The Lesson of Quantum Theory, edited by J. de Boer, E. Dal and O. Ulfbeck

Physics, Biological Computation and Complementarity

John J. Hopfield
California Institute of Technology
Pasadena, California, USA

Contents

1. The domain of physics in biology

Biology as we know it lies within a restricted domain of physics. The laws of elementary particle physics and cosmology and the history of the universe serve merely to determine the nature of a planetary environment. The dynamical equations of quantum mechanics and quantum electrodynamics (and their classical equivalents when appropriate) are the essential elemental laws of physics which lead to biology. Some physicists make claims that "we shall never understand life until we understand the origins of the elementary particles". But the real mysteries of biology lie in the way in which these dynamical laws of physics, and the substrate of electrons, photons and nuclei on which they operate, produce the complex set of counter-intuitive phenomena labeled with the term biology.

Biology is a problem in dynamics—an organism functions by irreversibly preying on the available free energy which it finds in its environment in order to maintain its dynamic state. Driven (or non-equilibrium) physical systems of simple components already show complex and almost unpredictable behaviors. Turbulence in fluid flow, deterministic chaos and fractal forms in snowflakes are a few of the complex phenomena which arise from the same simple physical laws and substrates that rule biology. Some of the basic unsolved problems of biology, such as the generation of complex forms in large systems, can already be seen in these simple systems. Our understanding of such problems in physics is far from complete. The physics of large systems in equilibrium is unified and simplified through our understanding of the derived or secondary laws of statistical mechanics and thermodynamics. There is no comparably general theory of strongly non-equilibrium systems. The major conceptual problems in biology have the additional complication that the biological matter is itself very complex due to a long and selective evolutionary history.

Biology is not a quantum-mechanical problem. Because the masses of nuclei are large compared to the masses of electrons, the adiabatic separation of electron and nuclear motion in the Schrödinger equation is usually adequate. The electrons can then be effectively removed from the problem, leaving a problem of only nuclear motion with effective interactions of some complexity between nuclei. This remaining problem of nuclear motion is partly in the classical regime, and partly a quantum-mechanical problem. The hydrogen stretching vibrations are of large quantum energy compared with kT at room temperature, and are rigid in the molecular dynamics. The rotational quantum numbers are large at room temperature, and the rotational motion is thus essentially classical, as is much of the translational motion. The molecular dynamics of liquid water at room temperature can be described in classical terms as the motion of rigid molecules having a complex set of two- and three-body forces between them. Accurate descriptions of the viscosity, rotational relaxation and dynamic neutron diffraction of liquid water have been obtained from the computed dynamics of such a model of water. Non-equilibrium problems such as the turbulence of water are in essence problems of classical physics, as is biology.

This is not to deny that there are intrinsic limitations in the knowledge with which we can know positions and momenta of nuclei. In this regard, the classical approximation ignores a quantum-mechanical limitation. But the essential mysteries, phenomena and complexity of biology are not a problem of Planck's constant. They are a problem of the large size of Avagadro's number combined with the non-equilibrium nature of the system. There do not seem to be any important larger quantum coherent aspects to living matter, contrary to the romantic hopes of some physicists of the 1930s.

2. Logical, physical and biological computers

Biology can be seen as a hierarchy of computations and computational devices. The translation of DNA into protein structure is a kind of computation. The construction of a complex organism from the instructions in DNA is also the following of an algorithm. The recognition of a familiar object is a neural computation. These diverse computations share some common characteristics because they share an evolutionary background. The species which exist in biology today are those which have survived the competition with other organisms and the cataclysms of weather and geology. At the other end of the size scale, the proteins which exist within a given species have survived a competition with different molecules which might have performed the same functions, and with the alternative of simply not having this function performed at all. The fundamental survival and competition problem for the organism or for the enzyme molecule is to develop an algorithm for predicting the future, or more accurately, to develop a behavior such that physical processes (or actions) taken now will be likely to be appropriate to the future environment in which the organism or molecule finds itself, and promote the survival or reproduction of the organism. The organism which survives best will be that which most clearly "sees" the consequences of its possible present actions. The prediction of the future from present information is a kind of computation. To understand complex aspects of biology beyond the descriptive level, it becomes necessary to think about the computational aspects—what is computation and how does biological computation differ from that which we think about in conventional computers. Before delving into neurobiological computation, we will illustrate some of the issues.

Computation has three conceptual elements: an input, an output and a "device", which reads the input and produces the output. The particular output, or range of outputs, is the consequence both of the particular input and the nature of the computing "device". The classic conceptual device of computational theory is the Turing machine. This machine has several internal states. It can read an input–output tape, shift the tape, write output on the tape and change the internal state of the machine. Universal computation can be performed by such machines. Turing machines are intrinsically digital (or logical) machines, having a small number of reading and writing symbols and a finite number of internal states. When computation is a physical process, as in a real computer or in biology, each of these elements has a physical manifestation.

The description of computers as logical devices has been thoroughly developed in the past 50 years. A computer is also a physical dynamical system, which follows the laws and limitations of physics. The object of a computer designer is to develop a real dynamic system which behaves as similarly as possible to some given logical design. Physical systems have noise and imperfections not envisioned in the simple logical view of computer function. As a result, considerable effort in hardware design is invested in a problem of no logical concern, namely trying to get reliable computation in a noisy and fault-filled world. Biology compounds this problem by adding a unique view of the nature of errors and of logic itself. We will examine these points by an example from protein synthesis.

A growing cell is constantly producing more proteins [see, for example, Watson (1976)]. In this process, the information in a strand of messenger RNA is used as an instruction for making a protein. The input tape consists of a single strand of RNA. The output is a protein, a linear polymer of amino acids which then folds into a functional three-dimensional structure. The mRNA is a polymer made up of four kind of units, A, U, G and C. The protein is a polymer made up of twenty different kinds of amino acids, glycine, alanine, tyrosine, ...

AUGGGUCCAAAGAGCCUGUGG....... UGA...... mRNA
Met Gly Pro Lys Ser Leu Trp stop protein

Protein synthesis is carried out at a polymolecular assembly of proteins and RNA called a ribosome, which is the "Turing machine" for the process. Many different molecules participate in the protein synthesis on the ribosome, including tRNA, GTP and a host of co-factors. The ribosome reads the mRNA input tape and inserts the appropriate amino acids into the protein in sequence. The logical operation performed might be described as the following instruction set:

(1) Read the next three bases on the mRNA molecule "tape".
(2) Look up the corresponding amino acid in the genetic code "dictionary".
(3) Add that amino acid to the protein.
(4) Shift the input tape by three bases.

The program contains a possible "stop" instruction in the code dictionary. The signal for starting is more complex.

The logical operation of reading a particular next codon in the mRNA is actually carried out in a chemical reaction which knows nothing about logic. The reaction might schematically be written:

$$\text{peptide-}n + \text{correct amino acid} \rightarrow \text{peptide-}n + 1.$$

Correct refers to the amino acid in correspondence with the next mRNA triplet. There is, however, also a competing reaction

$$\text{peptide-}n + \text{incorrect amino acid} \rightarrow \text{incorrect peptide-}n + 1.$$

Within the logical view of computation this competing reaction simply "doesn't happen". From the point of view of chemistry, the competing reaction can happen, but has a different energy barrier for taking place. The reaction is rather similar to the one desired, and any enzymatic system which allows the correct reaction to take place must also allow the incorrect reaction, albeit with rather slower rates. The equilibrium constant for adding the correct amino acid and for adding the incorrect one are essentially identical. Equilibrium processes—physical processes which take place sufficiently slowly—result in useless proteins being formed because a particular incorrect amino acid is as likely to be added as a correct one. The logic which we would like the biological system to display—correct amino acid only—is not possible without errors, and is possible with low but finite errors only by deliberately running the system out of equilibrium. In non-equilibrium processes, the choice between products can be made on the basis of rates. The biological process must be designed to make use of kinetics to obtain accurate logical calculation. The qualitative description of biochemistry—that reactions take place because molecules fit together, and other reactions do not because the corresponding enzymatic binding does not happen—consists of logical statements, and by being only logical overlooks the essential non-equilibrium element to real physical computation.

Ordinary computers also make use of dissipation to obtain their accuracy. The course of a computation might be described as a motion in computer state space, or in a physical phase space. The initial data define a starting point in that space, and the computer is supposed to follow an appropriate and determinate path to the solution, represented by another point in that space. This is illustrated by the solid line with arrows in fig. 1. The effect of noise and imperfections is to cause the flow to deviate from the desired direction, and an accumulation of noise or statistical imperfections will result in a wrong answer.

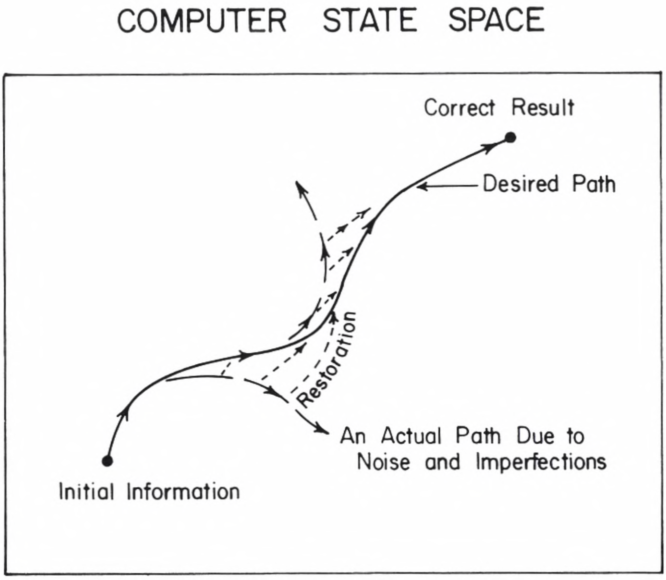

Fig. 1. The dynamical trajectory of a computing machine from initial information to the appropriate answer. In a noiseless perfect machine, the desired path would be directly followed. Noise and imperfections cause increasing divergence from this path, while restoration returns the trajectory toward the ideal one.

The way to avoid this problem is to introduce a physical process which squeezes the system back down onto the proper track in state space. If we can keep this compression going on, it will result in finding the correct answer in spite of noise. Compressing a bunch of trajectories in phase space down into a smaller volume is easily done by a dissipative process. This squeezing down is essential to avoiding errors in a system which has the physical possibility of making them. The reason that protein synthesis must be run out of equilibrium is exactly this necessity for compacting the occupied state space in a dissipative fashion to avoid errors. (In electronic digital computers, the electrical power necessary to do a computation is in principle bounded by this aspect, though in fact the power dissipation is far larger for reasons which have nothing to do with fundamental physics.) In digital computers, this idea is called restoration (von Neuman 1952, Mead and Conway 1980), and the fact that a nominal digital "one" (really a particular voltage level) will recover to the appropriate level after a transient perturbation is given to the circuit is an illustration of the presence of restoration. In thinking about the computations which are performed in neurobiology, we must expect that the system will show such restoration in an obvious fashion, or else the system would be unable to compute. Restoration—which gives a robustness against noise and computer imperfections–is a universal necessity to physical computers, while an irrelevancy in logical computers.

Even with restoration, biological computers are inaccurate. The intrinsic accuracy of protein synthesis of the simplest sort based on elementary chemical recognition, is about 1 error in 100 in difficult cases (Pauling 1957, Hopfield and Yamane 1980). The protein synthesis system also has a proofreading system (Hopfield 1974), which consumes more free energy and adds another layer of restoration, resulting in error levels on the scale of 1/3000, a drastic level from the point of view of electronic machines, but apparently good enough for biology.

The one additional oddity about biological computers is that they do not perform as well as possible. Streptomycin resistant mutants of bacteria proofread better (Yates 1979) and are more accurate than the normal wild-type bacteria, but are an artifact of an evolutionary history with streptomycin present in the growth environment. In the absence of streptomycin, the bacterium reverts to the wild type—which is less accurate. The same general kind of less-than-best accuracy has been also demonstrated in DNA synthesis in T4 bacteriophage by Muzyczka et al. (1972).

In biology, being as accurate as possible in computations may be a loosing proposition! Biology is not interested in perfect logic. The possibility of making progress through random accidents—called creative thought when they occur in neurobiology, or evolution and speciation when they occur in molecular biology—seems to be an essential part of biological computation.

3. *Neural computation*

We turn next to computation in neurobiology. The object is to understand how a set of neurons makes decisions, generates actions, generalizes and learns and profits from past experiences. The emphasis here must be on our understanding. To merely know the input–output relation of a set of neurons by exhaustive study is not satisfying. Worse, a clump of neurons with 100 input neurons and 100 output neurons might easily require more than 10^{40} bits of information to characterize the input–output relation, making exhaustive study impossible. It would be equally unsatisfying to know the neural hardware in sufficient detail to be able to simulate the hardware on a monsterous digital computer and predict correctly the behavior of a neural system. This would correspond to being able to simulate the behavior of a classical gas of complex molecules, without the conceptual understandings brought to such gases by statistical mechanics and thermodynamics.

Perhaps the largest computational burden placed on our brains is involved in visual perception. The end result of this computation is our decisions about what we have seen. (It is appropriate to emphasize decisions, for decisions are the essence of computation. Strictly linear systems do not truly compute, although they can be very useful elements in a computational system.) Given a flash exposure to a typical visual scene, we note the presence of a few familiar objects, some rough characteristics of each, such as color, general size, etc. The immense supply of almost non-meaningful information—more than 10^9 bits of information were processed by the retinal cells which begin this calculation—is compressed into significant perceptual information, of which there are probably only a few thousand bits. In a digital machine such a computation would be done by making a very large number of sequential decisions, but somehow the essence of biological decisions seems to be rather more holistic, collective, or Gestalt. We want to understand how such decisions are made.

Aspects of neural computations in higher animals which are particularly puzzling from a physics viewpoint include:

(1) The system makes very effective use of its computational resources, whether measured by speed of calculation or volume or energy considerations.

(2) In such systems, the computations done by the system are very resistant to damage to the neural system (fail soft).

(3) Emergent properties such as self-awareness seem to be present.

(4) In higher animals, neurobiology manages to function without a determined circuit diagram.

There are two reasons to think that progress might be made in understanding neural function. First, the system is large, the connectivity between neurons is large, the behavior is somewhat insensitive to the destruction or misfunction of components, and the calculation seems to have a somewhat holistic character. This suggests that collective effects might be involved, and that a search for collective effects in neural networks [Little (1974); Little and Shaw (1978); Hopfield (1982)] and neural computation might be fruitful. (Note that this is not the way that most chip designers work—they do not make use of collective effects, and would attempt to suppress them if they ever were to be noticed.) Second, the system cannot be as complex as it might appear. It is true that a general module of neurons with 100 inputs and 100 outputs could require 10^{40} bits to specify its behavior, and would require such a number if all we think of is to list them. But a general module would require also 10^{40} bits of information to describe how to build it. A simple module, which can be described in perhaps 10 000 bits cannot produce a general input–output relation. It must produce a very special kind of input–output relation, which can be only apparently complex, not truly complex. (In this same fashion, the random-number generators which are used in computers appear to provide highly random numbers, but in fact generate very special sequences because the generators can be described by short programs.)

4. Classical neurodynamics

We must understand what computational or circuit facilities a nervous system has at its disposal in order to see what is a mystery and what is trivial. The anatomy of a "typical" cortical neuron is sketched in fig. 2. The morphological and functional diversity of such cells is very large. We will briefly review the electrophysiology of such cells, which is covered in detail in textbooks (see, for example, Kandel and Schwarz 1981).

A small electrode can be inserted into a cell body, and the potential difference between the inside and the outside of the cell can be studied as a function of time. A typical result of such a recording is a baseline potential of about −90 millivolts, on which is superimposed a set of more or less stereotyped voltage "spikes" called action potentials, rising to a potential of about +50 millivolts, and being of about 1 millisecond duration. An action potential propagates from a cell body down an axon by an active regenerative process, and can thus propagate long distances without attenuation.

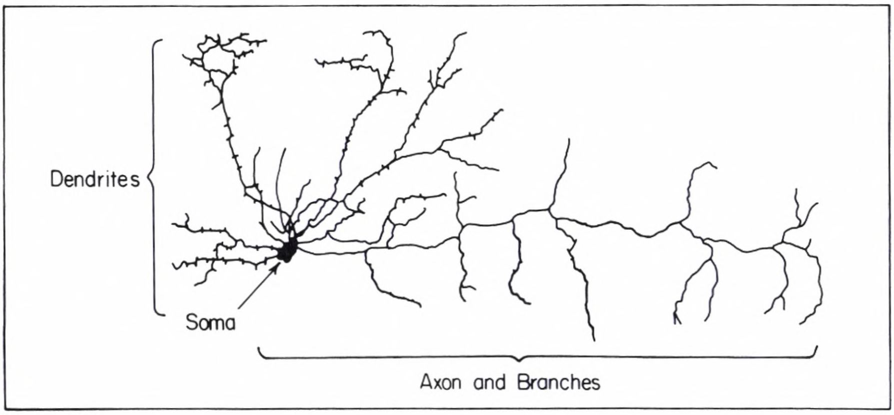

Fig. 2. A "typical" neuron from brain, showing the dendritic arbor (chiefly inputs), the cell body or soma, and the axon or chief output path.

The rate of generation of action potentials depends on the input to the dendritic arborization. Inputs to the dendrites produce current flows which depolarize the cell body, and if the cell body is depolarized enough, an action potential will be generated. This process can be readily simulated by passing a positive DC current into the cell body. The rate of firing (generation of action potentials) as a function of that positive current is sketched in fig. 3. Such curves have generally a sigmoid form, going from zero for large negative currents to a saturating maximal value of 100 to 1000 per second (depending on the neuron) for large positive currents.

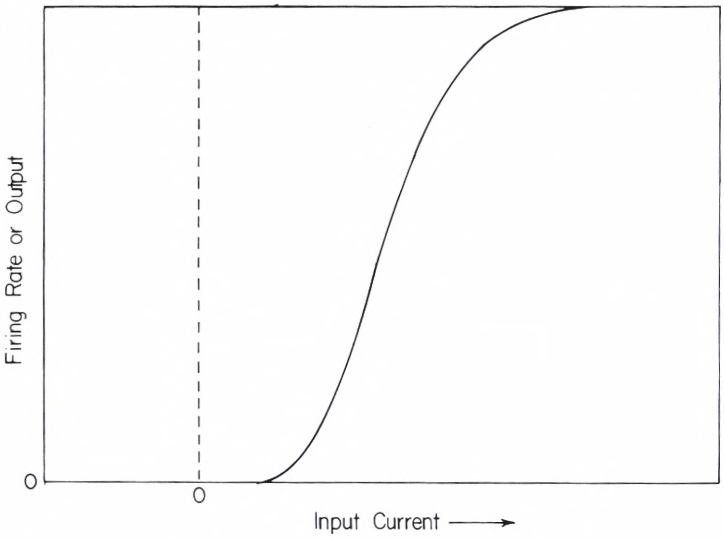

Fig. 3. The mean firing rate of a typical neuron as a function of the input current to the cell body. A sigmoid curve of this general shape is taken as the input–output relation $V = g(u)$ for the model neurons.

The ends of branches of axons of a cell are connected by synapses to the dendritic arborization of other cells. When an action potential reaches a synapse, a small amount of a chemical neurotransmitter is released at the axon terminal. Some of that transmitter becomes bound to receptors on the adjacent dendrite, and a current leak across the dendritic cell membrane is briefly produced. The detailed time history of this process depends on the type of cell, and on the particular neurotransmitter and receptor molecules involved, but is typically on the millisecond scale in simple motor neurons. This net current introduced into the cell body by such a process contains positive terms from synapses which excite the cell, and negative terms from synapses which inhibit (or suppress) the activity of that cell.

The electrical capacitance of a neuron is not negligible—cell membranes are only about 50 Å thick. The integrative time constant of a cell is appreciable, on the time scale of milliseconds. An individual action potential arriving at a synapse is generally not sufficient to result in the production of an action potential by the next cell. More typically, many action potentials must arrive (generally at many different synapses) on the dendritic arborization of a single cell within an integrative time constant before the post synaptic cell will fire.

A minimal set of properties which might be used to describe the functioning of a neural network might:

(a) Regard the individual action potentials as irrelevant quantization, and replace the true output of a cell i (a set of action potentials occurring at particular times) by a smooth, continuous property V_i describing the short-term average of the number of action potentials per unit time being sent out by that neuron. (Similarly, the discrete electrons are replaced by continuous charges and currents when one is dealing with electronic circuit behavior, and the quantization of electrical charge is generally neglected except when a noise analysis is done.)

(b) Describe the effective output V_i of a cell by a monotonic increasing sigmoid function $V_i = g_i(u_i)$ (where u_i is the input voltage of cell i) with $g(-\infty) = 0$ and $g(+\infty) = 1$. (The choice of 1 is only a convenient scaling.)

(c) Linearize the input dendritic arborization. The connection from neuron j to neuron i can then be described by a transconductance T_{ij}. The input to cell i can then be described as $\Sigma_j T_{ij} V_j + I_i$. I_i represents the input to a cell coming from sensory input (e.g. light falling on a light-sensitive neuron, resulting in a transduction from external light to an internal neural signal), external electrodes placed by a neurobiologist, or other neurons not being described at the moment.

(d) Represent the effect of the transmembrane resistance and capacitance of cell i by constants ρ_i and C_i. These approximations result in the following equations of motion for the state of the system of neurons:

$$C_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_j T_{ij} V_j - \frac{u_i}{R_i} + I_i,$$

$$u_i = g_i^{-1}(V_i).$$

These equations (Sejnowskii 1981, Hopfield 1984), describe a classical neurodynamics—classical in the sense of classical physics. They omit the effect of the quantization of the action potentials, the "pulse coding" by which many neurons communicate. They are also classical in treating the communication speed as infinite, neglecting the time necessary to propagate an action potential.

These are equations both of extreme complexity and also of drastic oversimplification. The complexities of these equations can be demonstrated by the properties of simple circuits of these elements. For example, two neurons can be connected together to make a bistable flip–flop circuit. Three or five neurons with negative couplings can be connected in a ring, forming a ring oscillator or clock. By choosing the correct offsets in the gain functions and the correct T_{ij}, a neuron can be set up so that its output is very nearly zero unless at least two other neurons which are driving it have outputs near one. This makes a logical circuit which performs the "and" operation. Continuing in this fashion, we can construct all the elements necessary to build the sequential logical operations and memory necessary for a universal computer. Any digital computer which can be built can be duplicated (as far as its logical operation is concerned) by appropriate choices of the set of g, T, I, R and C. The complexity of the behaviors which can be shown by these equations is beyond description, for they comprise all possible computations and computers.

In real brains, the synapses themselves change with time due to the activity of the network. Learning is believed to chiefly involve the modification of the synaptic strengths T_{ij}. A complete set of neurodynamical equations must also describe the rate of change of the synaptic strengths, an equation of motion for the T_{ij} themselves. Considerable experimental work and modeling of this procedure has been done, but cannot be reviewed here. The dynamic equations described above might describe the response of a brain to a new situation in the light of what the brain already knows, but does not include learning new facts or relations.

These equations are, at the same time, a gross oversimplification of the realities known from experimental neurophysiology and neuroanatomy. Individual action potentials are known to be important in some parts of neurocomputation, especially in early sensory processing in the auditory system and in the visual analysis of motion. Cells in invertebrates often display bursting behavior even when driven by a constant input, generating a train of equally spaced spikes followed by a period of no action potentials. The description of inputs from axons to dendrites is an idealization of motor neurons, while neurons in brains often show connections from dendrites of one cell to dendrites of another cell, and some connections where three processes (e.g. two dendrites and an axon) make mutual close contact. Signals in the dendritic arborization do not simply add—for there are rather more complicated interactions between inputs, and connections made close to the cell body may "veto" the inputs coming from more distant parts of that dendritic arbor by forming a shunting impedance path. Some general neurotransmitters have relatively longer-lasting or modulatory effects on neurons, and several time scales of integration are present.

The above list of emissions and oversimplifications could be easily extended to great length, making many neurobiologists feel that such a set of equations must be inadequate. But for looking at collective behaviors, it is essential to model a neuron as simply as possible, realizing that if any interesting computational properties occur in a simple system, the real system will certainly have these properties, and perhaps more. An excessively complex model of the capabilities of a cell would preclude the possibility of understanding the collective aspects of the system. So we are led to ask whether these very simple networks will do in a natural, elegant, and collective fashion some of the kinds of computation which biology seems to do uniquely well.

In conventional computer hardware, memory is located by address. Information is stored somewhere, at a particular address, and can be retrieved by an instruction to read the contents at that address. This is most obvious in the case of a magnetic recording medium, where the information is stored as magnetization at a physical location on the recording disc, and the address is a description of that location. It is equally true of the chips used in a simple semiconductor memory like a Read Only Memory (ROM), where there are two separate sets of signal wires to the chip. One is a set of address leads, into which information is sent about the locations to be read. The others are information leads, on which the information from the desired location is sent out. If we were in possession of part of the information stored at some address and wanted the rest of that information, but did not know the address, there would be no alternative but to examine each memory in turn. Partial information is not a key to more information in ordinary machine memory hardware.

Memory in animals seems to be of an entirely different nature. Each of us knows many individuals, and a particular acquaintance can be described by a large number of characteristics (eye color, height, face shape, accent, vocation, spouse, children, shared experiences, education, nationality, political views...). All these characteristics belonging to a single individual are somehow associated or linked in a memory. The entire set of features connected to one individual can be retrieved from an incomplete and partially incorrect set of information about that individual. "The short jovial tennis-playing physicist from MOT" is sufficient information to identify Prof. Feshbach, in spite of the fact that MIT has been misspelled and has hundreds of physicists and many tennis players. In this retrieval we have no notion of address. Any substantial subset of the information can be used to access the memory. Psychologists refer to such linkages as associative memory. Associative memory is a general and important feature of memory in all animals which learn, and the classical association paradigms of Pavlovian condition, often thought of in terms of higher animals, have been demonstrated by Gelperin (1983) and Sahley et al. (1981) on animals as simple as a garden slug (*Limax maximus*).

Although associative memories seem natural to biology, they are foreign to conventional computer hardware. They turn out to be natural for collective biological hardware, because the neurodynamics we have given describes a dynamic system, not a logical calculation, and the idea of associative memory is a natural construct in a dynamical system.

Consider an "information space" about people, with many dimensions. Each axis could be labeled with a characteristic such as height, weight, name of spouse, or age. In such a space, each person you know is represented by a point, whose location describes all characteristics of the individual in question. You know very few of all possible people, so the set of points which represent your memories is very sparse in this space. Suppose now that you are given partial information about some

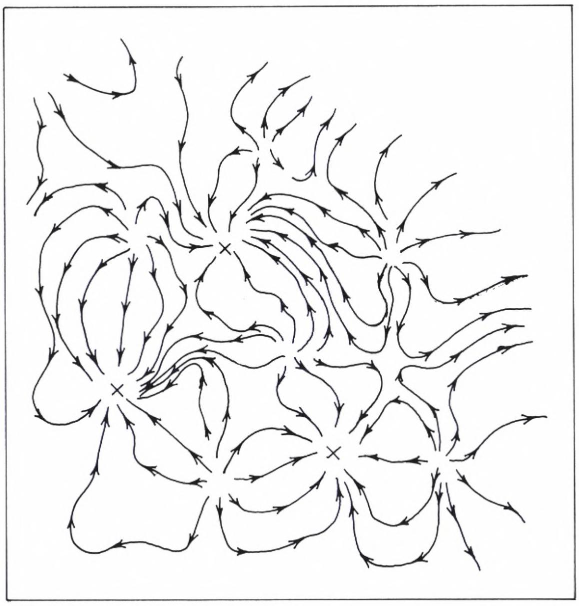

Fig. 4. A simple motion map of the change of the state of a dynamic system, characterized by a set of stable states. The stable states represent memories (in the case of associative memory), or more generally, possible answers in a computation.

particular friend. This information represents an approximate location in the information space. Presumably this approximate location is closer to the position of this friend in the information space than to the positions of other individuals. An associative memory functions by using this closeness to locate or move to the complete information. If a dynamical system is described by a flow characterized by stable points, as in fig. 4, the initial information could start the dynamical system at a location close to one of the stored memories. If the state-space motion is such that the motion goes to the most nearby stable point, the dynamical system would be functioning as an associative memory (Hopfield 1982). Many dynamical systems display state-space motions with multiple stable states. Any such system is a candidate for an associative memory if in addition the stable states can be placed wherever they are desired. Thus associative memory is not an esoteric property of computational systems. It is natural behavior of many simple dynamical systems.

In the case of the neuronal dynamics model, when the matrix of connections is symmetric, Hopfield (1984) showed that the flow is simple, and has no cyclic or strange attractors. Somewhat less stringent conditions will suffice. The symmetry of the connections is natural to the associative memory problem because the idea of simple association is itself symmetric. The essential step in this proof is the construction of a Lyapunov function, the "computational energy", which decreases monotonically during the neurodynamical motion. This energy is minimized (at least locally) by the state-space motion. Of course, the system of neurons is itself unaware that it is doing anything so grand as minimizing some global function—it merely obeys its own neurodynamical equation. This computational energy is simply our way of describing a global or collective property which enables us to understand the nature of the behavior of a large and complex system.

The idea of restoration is intrinsic to the notion of computation in a world of noise and error. The entire computation of this neural dynamics is intrinsically restorative, for the motion of the system is downhill on an "energy" terrain. Minor noise, or perturbations in the shape of the energy surface will not change the valley to which the state-space trajectory is descending, except when it starts at or traverses a logically ambiguous location such as a saddle point in the terrain.

Associative memories based on these principles of collective behavior and neurodynamics have been simulated on digital computers. The T_{ij} construction is itself amazingly simple. The T_{ij} can be built up by adding one memory at a time, as biology would require. The particular connections between i and j do not require any global information about the memory to adjust themselves, but only make use of information locally available from the "experience" of the network being held in a state to be remembered by the external input to it. Hebb (1949) postulated biological learning rules for synaptic connections which are of the nature that this associative memory would require.

Because the neuronal equations also correspond to elementary (but highly connected) electrical circuits of novel form, they also suggest the utility of making devices of such a structure. John Lambe first constructed such circuits, and there is a growing effort to investigate such devices in both electrical and optical formats (Lambe et al. 1985, Sivilotti et al. 1985, Jackel et al. 1986, Psaltis and Farhat 1985).

The secret to the simple flow pattern to stable states is the "computational energy" function which is always being decreased by the equations of motion. The operation of the collective circuit decision could be anthropomorphically described as the attempt of the circuit to find a minimum of this function. Computational problems can often be stated as minimization or optimization problems. When a mapping can be found which creates a correspondence between an energy function and the quantity which is to be optimized, then this kind of network may be effective in solving the problem. Many perceptual and reasoning problems from biology are of this minimization nature. What is the best route home, or the best pathway to move your hand in order to reach out and grasp the sandwich? Human speech is normally a continuous sound stream. Given such a continuous stream, where is the best place to interpret the breaks between words? (If you think this is not a problem, try to remember the first time you heard French spoken if your native language is not French). In the scene at which you are now looking, what is the best way of associating features together so that they will make reasonable objects? There is an immense field of optimization problems which biology must solve, and the fact that the neural networks are spontaneously capable of making a collective optimization decision suggests the way that a nervous system may go about doing such computations.

In an effort to understand the power of such networks in difficult computations, Hopfield and Tank (1985) took a non-biological optimization problem which is computationally very difficult (NP-complete) and examined the kinds of solutions which a neural network that we designed would generate. The object was not to see how a brain might solve this problem, but rather to study the computing power which an assembly of 100–1000 richly interconnected neurons could possess. The Traveling Salesman Problem investigated is defined as follows. Given a particular set of cities—Copenhagen, Stockholm, Oslo, Bergen, Malmö, Aarhus,..., in what order should one fly between the cities such that each city is visited once, one returns to the starting point, and the total distance flown is as short as possible? This problem is connected to biological computation of problems like word boundaries because in each case, there is a combinatorially larger number of possibilities which must be considered.

The simulation showed that a simple network of 900 "neurons" could find a good solution to a problem on 30 cities in a single convergence. The settling time for a set of neurons involved in such a computation would be 0.1–0.2 seconds. The network did not find the best solution. But of the more than 10^{30} possible solutions, the network found one of the best 10^7, rejecting poor solutions by a factor of 10^{23}. This is an immense amount of computation for such a small set of elements. A microcomputer, with an intrinsic speed about 10^5 times faster and containing 100,000 times more transistors, would achieve a comparable result in a comparable time. The fact that the neuronal network found a very good solution but not the best one is typical of biological computation. As was evident from the earlier molecular examples, biology does not insist on perfect solutions to computational problems.

The immense computing power of this small network comes from features which are explicitly those of neurobiology. First, large connectivity between neurons (or in engineering, amplifiers) is required. Brains have connectivities on the scale of 1000, while typical transistors in integrated circuits get inputs from only a few other transistors, and send outputs to only a few others. Second, the system operates in an analog mode, using the smooth, graded and nonlinear response of a neuron to its input. By contrast, the conventional digital machine emphasizes logical operations, ("on" or "off", 1 or 0), and suppresses as much as possible the fact that the fundamental hardware is itself an analog circuit. Third, a single collective decision is smoothly made by many neurons at once, rather than a sequence of minute logical decisions. The operation is as a dynamical physical system rather than as a logical one.

5. Beyond neurodynamics: complementarity

I turn finally to topics which I do not usually write about. They are subjects about which Niels Bohr would have asked, and on this centenary I am obliged to try to give answers—such as they are—to some queries which clearly come forth from the legacy of Bohr's essays.

Bohr's view of complementarity extended rather more broadly than the narrow confines of the wave-particle duality or the quantum-mechanical uncertainty relations. His view came from the understanding of quantum mechanics and its relation to classical systems, but did not fundamentally rely on quantum ideas. He asked whether there were other contexts in which the conflicting demands of observer and of normal unperturbed system behavior produced an essential paradox, and inability to know all about a particular system in the same fashion that we cannot ask simultaneously wave and particles questions. Such limitations on complete knowledge need not even necessarily arise from observer-experiment problems, but might arise in other situations. Bohr particularly raised this question with respect to the operation of biology, where he remarked (Bohr 1958, 1963):

> "The basis for the complementary mode of description in biology is not connected with the problems of controlling the interaction between object and measuring tool, already taken into account in chemical kinetics, but with the practically inexhaustible complexity of the organism."

This clearly expresses his view that issues of complementarity and the limits of knowledge in biology are of the nature that I described ror turbulence, and not fundamentally involved with quantum mechanics in a profound way. Bohr's (1958) view toward complementarity in biology evolved during his lifetime, as biology and the understanding of quantum physics also rapidly evolved. His later essays themselves seem somewhat complementary. The next paragraph attempts to combine the views expressed chiefly in several assays written in the last five years of his life (Bohr 1963).

Bohr viewed the action of the human brain as perhaps the most likely case of the occurrence of a complementarity limitation to the precision of knowledge of biological behavior. There appeared to Bohr to be intrinsic complementary aspects to the detailed study of the anatomy of a particular brain and the thought processes in that brain. Since the thought processes themselves change the microanatomy of the brain at the molecular level—how else could we remember our thoughts—observation of the function of the brain in a whole animal inevitably changes the anatomy of that brain. But an observation of the anatomy of the brain in sufficient detail to predict the functioning of that brain is so destructive as to preclude making the observations which would test the predictions. In this description, microanatomy and psychology become complementary views of a brain, to be united by a description which like the Schrödinger equation and its interpretation, should show why this limitation must forever remain.

What does a generalization of the classical neurodynamic equations in order to take into account thermal noise, and more realistic anatomy and physiology, suggest about such questions? If, with such generalizations, an adequate number of precise measurements were made (including studies of the properties of the modification processes) we could with sufficiently large computers predict the state of the neuronal system for a short time, and thus be able to simulate its behavior. Adequately precise information about a mammalian brain for such purposes might well require 10^{20} bits of information about a single brain, clearly a hopeless task. It might be too complex to ever simulate. This might be a fundamental limitation to knowledge of such systems. But while Bohr considered this aspect, it is not the complementarity which he viewed as most important.

His essential complementarity problem is whether even a complete set of electrophysiological measurements for simulation of neurons in complete detail can be connected to particular higher mental processes which are normally described in macroscopic or cognitive terms. There must be something observably different in the microscopic behavior of my neurons which occurs between the time at which I do not understand a mathematical theorem and the time slightly later when I do understand it. That difference can be studied by physiologists. That difference can also be pursued by psychologists asking questions about the nature of my understanding. But even if we can observe all the physiological microdetails, these microdetails may not necessarily lead to any knowledge of what "to understand the theorem" means. There may be no way to deduce the information available to the psychologist from the data obtained by the physiologist. In such a case, there would necessarily be parallel or complementary descriptions of mental phenomena. The analog in physics might be a phase transition so subtle that even though we can see macroscopically that the phase transition has occurred, the phase itself is so complex that it cannot be described microscopically by a list of symbols or words short enough to be read during a lifetime. In that case, a complete Hamiltonian, the equivalent to the equations of neural dynamics, would still need to be augmented by the complementary description of macroscopic physics in order to describe observable phenomena.

The complementarity issues just discussed have to do with the issue of computational complexity and complexity measures, a topic which has barely begun to enter physics. There is, in addition, one peculiar aspect of the mathematics of collective neuronal computation which has the feeling of complementarity. This involves the question as to whether we will be able to understand the logical process by which the brain computes.

Computation is, in the usual description, a logical operation. But when we look at the way problems like the one of the traveling salesman are solved on a neural network (Hopfield and Tank, 1985); it does not turn out to be a logical procedure. Figure 5a shows a representation of the state of the neural network at an intermediate time during the settling to a solution to such a problem. Each point in the array of squares represents the output of a particular neuron. When the square is as large as the one at the lower right, that neuron has an output of 1. When the system has found a good solution to the problem, each of the outputs will be either 1 or 0. Thus, in the final state of the system, the output will appear as in fig. 5b, with one 1 in each row and each column.

In this computational process, each of the neurons can be thought of as standing for a proposition. An output of 1 means that the proposition is considered to be true, and an output of 0 means that the proposition is not true. In fig. 5, the propositions (which are of the general form "the salesman should go to city D as the 7th city in the tour") are laid out in such a fashion that each proposition contradicts the other propositions in each row and column. An appropriate final state must have no more than one 1 in each row and each column.

What are we then to make of the intermediate state of the form shown in fig. 5a? At this time, the proposition that city D should be in position 8 has a weight of 0.25 while at the same time, the contradictory propositions that city D should be in position 7 and that city C should be in position 8 also have weights of about 0.25. The system might be characterized as simultaneously having partial belief in several mutually contradictory propositions. There is no obvious logical characterization of this state of partial belief in contradictory propositions. We cannot view this partial

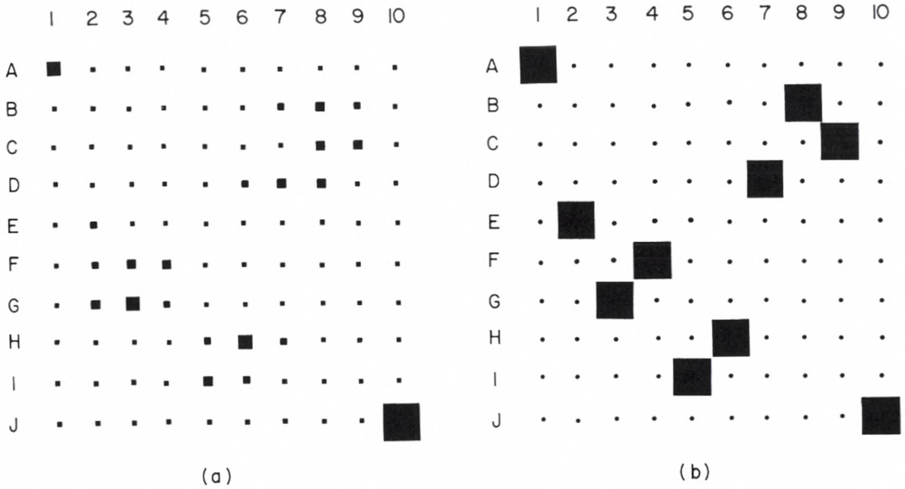

Fig. 5. The linear dimension of each square represents the activity of that neuron. The neuron in square A5 represents the proposition that city A should be visited 5th. (a) An intermediate state of the calculation. City J has been arbitrarily placed in location 10, so that the output of the corresponding neuron is of maximum size. (b) The final stable state of the network, represent the path AEGFIHDBCJ (and return to A).

belief in terms of probabilities, for the entire system is deterministic. There is a logical space in which a proposition is true or false, a space which for N propositions can be described as the vertices on an N-dimensional cube. The neural network computes by moving on a trajectory in the *interior* of this cube, while only the corners have logical meaning. Thus an insistence on finding correct logic during the decision process of the network destroys the computation as done by the network. The ability to use the non-logical interior of this space seems to be an important part of the power of these collective "neuronal" networks.

Is this a sensible way to go—to try to understand a brain by giving up simple logic and moving toward analog, analytic, and collective behavior? John von Neumann (1948) thought about the relation of the brain to the digital computers which he had been so influential in helping develop, and concluded:

> "All this will lead to theories (of computation) which are much less rigidly of an all-or-none nature... than formal logic. They will be of a much less combinatorial, and much more analytical, character."

He continued this theme in later lectures (von Neumann 1952), but did not find what he regarded as a satisfactory solution to the problem of how a brain computes.

6. *Summary*

We have learned a great deal about the chemical and electrophysiological basis of brain since Bohr wrote the cited essays. The way in which important aspects of brain function and neural computation could be of a collective nature is beginning to emerge. We see some simplified governing equations which are relatable to collective behaviors. These collective behaviors even in rudimentary form seem to show how associative memory and combinatorially difficult decision-making of the kind necessary in visual or auditory perception can be carried out with great effectiveness by a modest collection of neurons.

It is not clear where the limits of such pursuits will lie. Will we be able to explain some of the phenomena of cognitive psychology on the basis of details of microscopic neuronal behavior? Or will the macroscopics never be microscopically understandable? If Bohr's argument that a form of complementarity may be involved in brain turns out to be correct, the mathematical explanation for it will lie in the nature of computational complexity and the meaning of computation. Will there be a mathematical complementarity in brain? My own guess is that the complications of the system are not intrinsic, and that we will be able eventually to go astonishingly far toward understanding the language of psychology from a microscopic viewpoint. While weather prediction on the basis of molecular modeling is impossible, we do understand why there is weather and what storms are on the basis of an atomic approach to atmospheric phenomena. In this same sense, new approaches to theory in neurobiology should lead us at least to an understanding of the *existence* of psychological constructs on a microscopic basis, and potentially lead much further.

Postscript

The editors have expressed the hope that someone might read this book in connection to the second Bohr Centenary. The science described here will, 100 years from now, be amusingly naive. How science is done—who influenced whom, and how—may be of greater interest. I close on a personal note, which would otherwise be lost. I did not know Niels Bohr. John Wheeler was the strongest influence in my taking a position at Princeton, and Max Delbrück strongly influenced the Caltech scene and my interest in moving there. Both of these men were in turn involved with Bohr and Copenhagen. I have deep admiration for the imaginative scientific spirit of each of them. Reading the essays of Niels Bohr and listening to this symposium, I became increasingly aware that many elements I find admirable in these scientists are related to the spirit in which Bohr worked. But whether this similarity is due to influence or due to mutual affinity is not obvious. Francis O. Schmitt and the Neurosciences Research Program did much to stimulate my interest in neurobiology. The significance and liveliness of the many discussions about biological and network computation with David W. Tank is gratefully acknowledged.

This research was supported in part by NSF grant PCM-8406049.

References

Bohr, N., 1958, Atomic Physics and Human Knowledge (Wiley, New York).

Bohr, N., 1963, Essays (1958–1962) on Atomic Physics and Human Knowledge (Interscience, New York) p. 21.

Gelperin, A., 1983, in: Neuroethology and Behavioral Physiology, eds F. Huber and H. Markl (Springer-Verlag, Berlin) p. 189.

Hebb, D.O., 1949, The Organization of Behavior (Wiley, New York).

Hopfield, J.J., 1974, Proc. Natl. Acad. Sci. USA **71**, 4135.

Hopfield, J.J., 1982, Proc. Natl. Acad. Sci. USA **79**, 2554.

Hopfield, J.J., 1984, Proc. Natl. Acad. Sci. USA **81**, 3088.

Hopfield, J.J., and D.W. Tank, 1985, Biol. Cybern. **52**, 141.

Hopfield, J.J., and T. Yamane, 1980, in: Ribosomes, eds G. Chamblis, G.R. Craven, J. Davies, K. Davies, L. Kahan and M. Nomura (University Park Press, Baltimore) p. 585.

Jackel, L.D., R.E. Howard, H.P. Graf, B. Straughn and J.S. Denker, 1986, J. Vac. Sci. Tech. **B4**, 61–63.

Kandel, E.R., and J.H. Schwartz, 1981, Principles of Neuroscience (Elsevier, New York).

Lambe, J., A. Moopenn and A.P. Thakoor, 1985, Jet Propulsion Lab Publication, **November**, 85.

Little, W.A., 1974, Math. Biosci. **19**, 101.

Little, W.A., and G.L. Shaw, 1978, Math. Biosci. **39**, 281.

Mead, C.A., and L. Conway, 1980, Introduction to VLSI Systems (Addison-Wesley, Menlo Park, CA) p. 335.

Muzyczka, N., R.L. Poland and M.J. Bessman, 1972, J. Biol. Chem. **247**, 7116.

Pauling, L., 1957, in: Festschrift Arthur Stoll (Birkhäuser, Basel) p. 597.

Psaltis, D., and N. Farhat, 1985, Opt. Lett. **10**, 98.

Sahley, C., A. Gelperin and J.W. Rudy, 1981, Proc. Nat. Acad. Sci. USA **78**, 640.

Sejnowskii, T.J., 1981, in: Parallel Models of Associative Memory, eds G.E. Hinton and J.A. Anderson (Lawrence Erlbaum Associates, Hillside, NJ) p. 189.

Sivilotti, M., M.R. Emmerling and C.A. Mead, 1985, in: Conf. on Very Large Scale Integration, ed. H. Fuchs (Computer Science Press, Rockville, MD) p. 329.

von Neumann, J., 1948, in: Collected Works, Vol. 5, ed. A.H. Taub (Pergamon Press, New York, published in 1963) p. 304.

von Neuman, J., 1952, ibid, p. 354.

Watson, J.D., 1976, Molecular Biology of the Gene (Benjamin, Menlo Park, CA).

Yates, J.L., 1979, J. Biol. Chem. **254**, 11550.

Discussion, session chairman N.K. Jerne

Zimanyi: You have used a symmetric connectivity matrix T_{ij}. But in a neuronal network we have about as many inhibitory neurons as excitatory ones. The part of the connectivity matrix representing these types of neurons should be antisymmetric instead of symmetric. Howe can one overcome that problem?

Hopfield: Even networks which do not in detail look symmetric may be constructed equivalently to symmetric networks and possess the stability properties of symmetric networks. The symmetry assumed here serves to ensure viability of the mathematics. Asymmetric systems open a Pandora's box of totally uncontrollable mathematical behavior, but have a much richer computational structure.

von Weizsäcker: I have no quarrel with reductionism. Bohr used to explain complementarity by the difficulty of language. Can you explain the complementarity structure in language by your network mode? An example: My uncle, Victor Weizsäcker, a medical psychologist, who reminded me in many respects of Niels Bohr, once said: "When I ask you: 'what are you thinking now?', and you start to answer, you are already lying."

Hopfield: You are absolutely right in referring to Bohr's pointing out the language as an intrinsically complementary system; complementarity must be left out from microscopics when deriving collective results in a reductionist attitude.

Anderson: Going back to one of your first sentences about the complications in biology arising not from the size of Planck's constant, I should like to point to the size of Avogadro's number, a biological constant in itself, as it represents a piece of matter approximately our size. This may be the amount of matter necessary to produce the required complexity.

Berry: Let me return to your point concerning how thinking causes microscopic changes in physiology and your skepticism whether microphysiology could tell us when a person has solved a problem. Strictly, I should say "when the person *thinks* the problem is solved." I suspect you don't really believe that, because you and everybody in this room knows how solving a problem makes you feel better, quite literally. A problem must be awfully easy and awfully trivial to generate no such feeling at its solution. Consequently, do you not think the microphysiology should display some chemical signal when the person thinks he or she has found the solution?

Hopfield: Emotional states are relatively easy to recognize by chemical signals. I think it could probably be done now. But you would still not be able to tell what the person's understanding of the answer was or whether it was solved from your point of view.