ON THE ORIGIN OF MACROMOLECULAR SEQUENCES

HOWARD H. PATTEE

From the Biophysics Laboratory, Stanford University, Stanford

ABSTRACT The origin of the degree and type of order found in biological macromolecules is not adequately explained solely as an accumulation of genetic restrictions acquired through natural selection from otherwise unrestricted primeval sequences capable of self-replication, since the biological process of replication is itself dependent on the pre-existence of such order, and since the number of sequences that could have ever been tested by selection on the earth is an insignificant fraction of the number of unrestricted sequences which would be possible. Therefore the hypothesis is considered that replication and selection began from well ordered sequences, rather than random sequences. It is shown how the Turing concept of computation in fed-back, discrete-state automata can lead to the generation of order without pre-existing instructions, and how this computation can result in self-repeating, random-like, but well ordered sequences of great length. Macromolecular models of such computers are suggested on the basis of mechanisms proposed for the growth of eutactic polymers. Such self-replicating, mutable sequences may then evolve genetic control which is sufficient to accommodate the information accumulated by natural selection. The structure and function of enzymes and structural proteins is related to this model, and statistical evidence from known amino acid sequences is shown to be consistent with some degree of non-genetic ordering.

INTRODUCTION

A large body of evidence has accumulated which leads to the conclusion that some form of genetic information is transferred from parent to offspring by nucleic acids. According to the Sequence Hypothesis, this genetic information is stored and transmitted by the linear order of the bases in the nucleic acid molecules, and is expressed by some unequivocal process in the linear sequences of amino acid residues in proteins. Furthermore, it is commonly assumed that this information is transmitted in one direction only, from the base sequences to the amino acid sequences, where it remains; and that the three-dimensional configurations of proteins, as well as their aggregations and specific interactions in a given local environment, are a consequence of the particular linear order of their constituent amino acids (Crick, 1958).

Since there are about twenty common amino acids in protein and about four common bases in nucleic acids, there has arisen the "coding problem" which is generally approached as a search for a time-independent dictionary by which the genetic information in the nucleotide base symbols of nucleic acids may be formally translated into corresponding amino acid symbols of proteins. (Gamow, Rich, and Yčas, 1956; Crick, Griffith, and Orgel, 1957; Golomb, Welch, and Delbrück, 1958; Yčas, 1958; Woese, 1961). A great amount of experimental effort is also directed toward the search for reasonable biochemical mechanisms whereby a linear base sequence code in the nucleic acids may lead to the synthesis of specific linear sequences of amino acid residues in proteins. There is evidence that the parental genetic text is itself replicated by a template process involving a simple rule (Watson-Crick model).

WHAT IS THE QUESTION?

We find in none of the present theories of replication and protein synthesis any interpretation of the origin of the genetic text which is being replicated, translated, and expressed in functional proteins, nor do they lead to any understanding of the relation between particular linear sequences or distributions of subunits in nucleic acids and proteins, and the specific structural and functional properties which are assumed to result entirely from these linear sequences. According to present theories of replication, coding, and protein synthesis, all of the information necessary to order the linear sequences of amino acids in proteins is provided by the genetic sequences (Brenner, 1959). Furthermore, the random-appearing distributions of amino acids in non-structural proteins has been widely interpreted as evidence against any restrictions or expectations of sequence order independent of genetic instructions (Gamow, *et al.,* 1956; Crick, 1958). For example, Brenner (1957) has concluded, "As far as the coding problem is concerned, it now appears that all amino acid sequences are likely to be found and that it will not be possible to effect a 'decoding' by discovering restrictions in sequences." The genetic sequences are also commonly assumed to be unrestricted except as a consequence of similar pre-existing order in the parental genetic material. In fact biological "self-replication" has been defined by Pontecorvo (1958) to mean the process by which a collection of subunits ". . . which could arrange themselves in any one of all the possible sequences, becomes arranged into only one sequence *because of the pre-existence of such a unique sequence.*" We may say then, that most present theories of replication, coding, and protein synthesis state or imply the absence of rules, restrictions, or expectations of any order in the sequences of subunits in biological macromolecules, except in so far as such order pre-exists in genetic sequences.

According to the common view, the origin of all this genetic order which is believed to exist in linear sequences of bases in nucleic acids is attributed entirely to the independent evolutionary process of natural selection from the random or arbitrary errors in memory or replication of pre-existing sequences as well as to the random mixing of sequences in the process of reproduction. The primeval origin of genetic order is usually considered as a chance aggregation of subunits which for unspecified reasons is assumed to be capable of self-replication. It should be clear that the theory of natural selection implies only an eventual choice from a population of completed, functionally distinguishable sequences, capable of replication, and does not itself impose immanent rules or preferences on any particular order during actual synthesis.

If we accept the evidence that there was a time when there existed on the earth only a disordered reservoir of molecular subunits with neither genetic instructions nor macromolecular templates for ordering macromolecular sequences, we are therefore led to ask the following question: Can we explain the degree and type of order which is now observed in biological macromolecules without falling back at some essential step on a random or arbitrary choice of sequence or ordering process for which no further analysis is logically or empirically possible?

CAN NATURAL SELECTION ALONE ACCOUNT FOR THIS ORDER?

There is no question that the process of multiplication, variation and selection generates genetic information which now results in the choice of certain sequences from some enormous number of alternative sequences (Muller, 1958). Kimura (1961) has estimated the information accumulated by the process of natural selection as 10^8 bits, basing his calculation on values of gene substitution rates and the number of generations since the Cambrian epoch. If we assume an average generation time of the order of one hour over a period of 10^9 years, which would be appropriate for evolution of cells the size of bacteria, we would obtain a value of about 10^{12} bits. These estimates agree as well as can be expected with the limits of the information content of cells based on thermodynamic grounds (Linschitz, 1953), on the complexity of cell structure (Dancoff and Quastler, 1953), and on the information capacity of DNA (Muller, 1958).

Therefore, according to these coarse estimates, enough genetic information could, in principle, accumulate from the process of selection to define the order which is now found in a typical cell. However, it must be borne in mind that this represents the minimum amount of information necessary to *define* a certain state. In general it is not a sufficient amount of information to *produce* this state under arbitrary initial conditions. For example, the number of decisions or the amount of information necessary to locate a specified word in a dictionary depends on whether the words are printed in alphabetical order or in random order. Similarly, our ability to define the sequence in a nucleic acid molecule or protein with a given number of decisions, does not necessarily mean that we may expect to produce these macromolecules by a random search from a disordered environment with the same number of decisions. In general, only when a set of N equiprobable alternatives can be ordered according to some *pre-existing* rule, is the *minimum* amount of information necessary to select one of these alternatives equal to $\log_2 N$ bits. However, if no process can be devised to compare more than one alternative at a time with the one to be selected, then the selection process may require as much as N bits. The latter measure of information would appear to be in closer correspondence to that found in the process of natural selection.

In any case, there is also a very practical limit to the number of macromolecular sequences which can be tested by selection. Even if we make some unrealistically generous estimates of the geologic space and time available for the systematic production of different protein-like macromolecular species, it is difficult to conceive of more than about 10^{60} such macromolecules ever existing on the earth.[1] For even one small macromolecule the size of TMV protein there are about 10^{200} conceivable sequences, so that no matter what process of selection occurs, there will remain about 10^{200} such sequences which never could have existed on the earth, and consequently no significant fraction of the possibilities could ever have been selectively tested. In other words, if we assume that all conceivable linear orders of typical biological macromolecules are equally probable except for genetic restriction, we have no physically realizable selective process to reduce effectively the enormous number of such orders so that the occurrence of one particular sequence becomes an event with any reasonable probability. Therefore the assumption of non-genetic equiprobability of sequences must lead at some stage to a random or arbitrary choice between possible sequences which does not yield to further analysis. Finally, it should be clear that natural selection can effectively accumulate information only in self-replicating genetic systems which themselves are known to exist at the present time only in environments with a very high degree of pre-existing macromolecular order.

Therefore the process of random variation and natural selection does not adequately solve the problem which we pose; namely, to formulate an hypothesis for the origin of the degree and type of replicating order now found in biological macromolecules which does not make essential use of such improbable or arbitrary events that no reasonable experimental test can be devised to verify or disprove the hypothesis.

WHAT ARE THE CONDITIONS OF THE PROBLEM?

We shall begin by rejecting the assumption that in the absence of genetic control all sequences are equally probable chance events, and we shall consider the hy-

[1] We have taken the age of the earth as 10^{17} seconds, the generation time as 1 second, and a volume of macromolecules equal to the entire volume of the surface of the earth one meter thick, or 5×10^{23} cm^3, with 10^{19} macromolecules per cm^3.

pothesis that evolution by natural selection was not primarily a source of genetic information which produces order in otherwise random sequences, but rather that natural selection leads to increased genetic control, and hence random variability, in otherwise highly restricted sequences. Our first problem, therefore, is to provide a plausible mechanism for the natural occurrence of mutable, replicating sequences on which some process of selection can be expected to operate. We shall assume that there existed at one time on the surface of the earth a large reservoir of different types of distinguishable molecular subunits, such as nucleotides or amino acids, which are capable of forming similar pairs of strong chemical bonds so as to produce chain macromolecules. These molecular subunits are presumed to have arisen from elementary components through the action of non-specific energy sources, such as heat, ultraviolet radiation, or electric discharge as demonstrated to occur by Miller (1955) and others. We also assume that the necessary thermal or electromagnetic energy was available to allow the formation of bonds between these molecular subunits, but that these energy sources were not specific with regard to the ordering of these subunits in chains.[2] We shall not assume the spontaneous occurrence of the special biological property of self-replication as it is usually applied to the propagation of a fixed amount of pre-existing order by a process of copying, nor do we assume that sequence information must be transmitted from one type of macromolecule to another by a fixed code. We also exclude from this discussion the occurrence of an incredible event, such as the spontaneous, chance appearance of a functional nucleic acid or an enzyme molecule or any macromolecular template from a disordered collection of subunits. We must, of course, admit the philosophical possibility that the origin of living systems as we know them was essentially dependent on a fortuitous, random event; but since random events are not subject to further scientific explanation, we have chosen as a matter of strategy to investigate the possibility that there exist natural conditions under which life arises nearly inevitably. Therefore we shall try to make use of only those interactions which may be expected to occur with measurable probability among molecules in non-living aggregations of matter. However, before considering what type of interactions may be expected to result in ordered collections of molecules we must define more explicitly what we shall understand by the concept of "order."

ORDERING AND COMPUTING

The concept of "order" is so fundamental in any context that no simple definition is entirely adequate. For this discussion we shall call a given linear array of distinguishable subunits "ordered" in so far as we can find independent rules or in-

[2] We recognize the possibility that a particular energy or state of polarization may have favored one or another photochemical reaction. However, we exclude the possibility that this energy discriminates strongly in its action according to the order of the reacting sequences.

structions for generating such an array. By "rules or instructions" we shall include all forms of copying from both complementary templates or coded sequences, but it is our primary intention to include also the more general idea of a set of recursive operations which allow us to determine the next step in an ordering process given the present "state" of the ordered array.

In other words, we wish to include in our definition of "ordering" not only the present ideas of template replication and coding, but also the more general concept of "computation" which we use here in the sense originated by Turing (1936) in his paper on computable numbers. Turing theory is useful because it provides us with the most elementary logical rules of ordering which are general enough to produce any order which we may define, given suitable instructions. Many equivalent statements of these basic rules are possible, but for our purposes we may describe them as follows:—

1. READ or identify a specified elementary symbol which we may call the *input* symbol.
2. COMPARE or associate this input symbol with a second specified symbol which we may call the *state* symbol.
3. WRITE or add a new symbol which is specified by a rule determined only by the input symbol and the state symbol.
4. CHANGE or specify a new state symbol according to a rule determined only by the input symbol and the previous state symbol.
5. REPEAT these five steps.

These steps are very simple, but they form a logical basis for all digital computation, and as interpreted by Turing (1956), McCulloch and Pitts (1953), and Newell and Simon (1959), they may also be an adequate basis for the simulation of human thought processes. By a "computer" we shall mean any system which is capable of executing these steps, no matter what its physical representation may be.

Von Neumann (1951) has described a formal model of a finite-state automaton based on these operations which may be said to replicate itself in a very general sense. However, such a model must be of a certain "critical size" and would initially require an ordered linear array of external "genetic" instructions of considerable length to initiate its replication (Kemeny, 1955). The possible origin of such instructions was not a part of the problem which he posed. Other models of self-replicating systems (Penrose, 1959; Jacobson, 1958; Morowitz, 1959) should serve to illustrate that living matter as we usually recognize it is characterized not so much by its ability to replicate, but by the large amount of order which is there to be replicated. We consider the fundamental problem to be the natural origin of a high degree of order rather than its genetic replication.

AUTONOMOUS COMPUTERS

One might intuitively suspect that a simple process of computing, as defined by

TABLE I

SEQUENCE GENERATED BY THE BALANCE MECHANISM OF FIG. 1

B	B	A	B	A	A	B	A	B	B	A	A	A	B	B	A	B	B	B	B	A	B	B	A	B	A	B	B	A	B	B	A	A	B	A	A	B
A	A	A	B	B	B	A	A	A	A	B	A	B	B	B	B	B	A	A	B	A	B	A	B	B	B	A	A	B	B	A	B	A	A	A	B	A
A	B	B	B	B	A	A	A	B	A	B	A	A	A	A	B	B	A	A	A	A	A	B	A	A	A	A	A	A	B	B	B	B	B	B	B	A
B	A	B	A	B	A	A	B	B	A	A	B	B	B	A	B	B	B	A	B	A	A	B	A	B	B	A	A	A	B	B	A	B	B	B	B	A
B	B	A	B	A	B	B	A	B	B	A	A	B	A	A	B	A	A	A	B	B	B	A	A	A	A	B	A	B	B	B	B	B	A	A	B	A
B	A	B	B	B	A	A	B	B	A	B	A	A	A	B	A	A	B	B	B	B	A	A	A	B	A	B	A	A	A	A	B	B	A	A	A	A
A	B	A	A	A	A	A	A	B	B	B	B	B	B	B	A	B	A	B	A	B	A	A	B	B	A	A	B	B	B	A	B	B	B	A	B	A

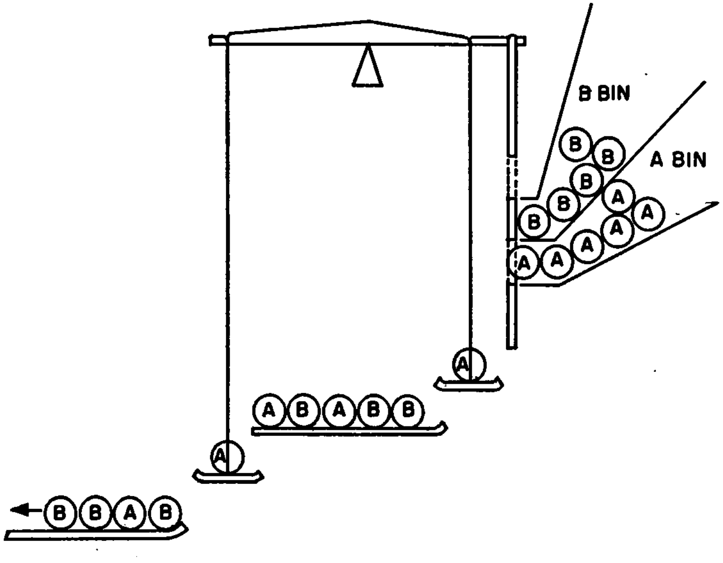

FIGURE 1 Mechanical model of an autonomous sequence computer which generates the sequence shown in Table I. The beam of the balance operates the gate so that when the pans are balanced an A ball enters, and when pans are unbalanced a B ball enters. A balls are heavier than B balls.

the five rules above, could not be expected to lead to a higher degree of order than is contained in the input symbols or instructions entering into the computing process. Indeed, it is just this suspicion which is largely responsible for our conceptual difficulty in explaining the origin of living matter with its peculiarly high degree of order from a primeval environment where very little order could have existed in macromolecules. This suspicion is justified by most of our common experience. However, there is a general type of recursive ordering process which may occur in computing configurations which operate according to simple rules but which do not contain specific input instructions. This process may be pictured as a logical feedback system which leads to a type of discrete-state oscillation where the output is an ordered, periodic sequence of symbols determined by the internal logical boundary conditions of the system and not on external instructions.

Before considering some possible molecular representations of such autonomous ordering systems, two simple examples of fed-back discrete-state computers may be instructive. Consider first the following schematic mechanical system (Fig. 1) in which an ordinary balance is fed through a gate with either an A (heavy) or B (light) ball depending on the state of balance at the moment. The entry of each new ball shifts the linear string of balls one ball to the left. The steps of computation are logically equivalent to those listed in the last section:

1. READ the weight in the left pan (A balls are heavier than B balls).
2. COMPARE weight in left pan with weight in right pan.
3. WRITE by adding type A ball if pans balance or type B ball if pans do not balance.
4. CHANGE state by shifting balls to the left.
5. REPEAT these five steps.

The last two steps are accomplished automatically by the WRITE operation. This mechanical feedback computer will produce a string of balls in the periodic sequence shown in Table I. As long as the bins are full, this mechanism will continue to repeat this sequence with a period of 127 balls. Neither the period nor the cyclic order taken over at least one period depends on the initial contents of the balance pans or troughs (if they are not initially all A balls) nor on which type of ball is heavier than the other.

The logical structure of such an autonomous computer is better illustrated by the equivalent representation shown in Fig. 2.

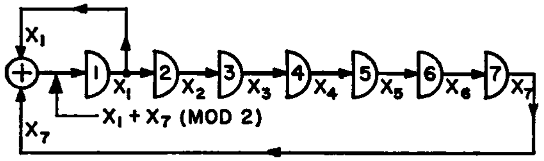

FIGURE 2 Schematic diagram of a binary feedback shift register which is logically equivalent to the mechanical sequence generator of Fig. 1, and which generates the same sequence as shown in Table I where A $\equiv$ 0 and B $\equiv$ 1.

The operation of this binary feedback shift register, as it is called, is logically the same as the balance model of Fig. 1, although we use different symbols, different state memory, and a different comparing and writing element. The computation of this configuration is defined as follows:—

1. READ the contents of register 7 (in this case $X_7 = 0$ or 1).
2. COMPARE the contents of register 7 with contents of register 1.
3. WRITE the result (in this case $X_1 + X_7$ modulo 2) in register 1.
4. CHANGE state by shifting previous register contents one register to the right.
5. REPEAT these five steps.

The successive contents of any register pass through the same sequence as is produced by the balance model when the symbols 0 and 1 are substituted for A and B respectively. Since the longest feedback loop in this diagram includes seven registers (or seven balls in the balance model), each of which may be in one of two states, the entire system can exist in no more than 2^7 distinguishable states. One of these states (all registers 0) is trivial, so that we may expect a period of at most $2^7 - 1 = 127$ symbols. These particular representations are therefore called *maximal period* autonomous sequence generators, and since one period contains all possible combinations of 2 elements taken 7 at a time, the cyclic order taken over one period is not dependent on the starting sequence, but only on the logical connections of the configuration.[8] Such systems may produce a large amount of order, considering the simplicity of the mechanism. For example, if the balance model of Fig. 1 could hold 101 balls in the trough between pans instead of only 5, and if the average mass of each ball was 1 mg, then the entire mass of the earth would be consumed making balls before one maximal period could be completed. Yet the entire computer could be built in a box of less than 1/10 cc in volume. We may say, therefore, that the fed-back, discrete-state automaton provides a logical system for generating self-repeating, random-like, but well ordered linear arrays with no external instruction and very little internal information. Of course all autonomous sequences are not maximal period sequences. In general, one particular autonomous configuration has many distinct periods depending on the number of its symbols, its rules, and the initial contents

[8] It may be shown that for a modular field of integers $J_p = (0, 1, 2, \ldots, p - 1)$ where p is a prime, and for linear operations, the maximal-period sequence will contain $p^n - 1$ integers, where n is the number of memory elements within the largest feedback loop. Furthermore, it may be shown that the number of different maximal-period sequences for modulus p, degree n configurations is

$$M_p(n) = \phi(p^n - 1)/n$$

where $\phi\ (k)$ is the Euler phi function defined as the number of integers less than k which are relatively prime to k. For example, if $p = 7$ and $n = 5$, a maximal period will consist of a sequence of $7^5 - 1 = 16{,}806$ integers, and depending only on the logical feedback chosen there can be produced 1120 different maximal periods of this length (see Elspas, 1959).

of its feedback loops. In these cases the state of the feedback loops represents the "genetic" instructions for a particular periodic sequence. An error in the addition of one subunit which subsequently enters a feedback loop may correspond to a mutation if it leads to a new sequence. However, the redundancy of such computed sequences is high, since many errors in subunit addition may occur without producing a mutation. In fact, it is the error-correcting ability of such sequences in communication systems which has been the main incentive for their study (*e. g.* see Hamming, 1950; Huffman, 1955).

Although the mathematical theory of linear autonomous sequential networks has been presented by Elspas (1959) and others using the theorems of Galois field theory (*e. g.* see Birkhoff and MacLane, 1947) the behavior of more general finite-state automata operating with arbitrary Boolian logical feedback or time-dependent logic structure is largely unknown. Much of the difficulty in analyzing such structures arises from the great generality of the Turing concept of computation, as well as the apparent complexity of the ordering produced by these systems. It is this characteristic ability of fed-back discrete-state automata to produce disproportionately complex order from simple rules which leads us to propose that these processes arising naturally at the macromolecular level may have contributed to the origin of the type of order which is typical of biological macromolecules.

CAN MOLECULES COMPUTE?

We know that a sufficiently well organized collection of molecules can certainly compute according to the rules we have stated, as for example in the balance model of Fig. 1, or for that matter in the brain. But these examples require in their computing mechanisms a high preexisting level of organization or ordering before they function as computers. The question must be stated more explicitly: Is there any natural configuration of simple molecules which occurs with a measurable probability from a disordered reservoir, leading to the properties of autonomous computing structures in the sense illustrated above?

Basically, a computer requires a set of conditional rules of order and a memory. We do not expect these characteristics from a state of matter which is as disordered as a liquid or as constrained as a three-dimensional crystal. The obvious place to look for mechanisms which produce linear order is in that state of matter where such order is found, namely, in chain molecules. A linear polymer is flexible and may fold back on itself in many ways, for example, by faking (Keller and O'Conner, 1958) or by helical coiling. If these folding interactions result in a determinative influence in the reaction with additional subunits, we have all the properties necessary for an autonomous computer. This mechanism has been proposed by Szwarc (1958) and Ham (1959) to explain the process of stereospecific polymerization (see Natta *et al.,* 1955). As an example of this type of interaction, Szwarc shows that the kinetics of the polymerization of amino acid

N-carboxyanhydrides as studied by Idelson and Blout (1958) may be explained in part by assuming that the addition of each subunit is influenced most strongly by the last-added subunit and the subunit one turn away in a helical configuration. This proposed growth process would be a macromolecular analogue of screw-dislocation growth in crystals (Frank, 1949) which has also been proposed on an even larger scale to account for the assembly of protein subunits in the tobacco mosaic virus rod (Sears, 1959). Commoner (1959) has also presented evidence that the ribonucleic acid and protein of tobacco mosaic virus grow synchronously, and suggested a hybrid, sequential growth process; however other evidence does not appear to support this model (see Gierer, 1960). The general idea of sequential interaction in growing chains has also been described by Crane (1950). However, none of these proposals has recognized the possibilities of sequence computation in such configurations.

As one of the simplest possible examples, consider a macromolecule growing in a helical configuration as shown in Fig. 3. We assume it is made up of two types of subunit, A and B. Whatever steric or reactive characteristics these subunits may have, it is reasonable to expect that the last subunit and the subunit one turn away will have the strongest influence on the selection of the next subunit to be added, since these two parts of the chain form the dislocation. If we pick an arbitrary rule that an A subunit is added when the dislocation is formed by similar subunits, and a B subunit is added when the dislocation is formed by dissimilar subunits, this macromolecule becomes logically isomorphic with the two other examples of autonomous computers which we have given, and consequently it will

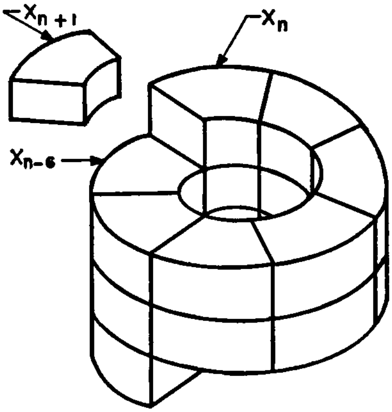

FIGURE 3 Schematic diagram of a macromolecular sequence computer. The type of subunit X_{n+1} to be added is determined by a rule depending on the type of subunit X_n and X_{n-6} forming the screw-dislocation. If there are only two types of subunit, A and B, and if an A subunit is added when $X_n = X_{n-6}$ and a B subunit is added when $X_n \neq X_{n-6}$, then this configuration is logically equivalent to the computers shown in Figs. 1 and 2, and it will produce the sequence shown in Table I.

order the subunits in the same linear sequence as the A and B balls were ordered by the balance model of Fig. 1 as shown in Table I.

There are, however, three important physical differences between this macromolecular model and the mechanical model. First of all, the large amount of information necessary to establish the order which must be built into the bins, troughs, and gate mechanism of the balance model before it can operate has been reduced in the macromolecular model to the information required to establish one turn in a helix of seven subunits. Therefore the initial spontaneous occurrence of the macromolecular model from a disordered collection of parts does not appear as improbable as the spontaneous occurrence of the balance model under similar initial conditions. Secondly, the macromolecular model does not make use of any subunits which do not appear in the output sequence. In other words, the subunits form a closed system which provides all the parts necessary for generating the sequence as well as the sequence itself. Therefore, unlike the balance model, such a macromolecule could effectively reproduce itself by the simple process of fission, since each complete turn is capable of regenerating the entire period from the same environment. Fission might be expected to occur spontaneously at the weakest link after the macromolecule had grown excessively long. And thirdly, also unlike the balance model, this model macromolecule has a natural three-dimensional structure. In this particular example, since the number of subunits per turn divides the linear period with a remainder of one, the three-dimensional period will be seven linear periods or 889 subunits, assuming that it grows with axial symmetry. However, since each subunit may be expected to have different packing properties, axial symmetry would not in general be preserved. For example, if the A subunit is larger in volume than the B subunit, the helix would develop an irregular supercoiling as it grows assuming no other steric restrictions exist. There is then the likelihood that this supercoiling may cause new self-interactions which alter the initial rules of subunit addition. In other words, the simplest molecular representation of such a computer may naturally involve state-dependent logic, and should lead generally to very complex three-dimensional structure. On the other hand, such a macromolecule might be constrained in its growth configuration by adsorption to a substrate so that it was not free to adjust its folding to accommodate all distributions of large and small subunits. In this case the initial rule of addition may be occasionally overridden by packing restrictions. For example, if the axis were held in a straight line, we might expect to find less variation in the average subunit volume along loci parallel to the axis than would be expected along the helix itself or along a random sequence.

The actual chemical mechanisms involved in such a growth process would of course be more complex and would involve other components. Natta (1959) has shown that the stereospecific polymerization of α-olefins requires the presence of both a metal organic catalyst and a highly crystalline substrate. If a growing polymer were to remain adsorbed to the surface of a crystal or to another folded polymer we could expect rules of ordering which depend on both the growing polymer and the substrate. For example, the substrate might provide the information necessary to determine initially a particular helical configuration. One might also expect the synchronous polymerization of a coupled system as shown, for example, in the double helix configuration of Fig. 4. This particular hybrid configuration could consist of chains using the same or different sets of subunits. The

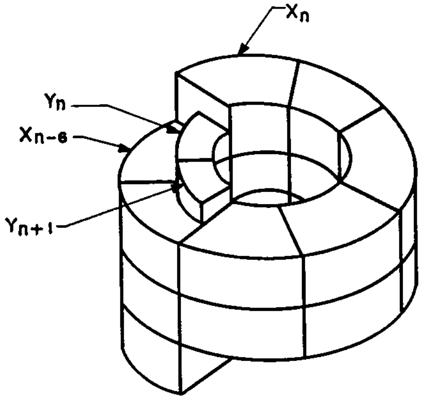

FIGURE 4 Schematic diagram of a hybrid macromolecular sequence computer. The type of subunit X_{n+1} to be added is determined by a rule depending most strongly on the types of subunits X_n and Y_n. The type of subunit Y_{n+1} to be added is determined by a rule depending most strongly on Y_n and X_{n-6}. To generate this sequence both chains must grow synchronously and all types of X and Y subunits must be available.

basic steps of computation are the same as before except that the "state" subunit could be given by one chain and the "input" subunit by the other. Only the synchronous growth of the two chains could be called autonomous. Furthermore, the growing hybrid could itself be absorbed on a structured substate which influences its three-dimensional configuration, and therefore to some degree its subunit selection.

Although the single macromolecule may satisfy the minimum requirements for an autonomous computer leading to long, random-like, but well ordered, self-repeating sequences, it is the hybrid associations of such conditionally restricted chains which suggest more closely the type of behavior found in biological systems. Thus we have a plausible hypothesis for the origin of aggregations of ordered macromolecules for which "replication" involves the repeating of the complex sequential behavior of the system as a whole, rather than simply the copying of each part by a pre-existing template. Furthermore, the individual components of such an aggregation may be expected to evolve those specific functions within the generation period which serve to increase the possibility of continued replication.

We shall discuss in the next section whether such a model is consistent with the evolution of genetic control as it is now observed.

We must emphasize that these examples are not presented as literal models of particular biological systems, but as illustrations of the ordering behavior of discrete-state, sequential computation which may occur in simple representations of fed-back or coupled molecular automata. This is in contrast to the behavior of simple time- and state-independent dictionary coding mechanisms which may transmit pre-existing order, but which specifically exclude all inherent organization or restrictions on which any process of computation is necessarily based.

THE EVOLUTION OF GENETIC CONTROL

The process of chance variation and selection is generally interpreted as a source of information which introduces order into replicating sequences which would otherwise be random or arbitrary (*e.g.*, see Muller, 1958). As long as it is assumed that in the absence of pre-existing genetic order there is only random disorder, the occurrence of living systems will inevitably pose two classical questions: (1) How did a disordered collection of elements which forms sequences with no restrictions produce, with reasonable probability, enough initial order to result in the general property of self-replication? (2) Assuming the general property of self-replication for all sequences, how did those particular sequences which now exist arise, with reasonable probability, from the set of all possible sequences? Instead of answering these difficult questions we have made the alternative assumption that genetic control is primarily a source of variability in otherwise highly restricted and well ordered sequences such as we commonly observe in crystals. The question which arises under this assumption is entirely different from the two above. We must now ask how aggregations of subunits which grow in well ordered arrays may develop naturally a capacity for chance variation in structure along with some process for preserving the variations which occur. To answer this question we have considered the general type of ordering which is essential for any type of discrete-state computation process exhibiting non-trivial behavior. We have shown how simple molecular configurations which occur naturally may produce well-ordered linear crystals with a rudimentary capacity for chance variation as well as a memory (*i.e.* the feedback loop) capable of preserving some of the variations which may occur. Replication in these ordered sequences is based, not on a printing press or template copying process, but on the recurrence of a set of discrete events which, in the absence of mutation, are closed and hence are periodic.

Under these conditions the increase of genetic variability itself becomes an evolutionary process depending on the selection of those growth configurations which have as much variability and memory capacity in the feedback path as is necessary to preserve the additional information accumulated in the total process of natural selection. Such configurations may arise from any process which decreases the restrictive self-interactions of the isolated macromolecule, and which thereby increases the capacity of the feedback path. This would be accomplished by surface growth on a substrate which holds the growing chain in a more open configuration, or by the intertwining of growing sequences which would lead to hybrid growth restrictions. By hybrid coupling of macromolecules which preserve the closure property essential for replication, there would evolve aggregations of macromolecules which are capable of autonomous growth only when suitably combined, but which might then individually acquire new configurations and functions when separated. From such aggregations the selective advantage of unrestricted variability would lead to the survival of those aggregations in which one type of macromolecule is not restricted in sequence by a specific secondary function, and which may then serve as an arbitrary memory; and in which other macromolecules by the evolution of optimum sequences could provide specific functions with maximum effectiveness. This would represent the simplest level of differentiation. According to this hypothesis, the occurrence of such an arbitrary memory sequence would not necessarily imply that it contains all the information necessary to determine completely each subunit of any associated sequence. That is, the possible order in one sequence may be unrestricted except by the complete replicative process, but this order is not necessarily sufficient to determine the detailed order of any sequence including its own. Thus, for example, a change of one base in a nucleic acid molecule which is sufficient to change one amino acid in a protein molecule may not be sufficient to select one amino acid from all the others. This possibility is consistent with the evidence given by Yčas (1961) showing that a coding ratio of one (one base corresponding to one of a subclass of amino acids) is consistent with the RNA and protein compositions of several plant viruses, as well as with the known alterations of amino acids resulting from mutations. The additional information which is then needed to uniquely determine each amino acid could be supplied by additional inputs which are not necessarily of genetic origin, or simply by the "state" of the growing configuration itself.

We therefore propose the following hypothetical picture of the origin of macromolecular sequences and the evolution of genetic control of these sequences: Beginning from an environment containing a non-specific source of energy, a disordered reservoir of distinguishable molecular subunits, and perhaps regular crystal surfaces with simple periodicities, there occur linear polymerizations with more or less random subunit distribution. As the length and concentration of polymers increase there arises a liquid-crystal state of matter in which the inherent self-interaction produced by folding and coiling gradually become strong enough to conditionally restrict or favor the addition of certain subunits at a given state of growth. As these restrictions increase through continued concentration or interaction of polymers, or by adsorption on surfaces, there arises enough self-interaction or feedback to cause some autonomous sequence generation. At this stage we must assume that the interacting subunits together with the restrictions involved in the computation satisfy some form of *closure* condition, *i.e.,* the property that all states of the growing configuration can react with one or another of the types of subunits which make up the growing configuration. This condition is no more than is required for the growth of any regular crystal. In simple three dimensional crystals this condition can usually be met by one or two different types of subunits. For the more flexible linear crystals a greater variety of subunits is to be expected.

As we have shown, a collection of distinguishable elements under these conditions can produce the type of self-replicating, mutable sequences on which the evolutionary processes of variation and selection are defined to operate. The variation may arise initially from errors in subunit addition, from modified folding or feedback interactions, by interaction with substrates, or by hybridization, one type of which was illustrated schematically in Fig. 4. This association of more than one type of sequence may then lead to differentiation of function, one type of sequence providing primarily variation and memory, the others providing primarily structural or metabolic functions. At this stage of complexity there is no point in speculating on chemical details, however the logical steps may be considered a little further.

The examples of sequence generators we have given illustrate that the amount of genetic control is measured, not by the length of the sequence period, but by the number of states in any feedback loop which lead to functionally distinct sequences. Therefore as genetic information gradually accumulates from the process of natural selection, the length of the sequence period need not increase, but the capacity of the feedback path must increase. It is important to consider the term "feedback path" here in the logical sense which may include the entire circuit of information used in the memory, growth, and replication cycle, and not necessarily only one turn of a helix, as in the examples. Some of the information stored in the memory component of this path may function more like the information in the program of a computer than like a master copy of all the sequences which might occur in the life of the organism. If this should be the case in biological systems, we must consider how we may separate the information accumulated by variation and selection and stored in an arbitrary memory, from the information inherent in the total aggregation which together form the autonomous replicating unit. We therefore consider what type of evidence may exist in the order or behavior of present day biological macromolecules which would indicate any trace of autonomous computation process that might have originated such order or that may still occur to some degree in living organisms.

DISCUSSION OF THE EVIDENCE

The most convincing experimental evidence for the occurrence of autonomous computation in any macromolecular system would certainly be its direct demonstration by performing a synthesis of long-period linear macromolecules with ordered subunits from a reservoir containing no such order. Should this be possible, the chemical nature of such macromolecules might indicate their possible role in the earliest biological systems. With the rapidly growing knowledge of the control of polymer structure, exploration of the possibility of molecular computation should be seriously considered.

In our present state of knowledge of natural biological sequences, to design a test for the occurrence of computation processes may be a difficult problem since the resolution of even simple, completely autonomous sequences into their logical structure can be a tedious process. Since we know that a large amount of genetic input now exists in the control of amino acid sequence, we can certainly not expect autonomous control except perhaps over the functionally less critical regions of protein chains. For example, the occurrence of genetically controlled, single amino acid substitutions in proteins is clearly inconsistent with completely autonomous computation processes at those regions where the substitution occurs. However, it would require much more evidence to show that this genetic control is itself sufficiently complete and autonomous to determine the choice of any amino acid at any position by a mechanism which is entirely independent of the state of the rest of the growing sequence, or the system as a whole. The present evidence of genetically influenced amino acid substitutions shows only that some detailed genetic control of protein sequence occurs. The nature and extent of this control remain largely a matter of conjecture.

On the other hand, there are several general observations on protein structure and behavior which are not so easily explained by a process of random variation and natural selection, as the sole source of order, expressing itself through genetic control of linear sequences by simple block codes. One of the most obvious over all characteristics of amino acid sequences in protein is their lack of apparent regularity. In fact, the statistical analysis of neighbor pair distributions in globular proteins of known sequence shows them to be indistinguishable from a random distribution (Gamow *et al.,* 1956; Yčas, 1958). We need not show how a random distribution arises from a series of random events, but the occurrence of a random distribution arising from well ordered events requires some discussion. Natural selection, even if it began with random sequences and depended on random variation, does not explain the persistence of a random distribution unless this distribution reflects some biological function. It has been suggested that randomness may have been preserved because this distribution maximizes the information capacity (Gamow, *et al.,* 1956); however, all estimates of the information content in proteins necessary for their known functions are an exceedingly small fraction of their theoretical capacity (Quastler, 1953; Augenstine, 1958), so that this explanation is not convincing. A related problem is presented by the occurrence in many enzymes of well ordered partial sequences of amino acids which may be completely removed with little or no effect on the activity of the enzyme (*e.g.*, see Anfinsen, 1959). In general, we should expect the process of natural selection alone to result in a degree of precision and uniqueness in the choice of each amino acid which corresponds in some way to the precision and uniqueness of the function of that portion of the sequence where it occurs. On the contrary, we find that some amino acid sequences are uniformly precise and unique over the entire protein, whereas the known functional specificities appear to be concentrated in relatively small regions of the chain. If the order of the subunits arises only by selection from random alterations, why is a precise sequence preserved through many generations when its loss or alteration would not strongly affect the known function?

The inverse problem arises for many structural proteins which show a wide variation in amino acid composition even though the function remains the same. For example, Lucas, Shaw, and Smith (1960) have concluded that the amino acid compositions of the silk fibroins from many species not only show a surprisingly large variation, even though the function of all the fibroins is essentially the same, but that there is no general correlation between the compositions of the fibroins from closely related species. How does the process of selection by random alteration of single amino acids explain the occurrence in two similar species of protein with the same function which have nevertheless evolved widely differing amino acid compositions?

These questions are answered quite naturally by a model of protein synthesis which at some stage involves a sequential computation process. In the first place, the occurrence of pair distributions of amino acids which are indistinguishable from a random distribution does not in itself allow us to conclude that the sequence is not generated by very simple rules, or that the possible sequences are unrestricted (*cf.* Brenner, 1957). It is one of the properties of some simple fed-back sequence generators that they can produce random-like order. Nor does this random-like order imply a large information capacity, for these sequences may be highly redundant in terms of their selective information content. However, in terms of structure or function such a sequence need not be so redundant, since within one period each segment which contains at least as many subunits as the feedback loop is structurally distinct from other similar segments.

The problem of well ordered, but "non-functional" sequences may also be understood by a sequential computation process. The simple examples of sequence generators we have shown illustrate clearly how the active site of computing is localized over only a few subunits at a time, but how the remainder of the sequence has had a similar precise and essential computing function. Once a particular active configuration has been attained, however, the record of the computation process necessary to reach that configuration is not likely to be needed for its activity. Therefore, when a precise and unique local configuration is necessary for certain reactive functions, as it appears to be for most enzymes, then it is reasonable to expect the computational steps leading to this configuration also to be uniformly precise and unique.

On the other hand, when the over all structural configuration is important, as in the case of fibrous proteins, the choice of subunits may not be critical as long as the three-dimensional form is maintained. In molecular sequential computers, it is possible for certain mutations to result in very large changes in subunit composition without altering the over all three-dimensional structure. Thus, according to a sequential computing model of protein synthesis we could reasonably expect functionally similar sequences with widely different compositions from closely related species, (*i.e.* species separated by only a few mutations).

Some degree of sequential ordering also provides a reasonable explanation for the synthesis of fibrous proteins such as collagen which show evidence of long and short range intramolecular periodicity (*e.g.* see Schmitt, 1959). Of course the positions of each amino acid, periodic or not, may also be explained in principle by the exclusive action of a coded nucleic acid template which contains all the information necessary to select separately each amino acid, and which accumulated this information entirely by selection from a random search process. However, according to our hypothesis of the evolution of genetic variability we would expect no more detailed genetic determination of each amino acid than is necessary to allow the variability consistent with the possible functions of the sequence. Any regularities or periodicities in sequence represent redundant information and therefore they need not be determined individually by genetic information alone.

Two more general properties of sequential computer models should be mentioned. First, it is clear that each subunit in a computer must be added in the correct temporal as well as spatial order with respect to its neighboring subunits; and second, all the subunits used in the computation process must be simultaneously available even though the end result contains only a few subunits. The present evidence that the position of addition of amino acids in a growing sequence is not random, but a steady sequential addition from one end of the sequence (Bishop, Leahy, and Schweet, 1960; Dintzis, 1961) indicates at least some transfer of information from the growing sequence which is necessary for continued addition of subunits. Whether this information has some influence on the choice of any subunits is not known. Furthermore, the evidence that all amino acids must be present for both ribonucleic acid synthesis as well as protein synthesis (Gale and Folkes, 1953) is suggestive of an interdependent RNA-protein sequential growth process.

We next consider how we might expect to find evidence of sequential computation from the known amino acid sequences. Until a larger sample is available it is unlikely that statistical analysis will yield convincing evidence for or against the computation process. However, we may at least assume a few simple configurations and test them for consistency with the known sequences. We shall first use the linear sequence delivered by the balance model in Fig. 1 to illustrate our approach.

If we knew nothing of the source of this sequence we might write it down as we did in Table I. By and large, this sequence satisfies the obvious tests for randomness, and in any case it is fair to say that its simple order is not immediately obvious. However, if we should write this sequence down with only six symbols per line, there would appear some patterns in the triangular placement of the symbol A, as shown in Table II.

TABLE II

PARTIAL SEQUENCE OF BALANCE MODEL WRITTEN WITH SIX SYMBOLS PER LINE

			.	.	.
A	B	A	B	A	B
A	A	B	B	A	A
B	B	B	A	B	B
B	A	B	A	A	B
A	B	B	A	A	A
B	B	A	B	B	B
B	A	B	B	A	B
A	B	B	A	B	B
A	A	B	A	A	B
A	A	A	B	B	B
A	A	A	A	B	A
.	.	.			

The order has become more obvious because this particular two-dimensional configuration brings together some of the subunits which were coupled by the computing mechanism to determine the following subunit. The triangle is the characteristic pattern in this representation for the symbol which represents an identity operation. Finally by writing the sequence in a helical array, as illustrated in Fig. 3, we may then discover the simple rules of computation. It is important to recall from the definition of a computer that although these rules may be simple, it is essential that they involve at least three classes of symbols; *i.e.*, the *input, state,* and *output* symbols. When searching for statistical correlations between symbols it is therefore clear that we must consider at least triplet distributions if we hope to find any clear evidence of a computation process in the sense it has been defined.

As pointed out earlier, the examples have been chosen primarily to illustrate the general concept of sequential, discrete-state computation. We should not expect such simple behavior from real polymers as occurs in these examples, since the physical interactions necessary for molecular subunit identification and reaction require more than abstract logical rules. Furthermore, none of the models has required more than the simplest logic; for example, only single feedback paths have been used. Finally we know that biological macromolecules probably require a large amount of input information in their synthesis which would easily mask any internal restrictions. The point we wish to emphasize by this simple model is that if there is any trace of order in biological sequences which originated at some stage by a sequential computing process, either in protein, nucleic acid, or more likely a hybrid of both, then this order should not appear unless the sequence is arranged in some configuration related to the growth configuration. Furthermore, we cannot expect to recognize computed order by the analysis of pair distributions alone, since pair distributions even in the simplest computers may be indistinguishable from the pair distributions of a random sequence. Therefore we must consider at least triplet distributions, or the possible interactions of more than two subunits.

Since the only configuration which is known to occur widely in different proteins is the α-helix (Yang and Doty, 1957; Kendrew *et al.,* 1960), we have analyzed the statistical behavior of amino acid sequences in tobacco mosaic virus protein (Tsugita *et al.,* 1960) and ribonuclease (Hirs, Moore, and Stein, 1960) observing the nearest neighbor triplet groups which would occur in the helical configuration with 2, 3, 4, and 5 subunits per turn. This has been done using matrices similar to those used by Gamow *et al.,* (1956) where the n^{th} subunit is listed on one axis and the $(n + j)^{\text{th}}$ subunit is listed on the other axis.[4] In the present case, however, we have entered the $(n + \mathrm{j} \pm 1)^{\text{th}}$ subunit in the matrix, where $j = \pm 1, \pm 2, \pm 3$, and ± 4. The plus and minus signs indicate reading from the N- and C-terminal subunits respectively. This allows us to find the triplet as well as the pair distributions. This type of analysis will indicate any obvious correlations among a given subunit, the nearest subunit one turn away in the helix, and the next subunit in a given direction along the chain. We know that a regular helical configuration along the chain can occur only for short segments since folding must occur and since residues such as proline do not fit in the α-helix. Even in the absence of genetic input, no simple computing rule could be expected to hold generally for such proteins.

In tobacco mosaic virus protein and ribonuclease we find that most pair distributions are indistinguishable from random distributions. There is some evidence of intersymbol influence in ribonuclease for the nearest axial neighbors on a helix of either three or four subunits per turn. The same result has been reported independently by Morgan (1960) for ribonuclease using a similar analysis.

[4] It has been pointed out by Yčas (1958) that this method is not valid if the frequency distribution of subunits is far from random. In TMV protein, cysteine, lysine, and tryptophane have lower frequencies than expected from a random selection. Omission of these amino acids made no significant difference in the fit of the pair distributions to the random distribution. However, the calculation of χ^2 for the triplet distributions was made using only 15 subunits instead of all 18. The distributions in other proteins will not be discussed in this paper.

However, as we have pointed out, in a model involving some sequential computation we may expect random *pair* distributions even though a closely related *n-tuple* distribution is far from random. This situation occurs in the tobacco mosaic virus protein sequence. Consider the pair distribution n, $n + 4$ as shown in Table III and the triplet distribution n, $n + 4$, $n + 5$, as shown in Table IV. We see that whereas the n, $n + 4$ pair distribution is random above the 50 per cent level of significance, the n, $n + 4$, $n + 5$ triplet distribution is non-random near the 5 per cent level of significance. The similar triplet association generated from the C-terminal end of the sequence, n, $n - 4$, $n - 5$, is a random distribution, with $\chi^2 = 0.08$ so that there is some evidence of an asymmetric correlation which is typical of any sequential generation process. The immediate cause of this statistical behavior may be traced to the six serine residues near the C-terminal end of the chain. If we

TABLE III

DISTRIBUTION OF AMINO ACIDS PAIRS, n, $n + 4$, IN TOBACCO MOSAIC VIRUS PROTEIN

Number of subunits per box	In TMV protein (observed)	In Poisson distribution (calculated)
0	204	202.0
1	94	95.4
2	20	22.5
3	5	3.54
4	1	0.42
5	0	0.04

χ^2 (calculated for three degrees of freedom at the 50 per cent level of significance) = 2.4
χ^2 (observed) = 1.76

TABLE IV

DISTRIBUTION OF AMINO ACID TRIPLETS, n, $n + 4$, $n + 5$, IN TOBACCO MOSAIC VIRUS PROTEIN (GENERATED FROM N-TERMINAL END)

Number of subunits in box	In TMV Protein (observed)	In Poisson distribution (calculated)
0	3225	3227
1	148	146.3
2	1	3.32
3	1	0.05
4	0	0.0006

χ^2 (calculated for two degrees of freedom at 5 per cent level of significance) = 5.99
χ^2 (observed) = 5.1 (Yates' correction applied)

TABLE V

PARTIAL SEQUENCE OF TOBACCO MOSAIC VIRUS PROTEIN WRITTEN WITH FOUR RESIDUES PER LINE

125			
ASP	ILEU	ASPN	LEU
ILEU	VAL	GLU	LEU
ILEU	ARG	GLY	THR
GLY	SER	TYR	ASPN
ARG	SER	SER	PHE
GLU	SER	SER	SER
GLY	LEU	VAL	TRY
THR	SER	GLY	PRO
ALA	THR		
	158		

write the tobacco mosaic virus protein sequence with four residues per line as a radial projection of a helix with four residues per turn as shown in Table V, we see that these six serine residues are adjacent to one another in a triangular pattern,¸similar to the patterns of the shift register model.

Unfortunately, the statistical samples are too small to draw strong conclusions from these interesting similarities with sequential computing models. They are given here primarily to illustrate one method of searching for possible restrictions. The ribonuclease sequence shows similar but less significant non-randomness in helical triplet distributions, but there also appears to be even longer range order (Lanni, 1960, 1961) which has not yet been associated with a sequential computation model.

We may conclude that the observed distributions of amino acids in some proteins suggest some degree of sequential computation, and that the data are consistent with a sequential growth process from the N-terminal end of the chain. However the sequence data are insufficient to provide any convincing conclusions. In any case we could not expect statistical data alone to indicate at what stage of synthesis any such order may be introduced. Only by more detailed knowledge of subunit interactions in simple configurations could strong evidence for or against autonomous restrictions be accumulated.

CONCLUSION

Biological macromolecules are recognized by the elaborate arrangements of their subunits. We know that these arrangements are being continually repeated by some ordering process within living cells. Some of this order, which is known in detail only for a few proteins, can be influenced by nucleic acid molecules, but the mechanisms and limits of this influence are not known. The current ideas of this ordering process assume the pre-existence of the same degree of order in nucleic acid linear base sequences, which are replicated by copying, and which express their order in the linear arrangement of amino acids in protein by a simple, dictionary-type code. Most codes which have been proposed account only for the transmission of linear order, are time-independent, and contain no intersymbol restrictions. Furthermore, information feedback is not considered in the ordering process, and the codes are independent of the state of the growing configuration and its local environment. Since these codes usually exclude any inherent restrictions or expectations of order which did not arise from pre-existing genetic order they do not in themselves lead to any explanation of the primeval origin of this high degree of order which is being replicated and coded in protein, or of the origin of the replication or coding processes themselves. Furthermore, we find in the process of natural selection from a random search process no plausible mechanism for effectively increasing the insignificant probability of producing one particular sequence from a set of otherwise unrestricted and unrelated sequences the size of a protein molecule.

We are therefore led to consider the possibility that the precursors of biological macromolecules were not random sequences, but naturally ordered crystal structures which resulted from the restrictions inherent in their growth similar to the restrictions which we find in all other crystallization processes. We indicate how the general logical process of discrete-state computation in simple configurations with feedback can assemble elaborate, repeating, well ordered sequences without requiring highly improbable pre-existing sequences from which to copy. A molecular representation of such a computer is suggested by the process of stereospecific polymerization, in which the choice of subunit is determined by conditional, state-dependent rules. From such redundantly ordered macromolecules, replicating by fission, the evolutionary process of variation and selection may then effectively accumulate as much genetic information capacity as produces survival value. In other words, instead of making the special assumption that in living organisms ordered macromolecular sequences are the evolutionary end result of the spontaneous, chance origin of genetic systems, we propose that the present genetic mechanisms themselves are the evolutionary result of the natural occurrence of ordered macromolecular sequences.

As a theory for the origin of macromolecular order, the fed-back, discrete-state computer model does not necessarily conflict with the present common ideas of genetic control of protein function through the use of the information contained in nucleic acid sequences; for as we have pointed out, according to the computer model, the accumulation of genetic information by natural selection necessarily leads to sequences with ever-increasing information capacity in the feedback path.

It is quite conceivable that this capacity may have grown so large that the only autonomous computation that now remains is through the entire process of variation and selection, which may certainly be considered as an information feedback loop from functioning proteins to nucleic acid sequences. Even if it should be the case that the entire, detailed ordering of every amino acid residue in all proteins is now under the complete control of the information generated *only* in this one great evolutionary feedback loop and stored *only* in linear base sequences, then our proposed hypothesis would still serve as a possible explanation for what would otherwise depend upon the initial chance occurrence of an enormously improbable computing configuration. On the other hand, we have presented evidence that the occurrence of long, well-ordered sequences with random subunit pair distribution is not in itself a sufficient reason to exclude the possibility of very simple, non-genetic rules and very short feedback paths in the synthesis of biological macromolecules.

Many biological as well as philosophical systems may be discussed in the light of a model for living organisms based essentially on fed-back discrete-state automata; however, at this point we shall only suggest how future empirical evidence may be expected to distinguish between sequences generated autonomously by computation and sequences translated from other pre-existing sequences by a dictionary code. A basic experimental difficulty in studying existing biological sequences is that both ordering processes may occur in the same sequence. According to our hypothesis for the evolution of genetic systems we may expect detailed genetic control to evolve only in those regions of the sequence in which the exact choice of residue is functionally critical. Therefore if one were attempting to demonstrate genetic control by correlating amino acid substitutions in critical proteins with nucleic acid structure, the experiment would be inherently biased in favor of finding just that which is sought. On the other hand, the existence of clear statistical regularities in biological sequences does not in itself imply inherent non-genetic restrictions in possible orders, since the regularity may be interpreted as only a reflection of some structure which has some unknown survival value and which therefore could have arisen entirely by random variation and selection. Since with our knowledge of computers we may in principle compute any sequence, and in our ignorance of the function of biological sequences we may attribute survival value to any sequence, we must add the condition that any sequence regularities shall be interpretable in terms of a chemically reasonable computer model if they are to support the computing hypothesis, or similarly, that these regularities shall be interpretable in terms of some reasonable biological function which has survival value if they are to support the selective theory for explaining the origin of these sequence regularities.

Finally, we suggest that the possible role of autonomous macromolecular sequence computation in the origin of life may be most effectively approached experimentally as a problem of macromolecular engineering under controlled, artificial environments, since the evidence of any such naturally occurring process may no longer exist on the earth. For example, a search might reasonably be organized for programmed macromolecular sequences with long period arising from a reservoir of well chosen small molecules and perhaps simple crystalline surfaces and short-period polymers. The chemical structure and behavior of any such macromolecules, should they occur, may give some indication of the role of simple computing processes in the origin of biological sequences.

This work originated at the Department of Medical Physics, Karolinska Institute, Stockholm, where the author was supported by a National Science Foundation Senior Postdoctoral Fellowship. The work was continued at Stanford University, supported in part by the National Institutes of Health, Research Grant A-2557-C3, United States Public Health Service.

Received for publication, April 27, 1961.

REFERENCES

ANFINSEN, C. B., The Molecular Basis of Evolution, New York, John Wiley & Sons, Inc., 1959, Chapter 6.

AUGENSTINE, L. G., Protein structure and information content, *in* Symposium on Information Theory in Biology, (H. Yockey, editor), New York, Pergamon Press, 1958, 103.

BIRKHOFF, G., and MACLANE, S., A Survey of Modern Algebra, New York, Macmillan Co., 1947, 409.

BISHOP, J., LEAHY, J., and SCHWEET, R., 1960, Formation of the peptide chain of hemoglobin, *Proc. Nat. Acad. Sc.*, **46**, 1030.

BRENNER, S., 1957, On the impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins, *Proc. Nat. Acad. Sc.*, **43**, 687.

BRENNER, S., The mechanism of gene action, *in* Biochemistry and Human Genetics, CIBA Foundation Symposium, (Wolstenholm and O'Conner, editors), London, J. & A. Churchill, Ltd., 1959.

COMMONER, B. C., 1959, A re-examination of theories, *Nature*, **184**, 1998.

CRANE, H. R., 1950, Principles and problems of biological growth, *Scient. Monthly* **70**, 376.

CRICK, F. H. C., 1958, On protein synthesis, *Symp. Soc. Exp. Biol.* (Great Britain), **12**, 138.

CRICK, F. H. C., GRIFFITH, J. S., and ORGEL, L. E., 1957, Codes without commas, *Proc. Nat. Acad. Sc.*, **43**, 416.

DANCOFF, S., and QUASTLER, H., The information content and error rate of living things, *in* Information Theory in Biology, (H. Quastler, editor), Urbana, University of Illinois Press, 1953, 263.

DINTZIS, H., 1961, Assembly of the peptide chains of hemoglobin, *Proc. Nat. Acad. Sc.*, **47**, 247.

ELSPAS, B., 1959, The theory of autonomous linear sequential networks, *Inst. Radio Eng. Tr. on Circuit Theory*, Vol. **CT-6**, 1, 45.

FRANK, F. C., 1949, The influence of dislocations on crystal growth, *Discussions Faraday Soc.*, **5**, 48.

GALE, E., and FOLKES, J., 1953, The assimilation of amino acids in bacteria, 14. Nucleic acid and protein synthesis in *Staphylococcus aureus*, *Biochem. J.*, **53**, 483.

GAMOW, G., RICH, A., and YČAS, M., The problem of information transfer from the nucleic acids to proteins, *in* Advances in Biological and Medical Physics, (J. Lawrence and C. Tobias, editors), New York, Academic Press, Inc., 1956, 23.

GIERER, A., Recent investigations on tobacco mosaic virus, *in* Progress in Biophysics and Biophysical Chemistry, (J. A. V. Butler and B. Katz, editors), New York, Pergamon Press, 1960, 299.

GOLOMB, S. W., WELCH, L. R., and DELBRÜCK, M., 1958, Construction and properties of comma-free codes, *K. Danske Vidensk. Selsk. [Biol. Medd.]*, **23**, 9.

HAM, G., 1959, Helical stereospecific polymerizations, *J. Polymer Sc.*, **40**, 569.

HAMMING, R. W., 1950, Error correcting and error detecting codes, *Bell Syst. Techn. J.*, **29**, 147.

HIRS, C. H. W., MOORE, S., and STEIN, W. H., 1960, The sequence of amino acids in ribonuclease, *J. Biol. Chem.*, **235**, 633.

HUFFMAN, D. A., The synthesis of linear sequential coding networks, *in* Proceedings of the Third London Symposium on Information Theory, p. 77, September, 1955; *also in* Information Theory, (C. Cherry, editor), New York, Academic Press, Inc., 1956.

IDELSON, M. and BLOUT, E. R., 1958, Polypeptides XVIII. A kinetic study of the polymerization of amino acid *N*-carboxyanhydrides initiated by strong bases, *J. Am. Chem. Soc.*, **80**, 2387.

JACOBSON, H., 1958, On models of reproduction, *Am. Scientist*, **46**, 255.

KELLER, A., and O'CONNER, A., Study of single crystals and their associations in polymers, *in* Configurations and Interactions of Macromolecules and Liquid Crystals, *Discussions Faraday Soc.*, No. 25, 1958, 114.

KEMENY, J., 1955, Man viewed as a machine, *Scient. Am.*, **192**, 4, 58.

KENDREW, J. C., DICKERSON, R. E., STRANDBERG, B. E., HART, R. G., DAVIES, D. R., PHILLIPS, D. C., and SHORE, V. C., 1960, Structure of myoglobin, *Nature*, **185**, 422.

KIMURA, M., 1961, Natural selection as the process of accumulating genetic information in adaptive evolution, *Genetical Research*, **2**, 127.

LANNI, F., 1960, Analysis of sequence patterns in ribonuclease, I. Sequence vectors and vector maps, *Proc. Nat. Acad. Sc.*, **46**, 1563.

LANNI, F., 1961, Analysis of sequence patterns in ribonuclease, II. Primitive groups, their coordinations, and periodicity, *Proc. Nat. Acad. Sc.*, **47**, 261.

LINSCHITZ, H., The information content of a bacterial cell, *in* Information Theory in Biology, (H. Quastler, editor), Urbana, University of Illinois Press, 1953, 251.

LUCAS, F., SHAW, J. T. B., and SMITH, S. G., 1960, Comparative studies of fibroins, *J. Molecular Biol.*, **2**, 339.

MCCULLOCH, W. S., and PITTS, W. A., 1953, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophysics*, **5**, 115.

MILLER, S. L., 1955, Production of some organic compounds under possible primitive earth conditions, *J. Am. Chem. Soc.*, **77**, 2351.

MORGAN, R. S., 1960, Concerning the ribonuclease sequence, *J. Molecular Biol.*, **2**, 243.

MOROWITZ, H. J., 1959, A model of reproduction, *Am. Scientist*, **47**, 261.

MULLER, H. J., 1958, Evolution by mutation, *Bull. Am. Math. Soc.*, **64**, 137.

NATTA, G., 1959, Kinetic studies of α-olefin polymerization, *J. Polymer Sc.*, **34**, 21.

NATTA, G., PINO, P., CORRADINI, P., DANUSSO, F., MANTICA, E., MAZZANT, G., and MORAGLIO, G., 1955, Crystalline high polymers of α-olefins, *J. Am. Chem. Soc.*, **77**, 1708.

NEWELL, A., and SIMON, H. A., The simulation of human thought, Rand Corporation Report P-1734, June 22, 1959.

PENROSE, L. S., 1959, Self-reproducing machines, *Scient. Am.*, **200**, 6, 105.

PONTECORVO, G., 1958, Self reproduction and all that, *Symp. Soc. Exp. Biol.* (Great Britain), **12**, 38.

QUASTLER, H., The specificity of elementary biological functions, *in* Information Theory in Biology, (H. Quastler, editor), Urbana, University of Illinois Press, 1953, 170.

SCHMITT, F. O., Interaction properties of elongate protein macromolecules with particular reference to collagen, *in* Biophysical Science—A Study Program (J. L. Oncley, editor), New York, John Wiley & Sons, Inc., 1959, 349.

SEARS, G. W., 1959, Growth of tobacco mosaic virus particles, *Science*, **130**, 1477.

SZWARC, M., 1958, A new approach to the problem of stereospecific polymerization, *Chem. & Ind.*, No. 48, Nov. 29, 1589.

TSUGITA, A., GISH, D. T., YOUNG, J., FRAENKEL-CONRAT, H., KNIGHT, C. A., and STANLEY, W. M., 1960, The complete amino acid sequence of the protein of tobacco mosaic virus, *Proc. Nat. Acad. Sc.*, **46**, 1463.

TURING, A. M., 1936, On computable numbers, with application to the Entscheidungsproblem, *Proc. London Math. Soc.*, **2-42**, 230.

TURING, A. M., Can a machine think?, *in* The World of Mathematics, Vol. 4, (J. R. Newman, editor), New York, Simon and Schuster, 1956, 2099.

VON NEUMANN, J., The general and logical theory of automata, *in* Cerebral Mechanisms in Behavior (L. E. Jefress, editor), New York, John Wiley & Sons, Inc., 1951.

WOESE, C. R., 1961, A nucleotide triplet code for amino acids, *Biochem. and Biophysic. Research Communications*, **5**, 88.

YANG, J. T., and DOTY, P., 1957, The optical rotatory dispersion of polypeptides and proteins in relation to configuration, *J. Am. Chem. Soc.*, **79**, 761.

YČAS, M., The protein text, *in* Symposium on Information Theory in Biology (H. Yockey, editor), New York, Pergamon Press, 1958, 70.

YČAS, M., 1961, Replacement of amino acids in proteins, *J. Theoret. Biol.*, **1**, 244.