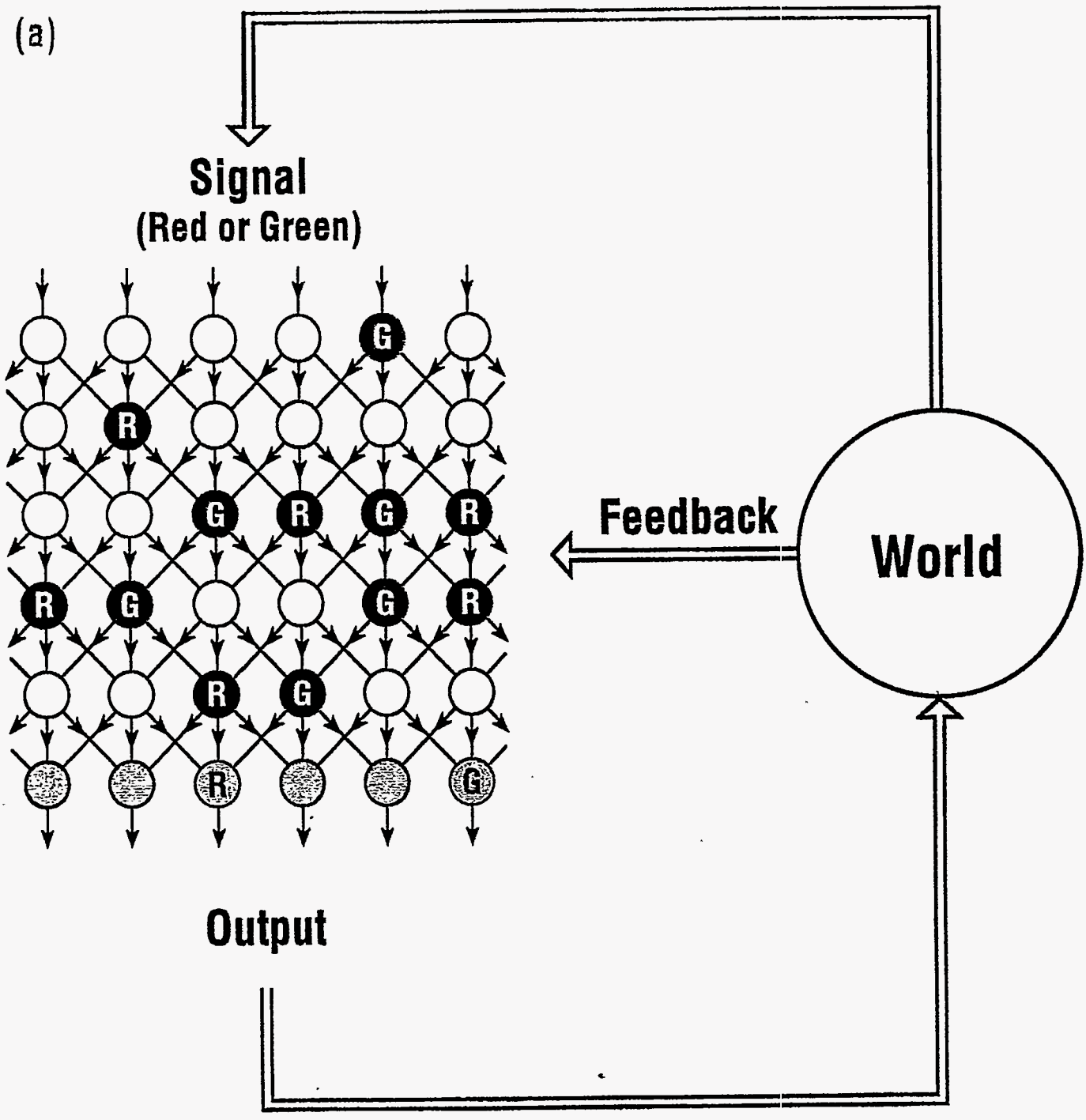


Figure 1

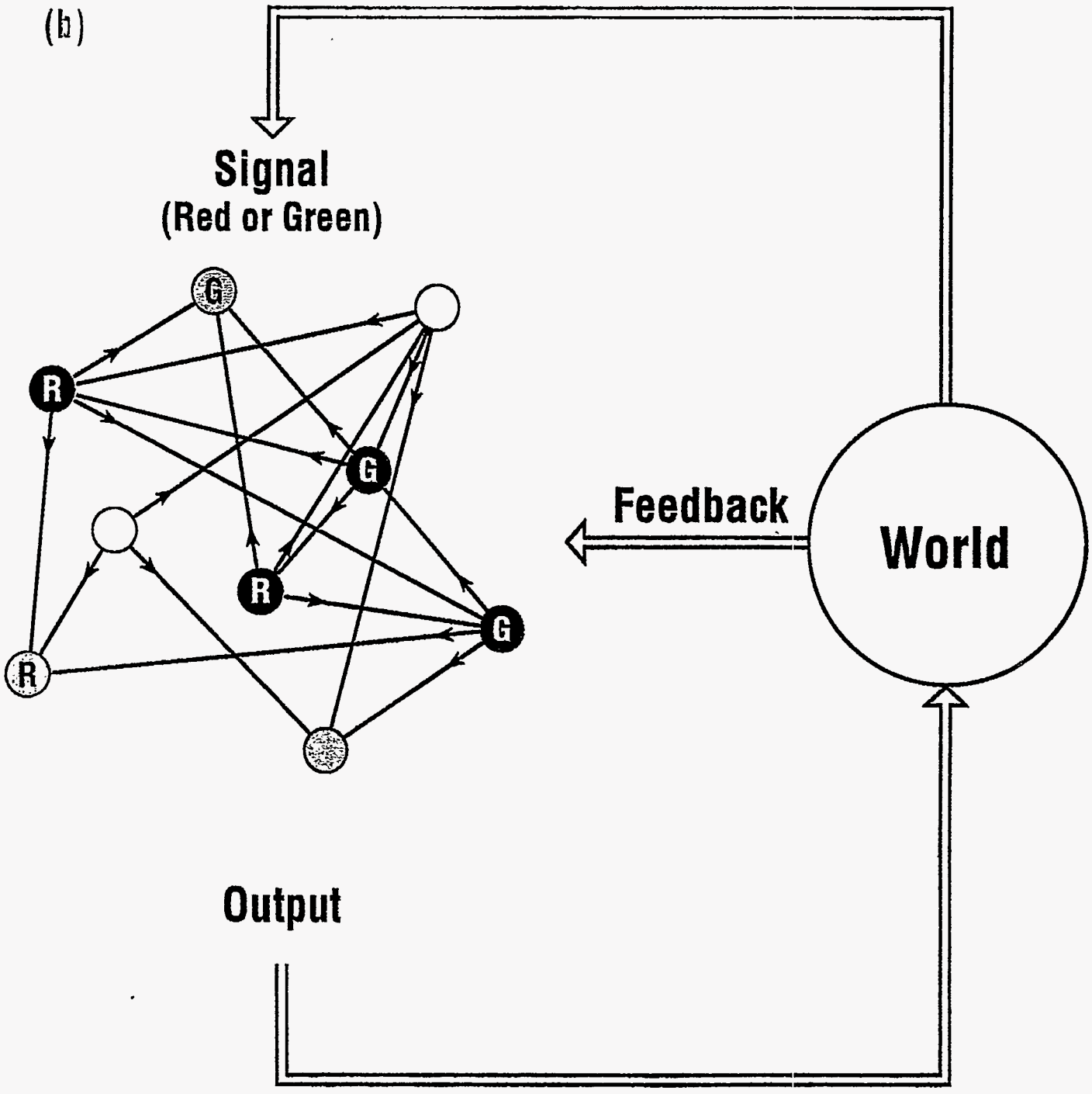
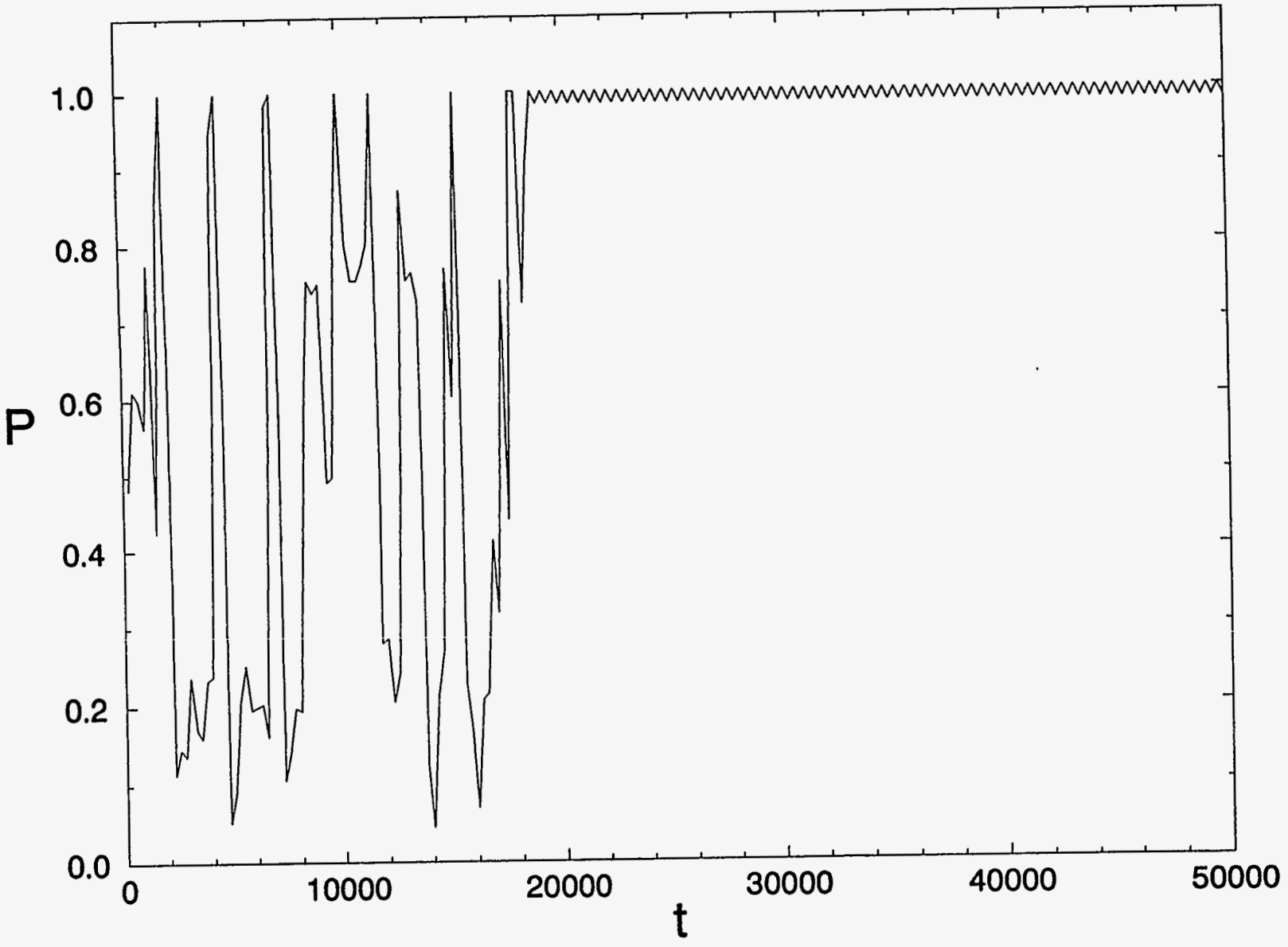
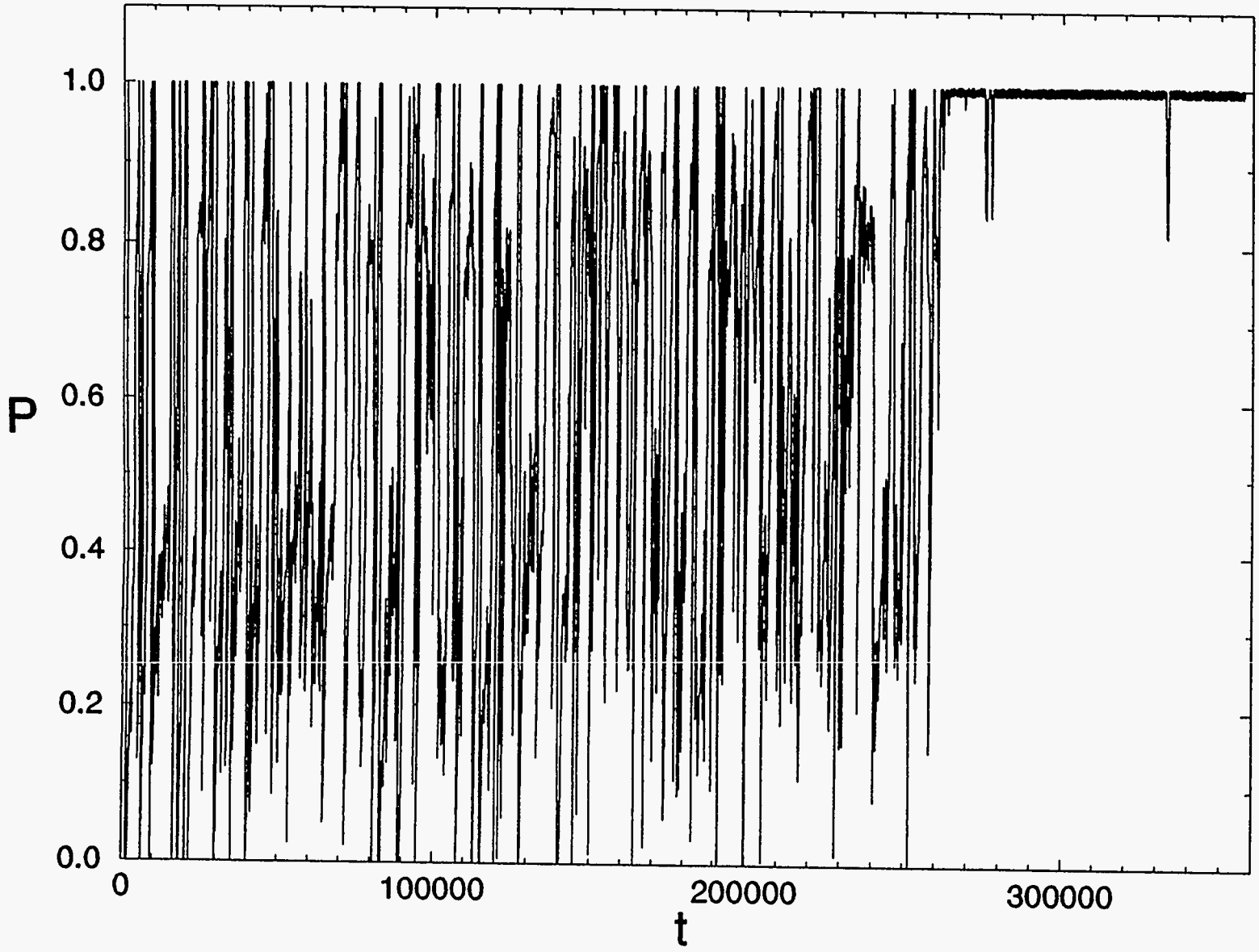


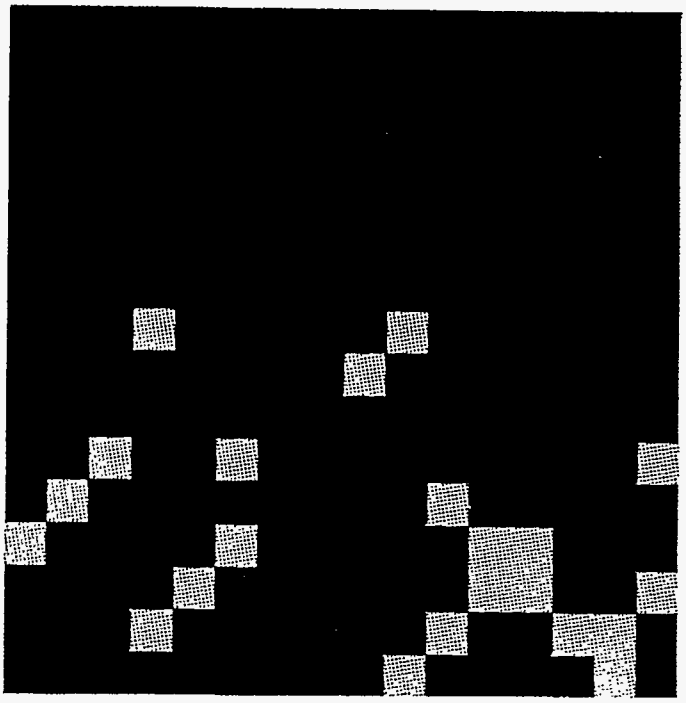
Figure 1

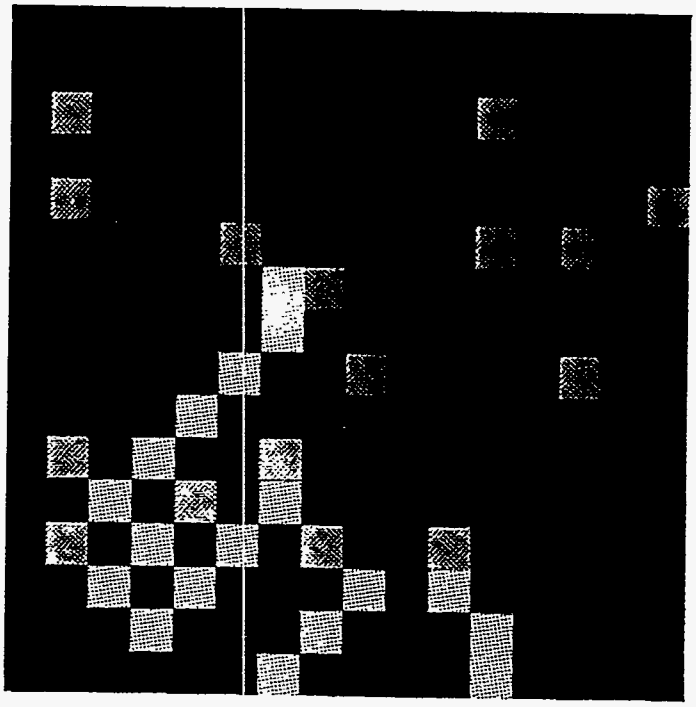
Classic papers D & Bak, P., & Ástrom, P.


1.0
0.8
0.6
0.4
0.2
0.0
P
0
10000
20000
30000
40000
50000
t

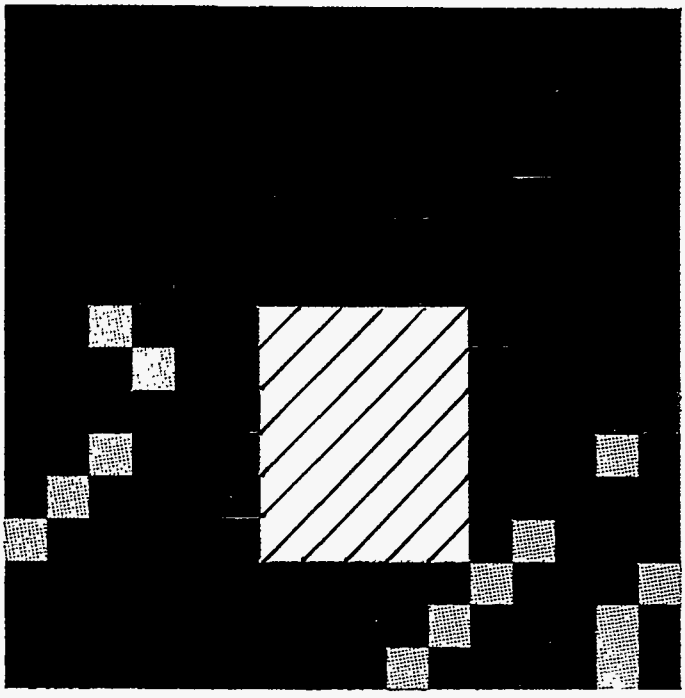

1.0
0.8
0.6
0.4
0.2
0.0
P
0
100000
200000
300000
t

(a)





(c)



(d)

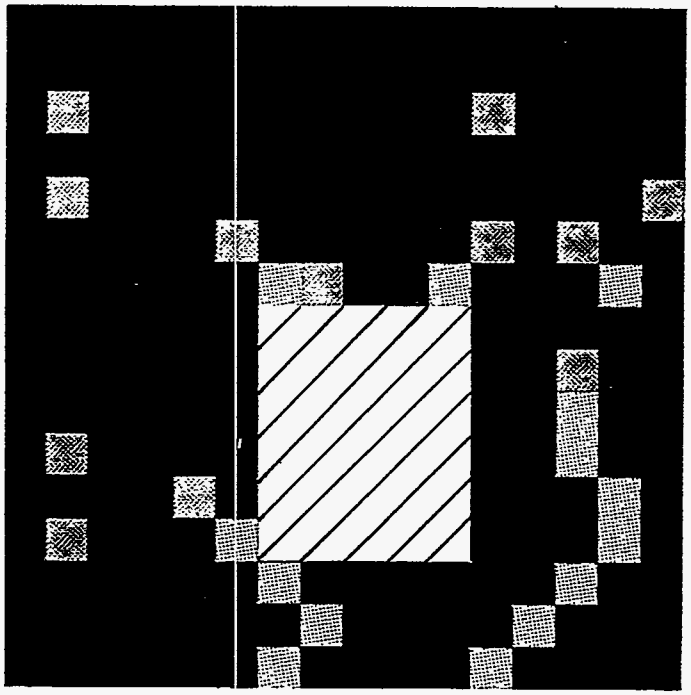
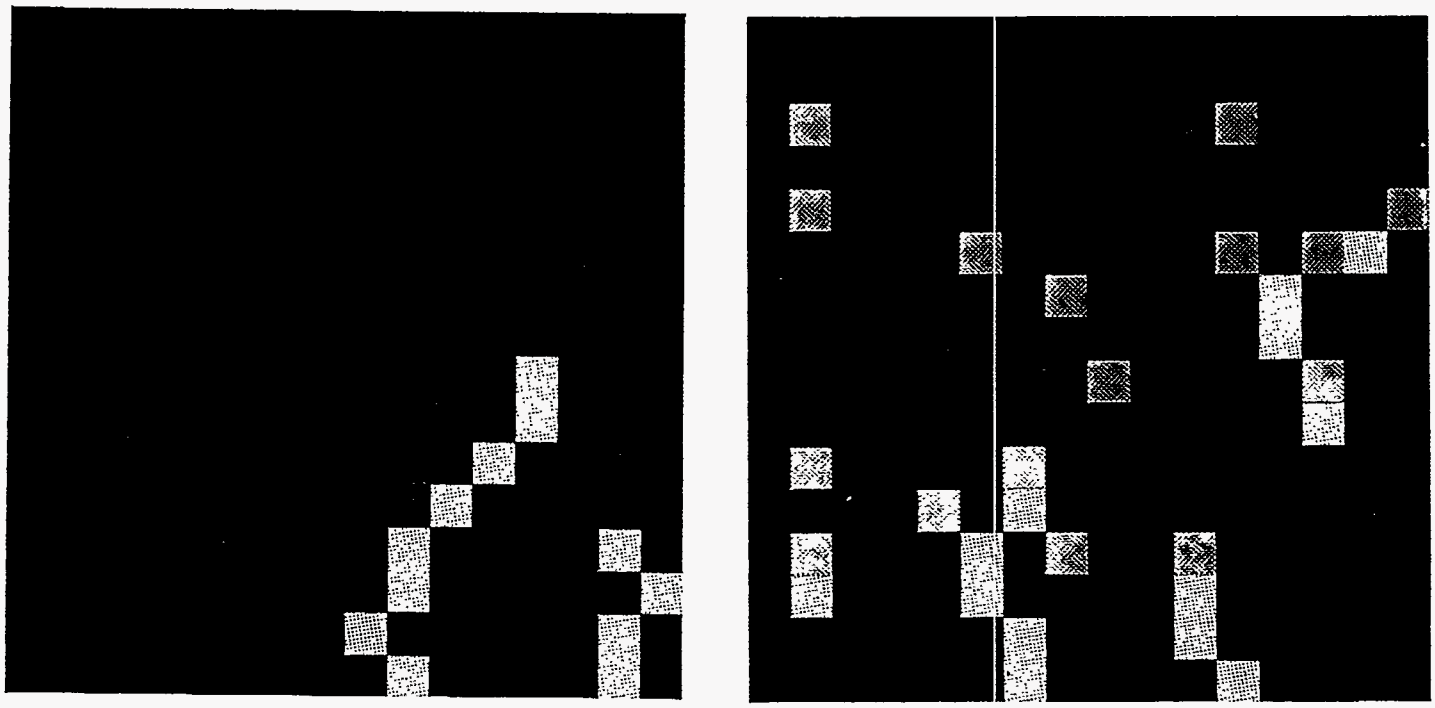


Fig. 3



(g)

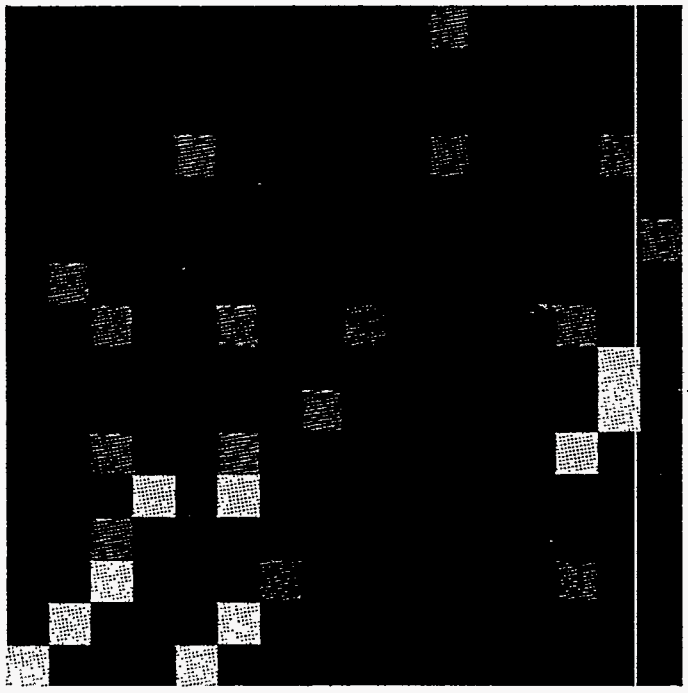
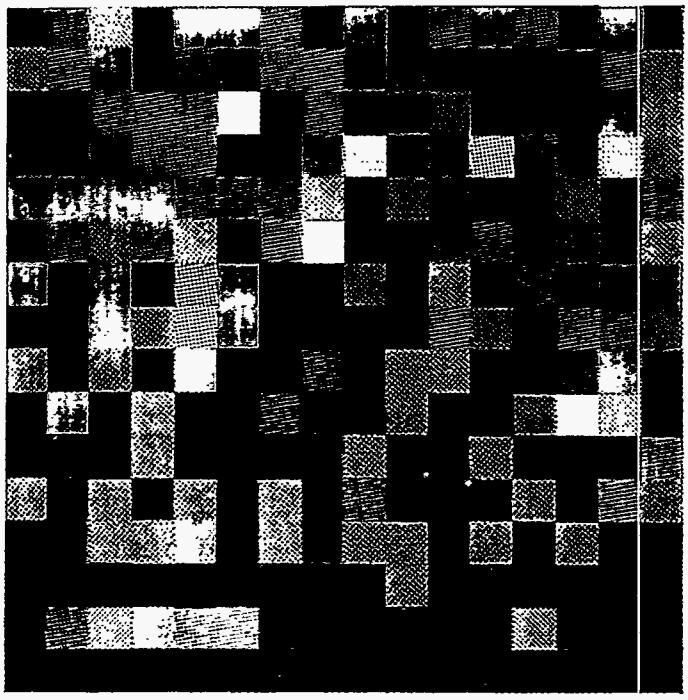


Fig. 3



(i)

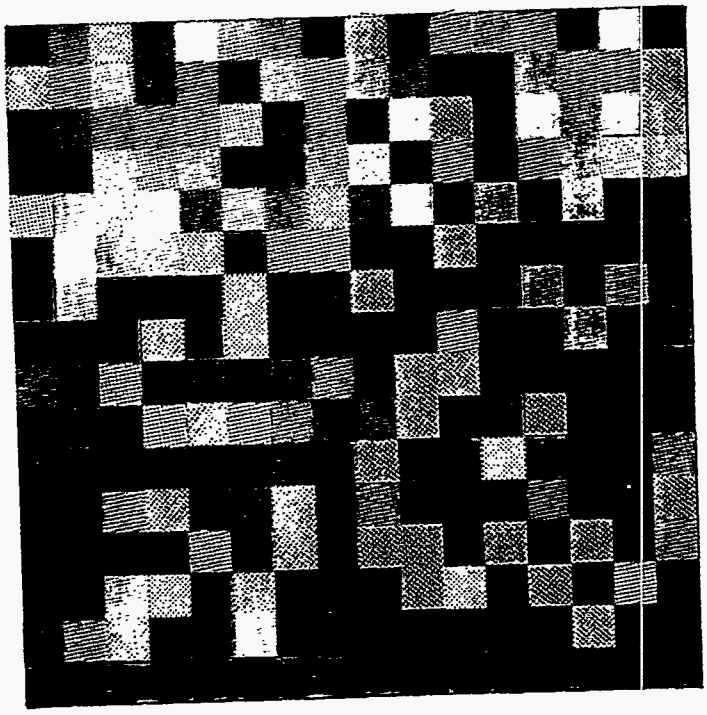
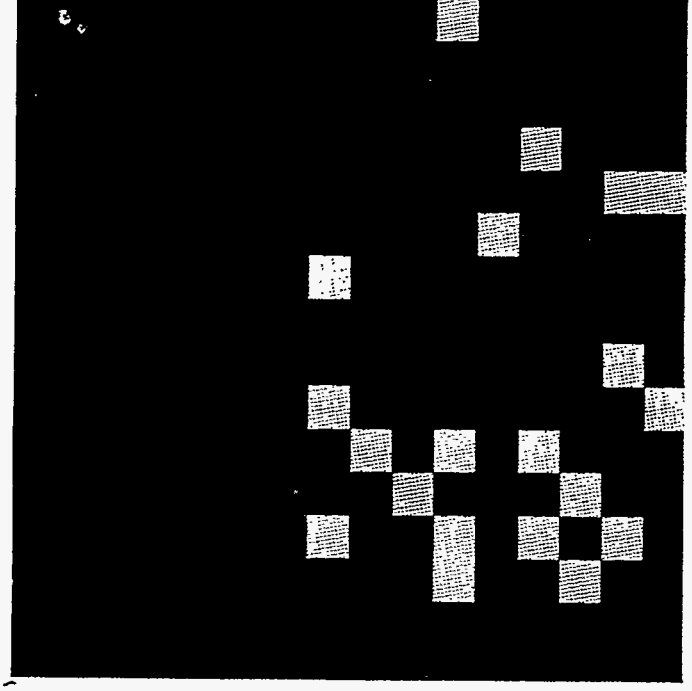
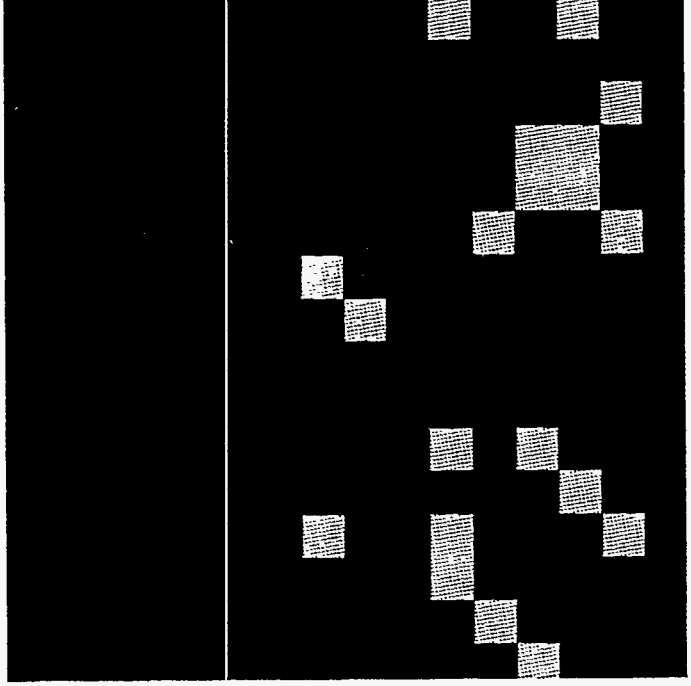
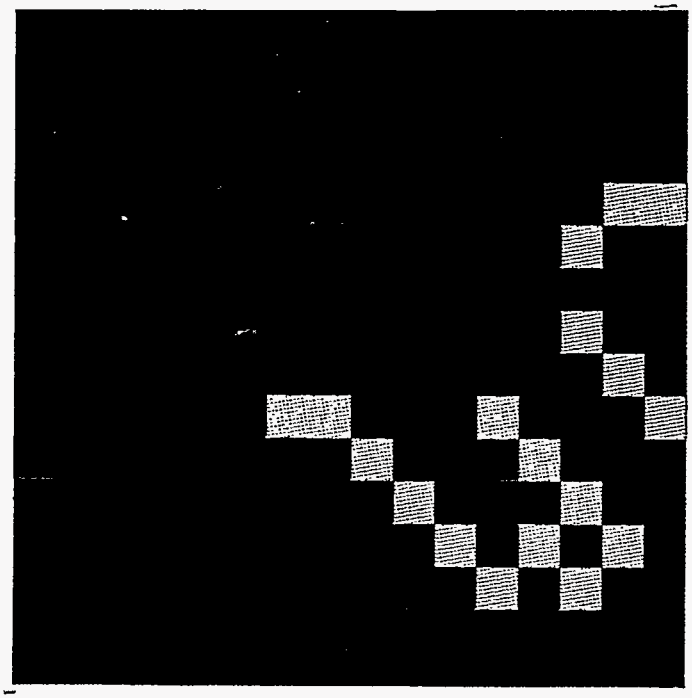
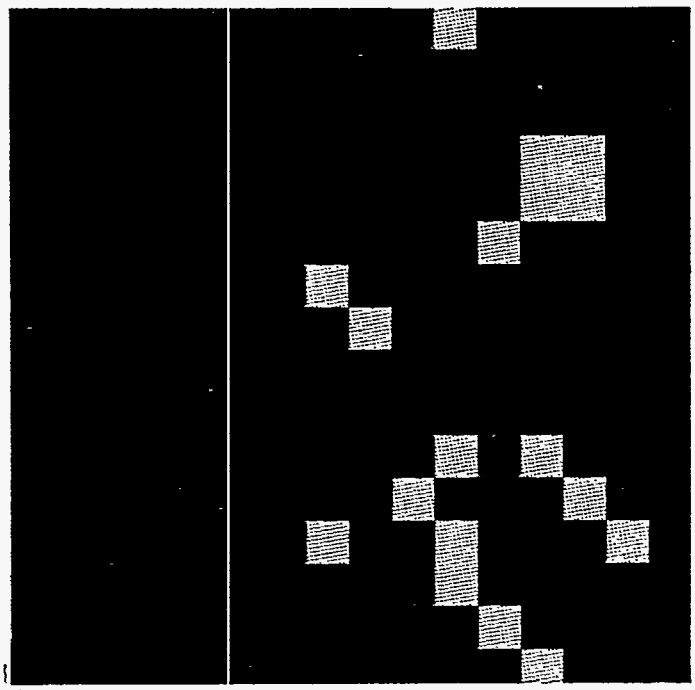


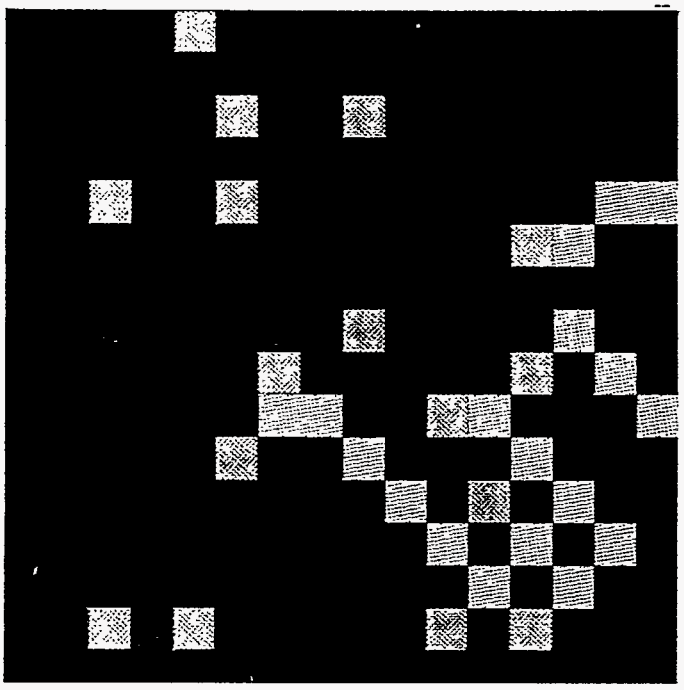
Fig. 3

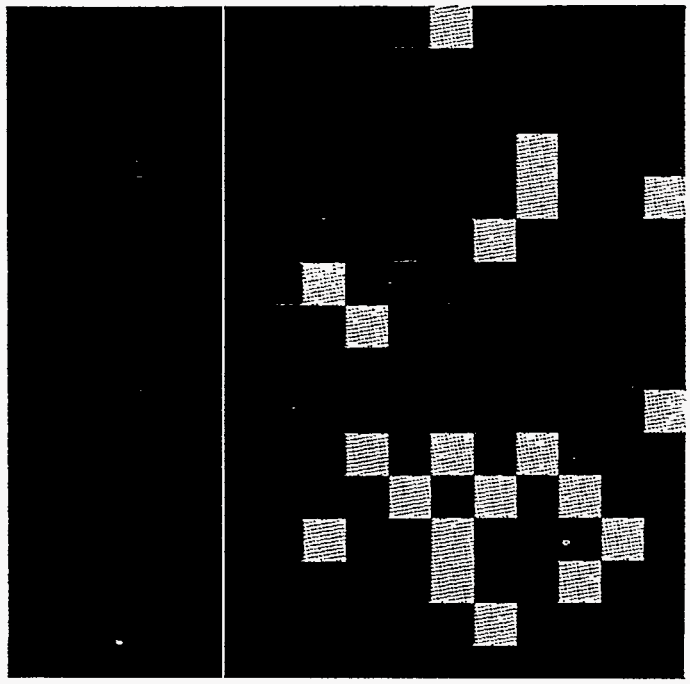


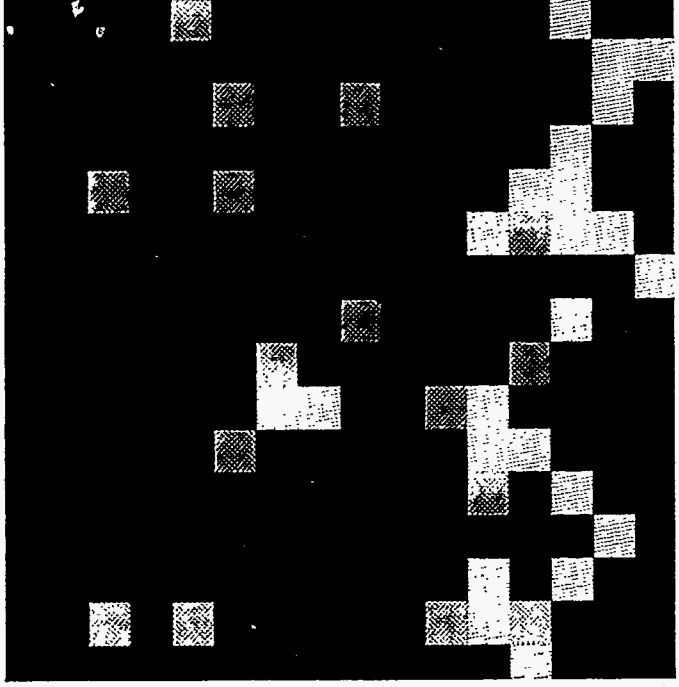


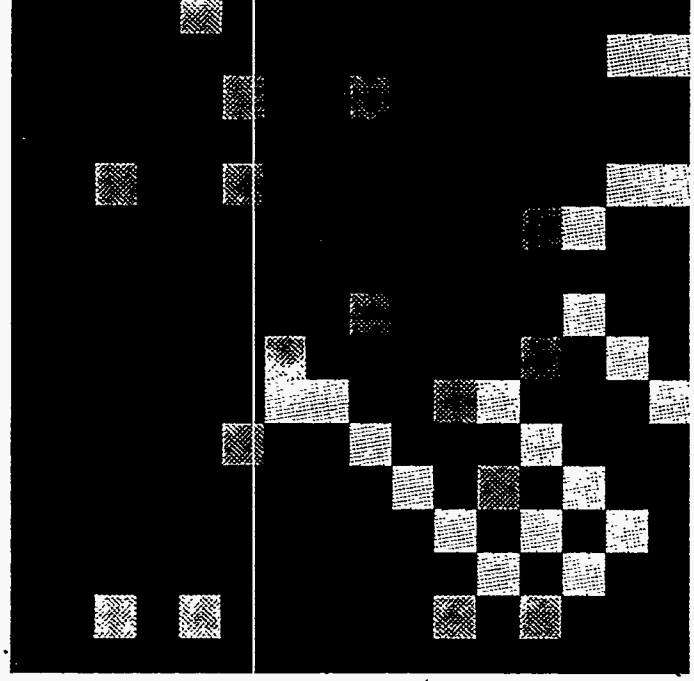


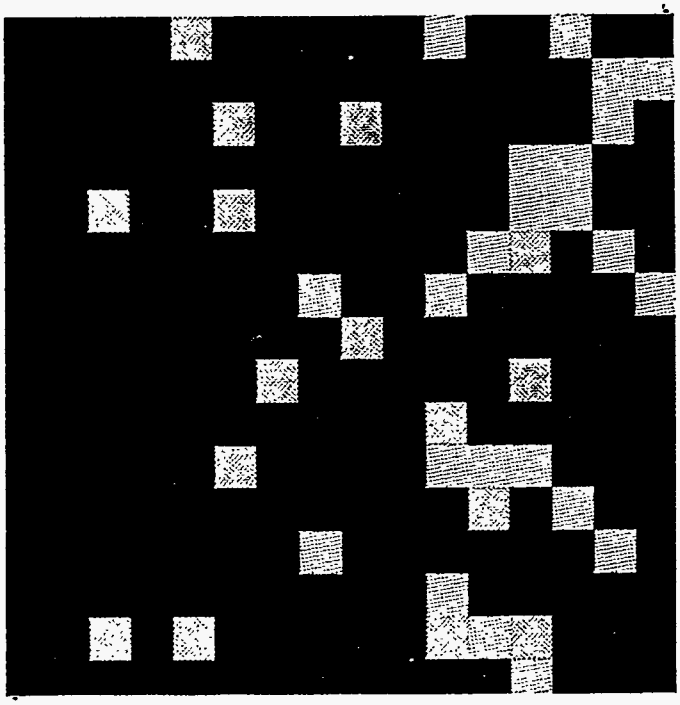


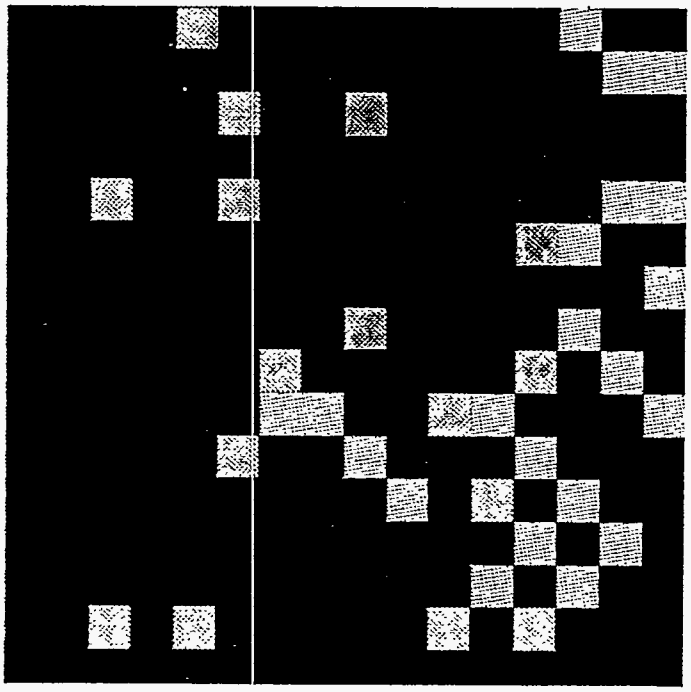


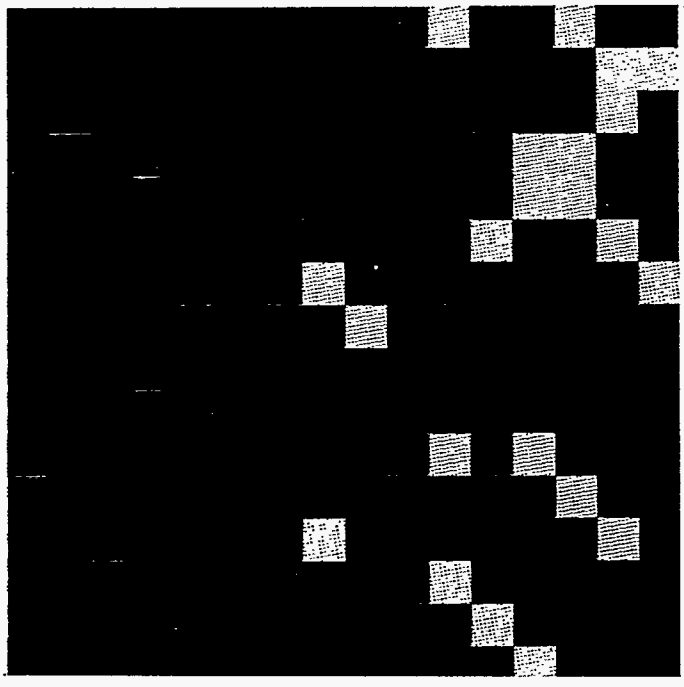


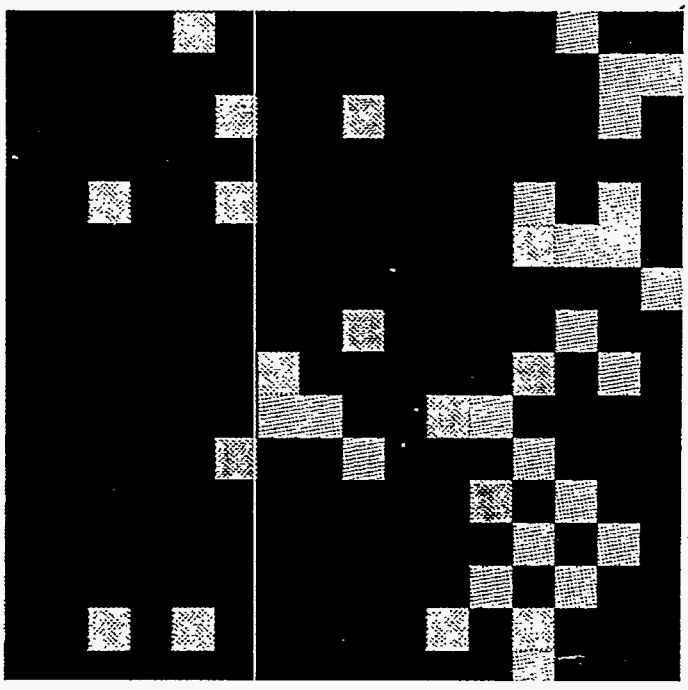












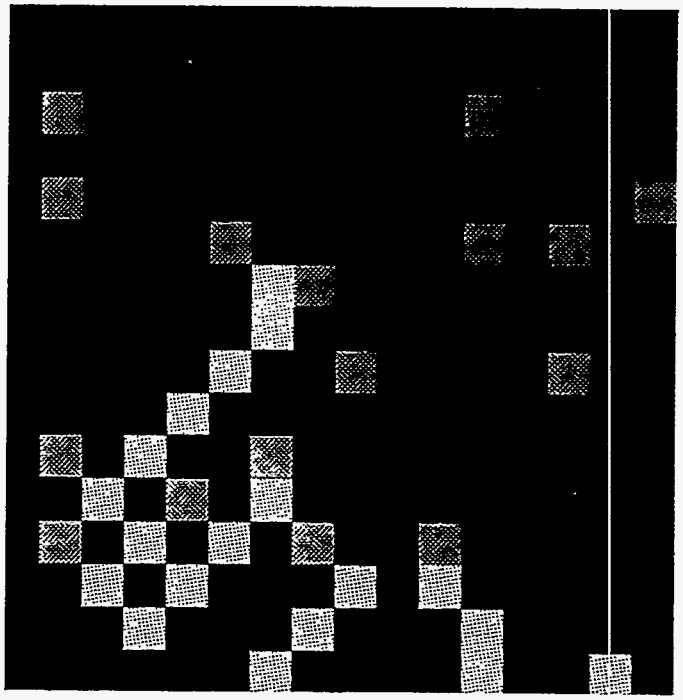


Fig. 4

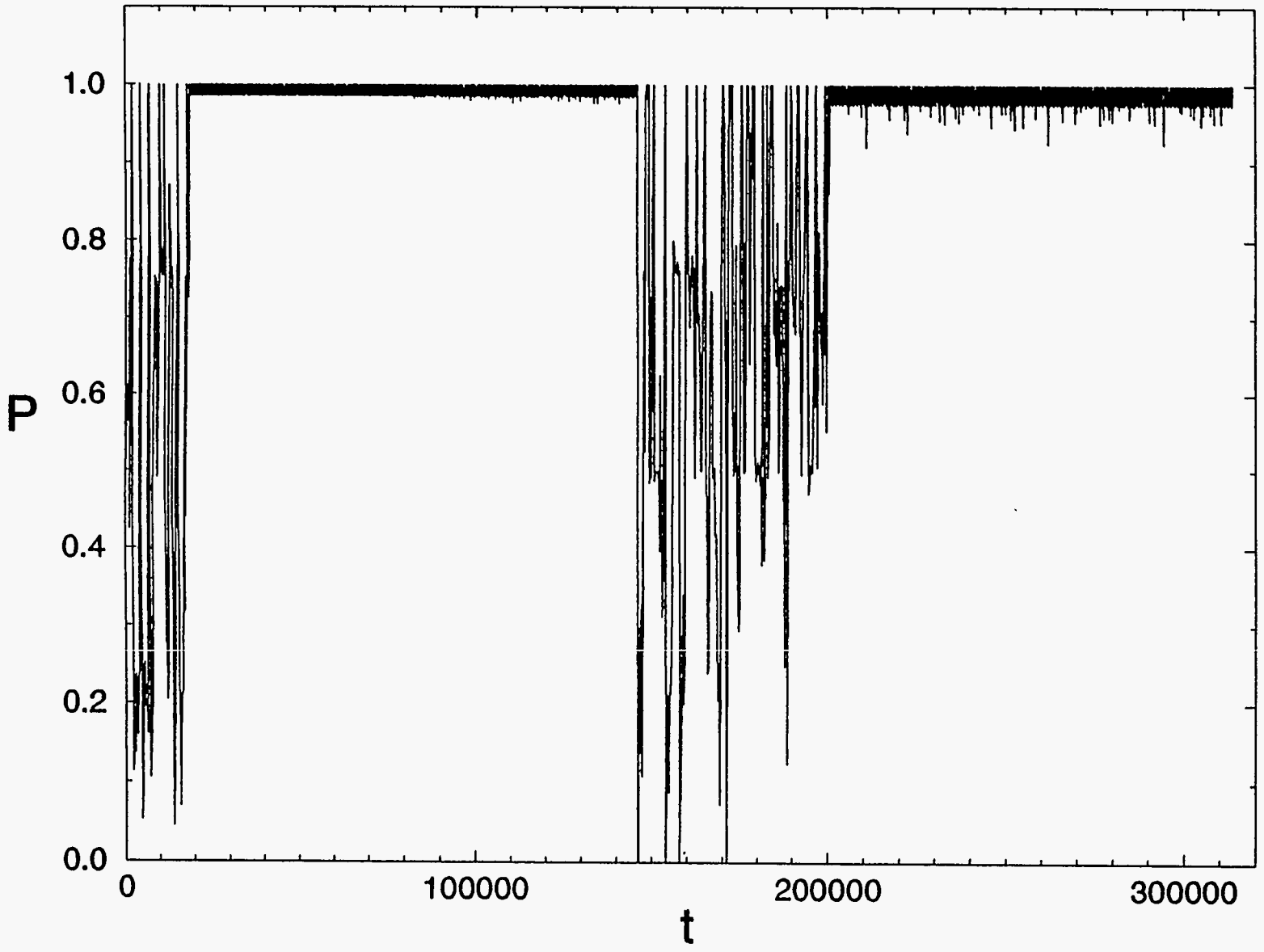


Figure 5

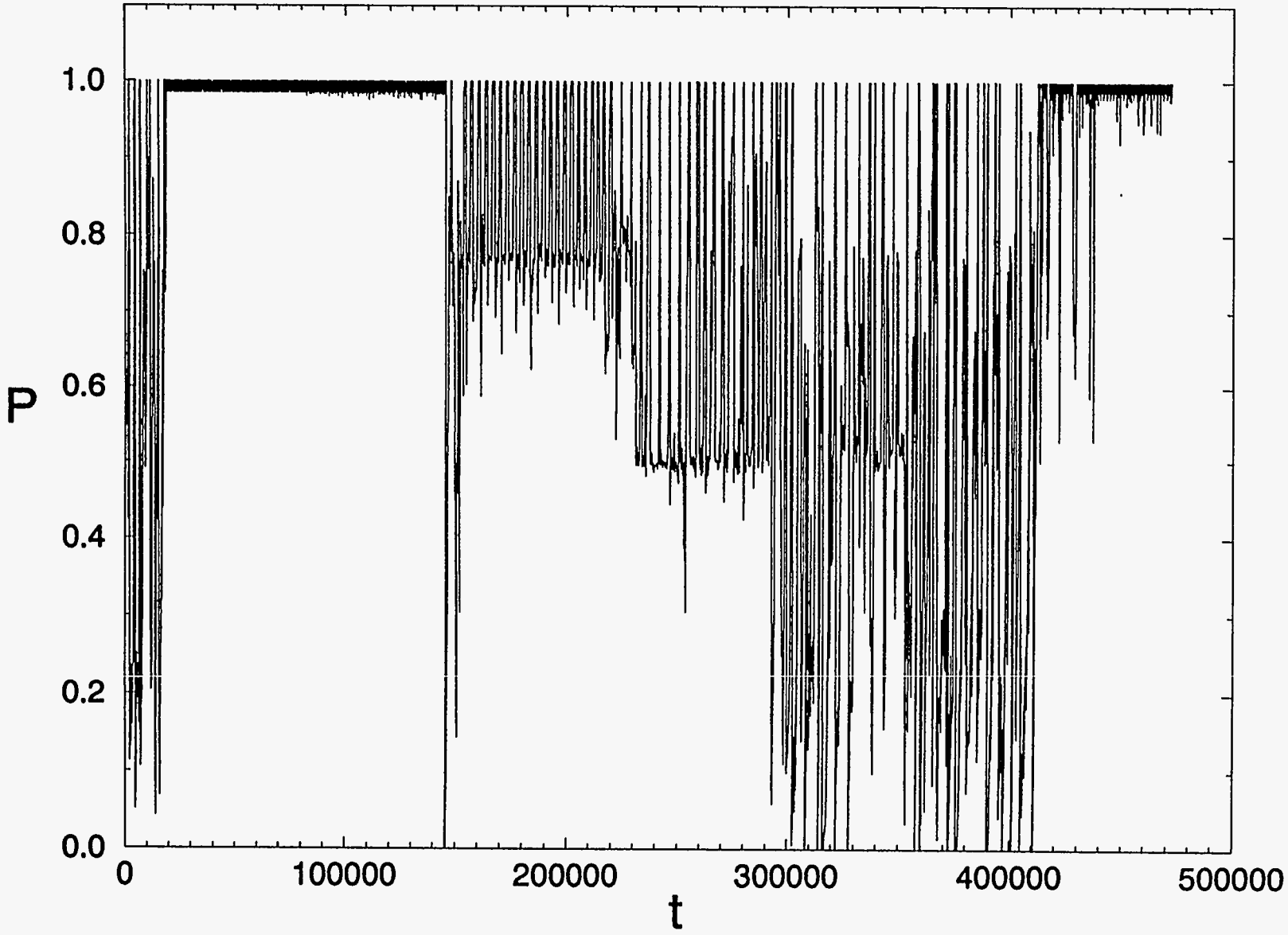


Figure 6