On the Problems of Evolution and Biochemical Information Transfer

ALEXANDER RICH

Department of Biology,
Massachusetts Institute of Technology,
Cambridge, Massachusetts

I. Introduction

Perhaps the most striking characteristic in the development of biochemical understanding during the past decade has been the discovery that nucleic acids play the central role in the transmission of molecular information. Although a great deal of data had been compiled on pathways of intermediary metabolism by the early 1950's, relatively little was known about those reactions which govern the over-all flow of metabolism. In contrast to that situation, we now have a reasonably coherent explanation of the replication and transmission of genetic information and the process whereby this information leads to the ultimate regulation of metabolism. At the very heart of the system are the nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). These molecules now occupy such a central role in biochemistry that it is difficult to realize the extent to which their properties and functions were unknown as briefly as a decade ago.

Inevitably the development of a more comprehensive under-

standing concerning the role of the nucleic acids is accompanied by a large number of questions concerning the manner in which these molecules came to occupy such a commanding position in biochemical systems. In short, what has been their evolutionary history? Unfortunately, we have little experimental data at the present time with which to answer these questions. Nonetheless, a study of the over-all pattern of genetic replication and genetic expression leads one to ask certain direct questions and to suggest possible evolutionary routes which might lead to the present-day system. Several speculations and hypotheses about the evolution of the nucleic acids are developed in this article. Some of them may prove useful in stimulating further discussion and they may also suggest directions for an experimental approach to the problem.

II. An Outline of Nucleic Acid Function

A brief and simplified outline of the functions of the nucleic acids is schematically illustrated in Fig. 1. The DNA molecule is the major carrier of genetic information. When we use the term "information" in conjunction with the nucleic acids, it is often synonymous with the word "sequence." The nucleic acids are built in a linear chain containing four major nucleotide components and we believe that the "information" resides in the order of these four units. This is analogous to the way in which "information" in a sentence is determined by the ordering of the letters. Thus we can say that the molecular language of the nucleic acids is written with a four-letter alphabet.

Broadly speaking the DNA molecule has two functions: first, it must reproduce itself during cell division so that its information can be transmitted to daughter cells; and secondly, it must express this information by influencing the metabolism of the cell. According to our present understanding, the major mechanism for metabolic control is exercised through regulating the supply of protein molecules which act as catalysts. In the presence of these enzymes, chemical reactions proceed readily; in their absence these reactions proceed very slowly if they occur at all. Only in the past few years has the mechanism underlying these two functions of DNA become clear. The work of Kornberg and his collaborators (see Bessman *et al.*, 1958) has demonstrated that the replication of DNA is governed by the activity of an enzyme, DNA polymerase. Given a primer of DNA, the enzyme has the property of separating the twisted

chains in the double-stranded molecule and organizing a series of complementary nucleoside triphosphate bases along each strand which are then polymerized to produce two identical molecules.

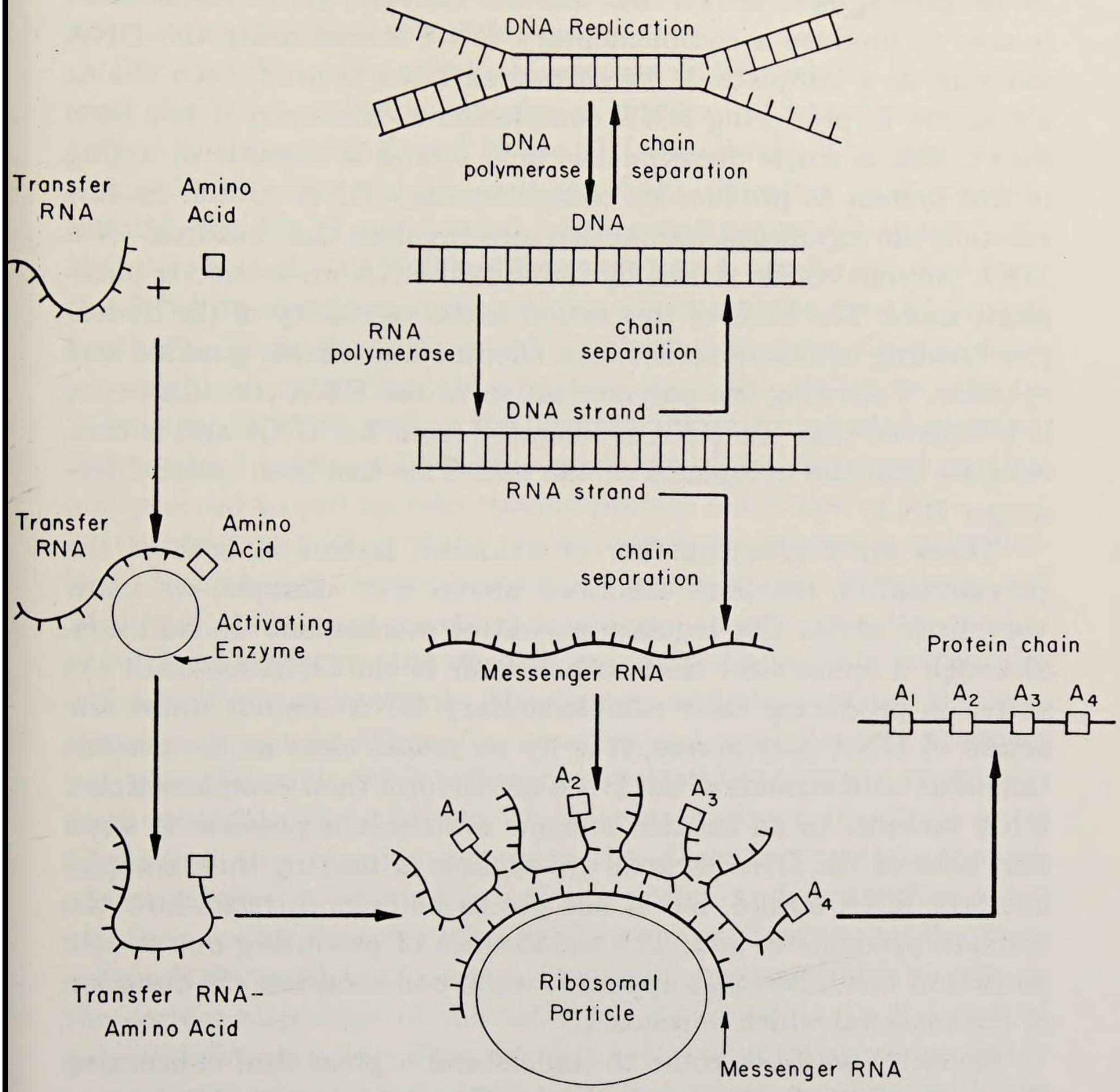

FIG. 1. A schematic outline of nucleic acid function. The ladder-like figures represent two-stranded nucleic acids and the bases are represented by the short cross lines.

The experiments of Meselson and Stahl (1958) demonstrate *in vivo* the separation of the two strands of the parent DNA molecules. The basic requirement in this replication mechanism is the "reading" or recognition by a deoxynucleoside triphosphate of a complementary site on one of the separated DNA chains. Thus the adenine

residue is hydrogen bonded specifically to thymine, and the guanine residue specifically to cytosine.

Recently Weiss (1960) and Hurwitz *et al.* (1960), as well as other investigators, have shown that another enzyme, RNA polymerase, is able to produce a complementary RNA strand using the DNA molecule as a template. If two-stranded DNA is used, both chains are active in producing RNA complements. However, it has been shown that a single deoxynucleic acid strand is capable of acting in this system to produce its complementary RNA strand. In this reaction the significant interaction also involves the "reading" of a DNA polynucleotide strand by individual RNA nucleoside triphosphate units. The basis of this action is the specificity of the hydrogen bonding interactions between adenine and uracil, guanine and cytosine. Following the polymerization of the RNA strand *in vivo*, it is believed that the RNA is liberated from the DNA and is then released into the cytoplasm of the cell. This has been called Messenger RNA.

There are a great number of unknown factors in both of the polymerization reactions described above. For example, we know very little about the regulating control mechanism. In addition, although it seems clear that both strands of the DNA molecule are active in producing their complementary DNA strands under the action of DNA polymerase, it is by no means clear at the present time that both strands of the DNA act to form their complementary RNA strands. In an isolated enzyme system it is possible to show that both of the DNA strands are capable of making their complementary RNA strand, but it has yet to be demonstrated how the reaction proceeds *in vivo*. The significance of producing one or two strands of RNA becomes apparent when one considers the function of the material which is released.

Recently we have come to understand a great deal concerning the mechanism of protein synthesis. The most remarkable feature of the protein synthetic process is the extent to which the nucleic acids have a central role in organizing the assembly of amino acids into completed protein chains. This is also shown diagrammatically in Fig. 1. Short lengths of RNA, called Transfer RNA, containing some 90 nucleotides have been shown to react with the amino acids to make an RNA-amino acid complex. This reaction is carried out on the surface of an activating enzyme. At the present time the evidence suggests that there is a specific Transfer RNA molecule

for each amino acid, as a specific activating enzyme. The activating enzyme is specific in that it only accepts a particular amino acid (actually, an amino acid-adenylate), and a particular Transfer RNA molecule, presumably because it can detect or "read" a segment along the polynucleotide chain.

The newly polymerized Messenger RNA which is released from the DNA molecule contains a sequence of bases complementary to the sequence found in the DNA molecule. This material is metabolized very rapidly and has been detected only recently through the use of rapid pulse labeling techniques (Brenner *et al.*, 1961; Gros *et al.*, 1961). The Messenger RNA can be followed from the DNA to the ribosomal particles which are the sites of protein synthesis. These particles, roughly spherical in size with a molecular weight of approximately 4 million, have the ability to take up Messenger RNA as well as the Transfer RNA-amino acid complex. Polymerized polypeptide chains come out of the ribosomal particle and proceed to coil up into specific protein molecules.

Our understanding of the protein synthetic process is quite limited. Nonetheless, recent work by Nirenberg (see Nirenberg and Matthaei, 1961) has provided significant insight into the mechanism. He has shown that one can use synthetic polyribonucleotides which will act as synthetic Messengers, and the product is a synthetic polypeptide. Thus, for example, polyuridylic acid produces polyphenylalanine, while other ribonucleotide polymers or copolymers stimulate the addition of different amino acids. It is quite likely that this reaction occurs in the ribosomal particle through an interaction in which some of the purine and pyrimidine bases attached to the Transfer RNA find an appropriate site on the Messenger RNA so that there is a "reading" process which results in the correct alignment of the activated amino acids. In the case of polyuridylic acid, the Transfer RNA for phenylalanine probably has a series of adjacent adenine residues which combine with the uracil residues. After alignment, the amino acids are polymerized together starting from the amino end to form a protein chain.

A great deal of effort has been directed toward understanding the nature of this reading process, and an important quantity in this discussion is called the "coding ratio"; that is the number of nucleotides on the Messenger RNA molecule which interact with the nucleotides of a given Transfer RNA to determine the specificity of that site. Assuming that all four nucleotides are active, a pair of

nucleotides could only define 16 amino acids. Since there are about 20 major amino acids, it is believed that at least three nucleotides are needed in this process. Recent genetic experiments by Crick and his collaborators (1961) strongly suggest that the coding ratio is three. However, it is quite likely that the ratio will be determined in the very near future using systems such as those described above with synthetic Messenger RNA. In this regard it should be pointed out that although it is commonly assumed that the coding ratio is the same for all amino acids, this is not necessary, since Nature may have developed a system in which amino acids fall into different classes utilizing a different number of nucleotides for coding the different classes. However, this probably is unlikely in view of the genetic experiments cited above (Crick *et al.*, 1961).

We have mentioned the possibility that Messenger RNA may be made *in vivo* as complementary copies of one or both strands of DNA. If both strands are active, then the DNA would produce two RNA strands which are complementary to each other. Only one of these might be active in protein synthesis, and the other strand might be a component of the control or regulatory system. However, if both strands are coded for protein synthesis, we are faced with the possibility that proteins are manufactured in pairs related to each other by the two complementary Messenger RNA strands which are believed to determine amino acid sequence in the ribosomal particle. An alternative possibility is that the two complementary strands *both* synthesize the same protein; this could exist if there were a special type of degeneracy in the code. A given amino acid might have more than one triplet of nucleotides as its code letters, but they would have to be complementary. Thus, leucine might be AUG as well as CAU where the letters stand for the nucleotides in Transfer RNA which determine the code for leucine. Under such special conditions, complementary strands of Messenger RNA would synthesize identical proteins. It is quite likely that we will soon be able to answer such questions concerning the code through the use of synthetic Messenger molecules.

Using this brief outline, we can draw attention to a few generalities about nucleic acid function. First of all, the most important operation is the "reading" of one nucleic acid unit by another, which arises out of the specificity of the purine-pyrimidine interactions. Although many different types of hydrogen bonding interactions are possible among purines and pyrimidines (Rich, 1959), the most

important of these appears to be the pairs which are seen in DNA. Thus, the AT (or AU) pairing (Fig. 2a) and GC pairing (Fig. 2b) provide the specificity in DNA replication, RNA production, and probably are also used in the alignment of the activated amino acids on Messenger RNA during protein production. This is the most characteristic feature of the nucleic acids, and the stereochemistry of the molecules is especially well adapted for this type of interaction. The flat, unsaturated purine and pyrimidine rings interact with considerable van der Waals stabilization when they are stacked on top of each other with their flat surfaces opposed. The sugar-phosphate backbone has a preferred configuration such that the molecules can form helical aggregates around the centrally situated pile of purines and pyrimidines. Finally the detailed system of two or three hydrogen bonds between the purines and pyrimidines have sufficient specificity to determine the choice of a complementary base.

Information-containing molecules may be defined as polymers in which the sequence of residues is not random or regular, but rather has a specific and often complex pattern. All but a few of the proteins would fall in this class, and many of the nucleic acids but probably not all of them. Thus, ribonucleotide polymers which may be formed in bacterial cells through the action of polynucleotide phosphorylase probably have sequences which are close to random, since they are not made from a template. Possibly one can make a hypothetical generalization, namely that *all* information-containing polymers in biology are made by transferring sequence information from the nucleic acids. It is probable that this is true for the proteins, and perhaps this hypothesis may prove useful in understanding some aspects of polysaccharide synthesis. Several different types of sugar molecules are found in polysaccharides, and some of these polymers may contain information in the sense that they are not random or regular copolymers. This may be true for some of the cell wall polysaccharides or the blood group-specific polysaccharides. It may be that further investigation will disclose an RNA-dependent polymerization mechanism not too unlike that seen in protein synthesis.

We can use the term "information transfer reactions" to include those molecular interactions which are important in determining the sequence of information-containing polymers. According to the outline above, all of the information transfer reactions discovered up to the present involve the interaction of nucleic acids with other

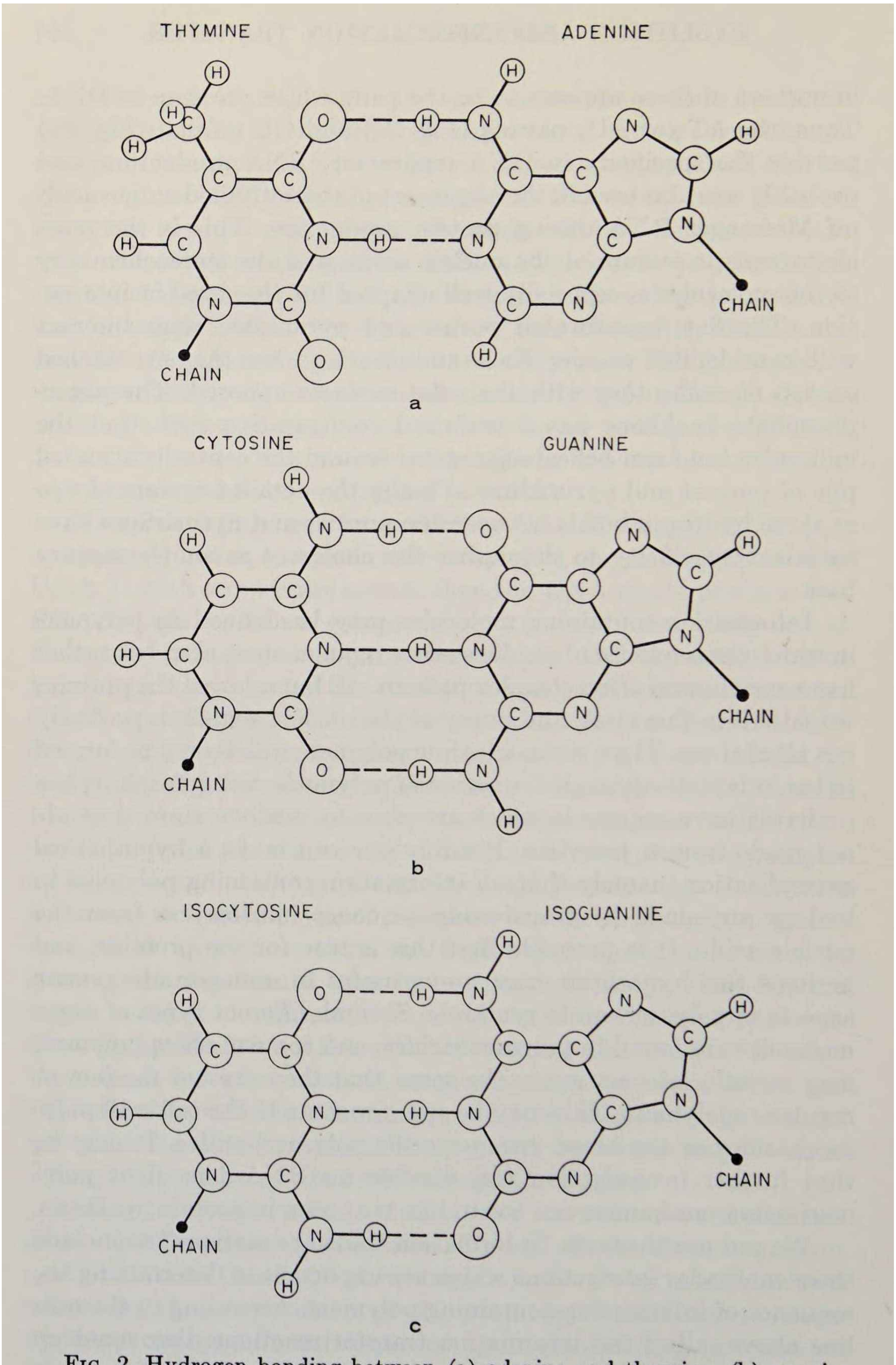

FIG. 2. Hydrogen bonding between (a) adenine and thymine, (b) guanine and cytosine, and (c) isoguanine and isocytosine.

nucleic acids or their components. This statement has one outstanding exception, namely the interaction of the Transfer RNA with its specific amino acid on the activating enzyme. This reaction does not necessarily involve a nucleic acid-nucleic acid interaction, and it is, of course, crucial in determining the correct sequence of amino acids within protein molecules.

III. Chemical Evolution and the Origin of Life

There are two levels on which one can begin to discuss the problem of the origin of life. The primary level consists of trying to understand the natural phenomena which produced a variety of molecular species during the early years in the evolution of our planet. Recently, several experimental approaches to this problem have been made by Miller and Urey (1959), and others. In general these experiments are designed to study the formation of biologically important substances such as amino acids in a primitive reducing atmosphere which is believed to have existed in the early stages of the evolution of the earth. Under the influence of an electric discharge through such an atmosphere, a variety of amino acids have been isolated. These experiments thus point to a particular route or reaction mechanism which would have made possible the accumulation of a large variety of organic molecules which gradually increased in concentration in a primitive sea.

During this period of what has been termed "chemical evolution," there would be a steadily increasing number of complex molecules, because the accumulation of simpler substances, such as some of the more stable amino acids, means that they could then become reactants leading to the formation of other organic substances. As yet there has been only limited experimental work in studying possible routes for the synthesis of nucleotide units by means of such reactions. However, recent experiments have demonstrated that primitive atmospheres are also capable of producing adenine (Oro, 1961) or uracil (Fox and Harada, 1961). Thus it is not unreasonable to believe that molecules as complex as nucleotides could have been made by nonenzymatic processes which utilize either radiation energy from the sun or electric discharges in a primitive atmosphere and it is likely that they will be observed with further experimentation. For the purpose of this discussion, it will be assumed that primitive seas some three billion years ago contained molecules of this complexity at a stage which antedates a

molecular organization which we would call "living" by present-day standards.

The second level to be considered in the origin of life is the development of replicating, information-containing polymer molecules. What system one prefers to call "living" is, of course, arbitrary, but perhaps a reasonable description of a primitive "living" system might be one in which there is an autocatalytic replication of an information-containing polymer which utilizes monomeric components from the environment and incorporates them into replicas of itself. This system, once started, would then be subject to all the modifying influences of the environment and we could call this the beginning of life as we know it.

There are two characteristic types of large molecules which are found in living systems, the proteins and the nucleic acids, and they can each in a sense serve as the basis for a theory concerning the origin of living systems. In each case, one imagines that energy derived from the sun, either in terms of radiation or electric discharges serves directly or indirectly to create conditions which bring about the polymerization of amino acids or nucleotides to form polymer molecules. The events which might happen subsequent to the formation of either primitive proteins or nucleic acids are quite distinct and we can examine them separately.

IV. Did Life Originate with Protein Molecules?

According to this theory emphasis is placed on the importance of the catalytic role played by primitive protein molecules in creating the necessary environment for self-replication. It is assumed that polypeptides are made by a random polymerization of amino acids and among these are found a few molecules which are catalytic and can act as enzymes in assembling the nucleotide units which we assume to be present in the primitive sea. At the same time another enzyme catalyst is formed, also by random assembly, which facilitates the replication of the primitive polynucleotide chain. Thus it acts as a primitive nucleic acid polymerase, organizing complementary mononucleotides along the chain leading to the formation of a two-stranded molecule. This can be followed by a separation of the two individual strands so that the process can continue. However, such a system has not yet coupled the synthesis of protein enzymes with the informational content of the nucleic acid, and this cannot be accomplished by a random amino acid polymerization.

Here is the greatest weakness in a theory of this type, since it does not really explain this most difficult step, the evolution of nucleic acid-controlled protein synthesis. However, another type of theory can be described as the basis for the origin of life which places the nucleic acids in a more central role.

V. Polynucleotides as the Origin of Living Systems

Theories of the biochemical origin of life on this planet were seriously developed about 20 to 30 years ago. At that time, it was quite clear that proteins were the most characteristic molecules in living systems and that their specific catalytic properties were essential to the functioning of biochemical systems. Accordingly much attention was devoted to ways of developing primitive protein molecules through nonliving agents. However, the almost explosive development of our understanding about the role of the nucleic acids during the past decade has made it imperative for us to reformulate theories about the origin of life in order to place the nucleic acids in proper perspective. As discussed above, the sequence of amino acids in proteins is, in a sense, a derivative of the sequence information encoded in the order of nucleotides in some part of the nucleic acids. Accordingly, it may be more reasonable to consider a theory of the origin of life in which the nucleic acids were developed as the primary agents.

Here we imagine a large number of nucleotide monomers among other molecules which are floating freely in a primitive sea, having been created by chemical reactions promoted by the products of ultraviolet radiation or electric discharges. We postulate that these nucleotide units can be assembled in random chains to make primitive, single-stranded polynucleotides in the absence of any protein catalysis. This is not a completely absurd hypothesis since we are able to do this chemically at the present time, although under somewhat special conditions. Methods for producing polynucleotide chains have been developed recently by Khorana and his associates (see Tener *et al.*, 1958) as well as by Schramm *et al.* (1961). Both of these polymerizations depend upon the use of strong dehydrating agents in a nonaqueous medium. In the presence of the dehydrating agents, water is split out of the nucleotides and they are joined together to form polynucleotides. In the case of the synthesis by Khorana, molecules of molecular weight up to five thousand or so have been formed while Schramm's recent synthesis has yielded

macromolecules of molecular weight greater than fifty thousand. Further experiments of this type should tell us more about the conditions under which such a polymerization occurs. However, we now know that it is possible to make polynucleotide molecules in the absence of a catalytically active protein.

The next step is the development of replicating nucleic acids. We postulate that the primitive polynucleotide chains are able to act as a template or as a somewhat inefficient catalyst for promoting the polymerization of the complementary nucleotide residues to build up an initial two-stranded molecule. This is in a sense the same system as that whereby a single DNA strand is able to make new RNA or DNA strands by the stepwise addition of complementary nucleotides which are subsequently polymerized together. However, we postulate that this may happen in a primitive environment in the absence of protein catalysts. We might ask if we have any evidence that a reaction of this type could occur. It would be of great interest if we could demonstrate experimentally that single polynucleotide chains have a tendency to bind or act as condensing sites for their complementary bases or nucleotides. Experiments of this type could be carried out at the present time. However, a very interesting effect was noted recently by Schramm *et al.* (1961) in the chemical polymerization of polyuridylic acid. He found that the addition of polyadenylic acid increased the rate of polymerization of polyuridylic acid over tenfold. This suggests that the polyadenylic acid is acting as a condensing site for the complementary uridylic acid residues which are then polymerized more rapidly. Again, further experiments could be carried out on this type of nonenzymatic polynucleotide catalysis to determine the limits of specificity.

What is being pointed out here is an important difference between proteins and nucleic acids. There are significant stereochemical reasons why the polynucleotides can act as their own catalysts for self-replication. However, there are no analogous reasons for believing the polyamino acids have this ability to reproduce themselves.

At this stage we imagine a system which contains nucleotide monomers together with polymers which, under some conditions, are able to make two-stranded versions of themselves using a system of complementary hydrogen bonds. This nonenzymatic replicating system would undoubtedly be very inefficient and slow. It is possible that the two-stranded form of the nucleic acid might then be thermally denatured by heat from the sun and separated into

single strands again. The process can then continue again with the manufacture of an increasingly larger number of nucleic acid chains. The concentration of these nucleic acid polymers might become appreciable since the rate of chain destruction might be as slow as the rate of chain production.

Now we must consider a possible modification of this primitive chain replicating action. We imagine that the mononucleotide units which are about to be incorporated into a polynucleotide chain occasionally have other small molecules attached to them. Thus, for example, if the oncoming nucleotide unit is a ribose nucleotide, the 3′- and 5′-hydroxyl groups might be used in polymerization, but there is an additional hydroxyl group found at position 2′. It might have attached to it another type of molecule which would then remain attached to the growing polynucleotide chain. In particular this residue might be an organic acid linked to the hydroxyl group by an ester linkage. It might also be an amino acid attached through the same linkage. At this stage we theorize this would be the beginning of a prototype system in which the polymerization of a nucleic acid molecule is coupled with the assembly of a series of amino acids. The amino acids might then be subsequently polymerized once they are organized in a linear assembly by their attachment to the adjoining nucleotides. In a sense it is this juncture in any theory of the origin of life which will present the greatest difficulty, since in order to develop the type of information-transferring system seen in biochemical systems, we must have a coupling of the nucleotide sequence with an amino acid sequence in proteins. However, to do this we must evolve specificity in amino acid sequence in order to develop those catalytic functions which make the system operate. In short we must try to develop a method for evolving the activating enzyme function.

Let us explore this point in more detail. We could imagine that a random polynucleotide chain, acting as a condensing agent and randomly polymerizing amino acids, could produce a molecule which specifically attaches glutamic acid to a particular nucleotide such as the adenine unit (or to a triplet of nucleotides, or even larger). The chance development of this catalytic molecule might give rise to large numbers of adenylic-glutamic units, but these would not aid in the recreation of the original activating enzyme, since the latter was produced by a random association of amino acids with nucleotides. Hence, such a mechanism could not "go critical" and

lead to the synthesis of specific protein molecules governed by the polynucleotide sequence. Thus we must look to alternative methods for developing an activating enzyme.

Let us imagine one way in which this could have developed. For example, the initial form of this contaminated nucleic acid synthesis might have individual amino acids attached to individual nucleotides. Thus the coding ratio for such a primitive system is one. If four nucleotides were used in this primitive system, then such a mechanism could produce only random polypeptide chains if any amino acid could be attached to a particular nucleotide. However, we might imagine conditions in this primitive environment in which there was some selectivity. For example, the adenine nucleotide might have a preference for attachment to glutamic acid, and so on for the other three bases. This might lead to the development of primitive polypeptide chains which have predominantly four different amino acids, namely those four which have the greatest preference for attachment to the four individual nucleotides. In such a hypothesis, the basis of selectivity must lie with some intrinsic stereochemical property or reactivity of the given amino acid with a given nucleotide in the absence of protein catalysts. At the present time there is no information about specificity of this type. It would be of interest to see whether there are any affinities among the different nucleotides toward certain of the amino acids. For example, are certain amino-acyl nucleates formed preferentially in a chemical synthesis? This problem could be subjected to investigation in the laboratory at the present time. Alternatively, we cannot overlook the possibility that other components of the primitive environment, such as the surface of minerals or clays, could serve in the role of a selective catalyst.

The first important step in the development of this system must be the creation of prototype polypeptide catalysts which act as activating enzymes to reinforce the crude selectivity which prevailed initially, as in combining adenylic acid with glutamic acid in the example cited above. Once such primitive activating enzymes appear, the system starts to exhibit specificity, and there begins to be a relation between nucleotide sequence and amino acid sequence.

It is very difficult to overestimate the pivotal role of the activating enzymes in any evolutionary theory. These are the molecules which relate polynucleotides to polyamino acids and the entire mechanism of specific synthesis hinges on their evolutionary devel-

opment. These enzymes had to be among the first made in an evolving biochemical system; before they appeared there would only have been random polyamino acid polymers.

To get the nucleic acid-protein system fully coupled, we imagine that a small number of randomly arranged polynucleotide chains begin to form polypeptide chains which act as activating enzymes, so that the nucleic acids are "read" in a unique manner. If the initial coding ratio were 1, then four such enzymes would be necessary if four nucleotides were present in this primitive nucleic acid. If the primitive nucleic acid were composed of only 2 nucleotides, as discussed below, then only two activating enzymes would be necessary. This is a minimal number, since more than one enzyme is needed to obtain specificity.

Once such a system is developing specific amino acid sequences which are coupled to nucleic acid sequences, then we may consider that "life" has started. Some of these primitive nucleic acid chains may have a nucleotide sequence which leads to a polymerase enzyme, and the production of that type of nucleic acid would be stimulated preferentially over others; in short, natural selection on the molecular level can commence. If the primitive coding ratio were 1, we can see that considerable advantages would develop with a mutant in which nucleotides were taken as pairs (or triplets) in the activating step, since they could then incorporate other amino acids and develop polypeptide chains of greater chemical subtlety.

The major thesis in this development is the idea that what we may call "life" started with the coupling of nucleic acid polymerization and amino acid polymerization. This process made an important step forward when specificity of protein synthesis was developed in the system through the discovery of activating enzymes. However, the prototype of this reaction may have been in a form quite different from that which we observe today.

VI. The Trend Toward Increased Complexity

The study of evolution on the molecular level is still very new. Only in recent years have we developed the tools which are necessary to determine the sequence of amino acid residues in the proteins. Thus we can now begin to study their molecular evolution, and it is quite likely that this type of effort will be continued for many years. However, at the present time, we do not have guide lines for our thinking, that is, general concepts about molecular evolution

which may prove of utility in understanding and interpreting the results.

We have a considerable body of information about macroscopic evolution in terms of the changes, with time, of organisms and the development of new species. From this we see that the mechanism of evolution is such that slow refinements are usually made within a given organism in relation to its environment which lead toward increased efficiency of reproduction. Occasionally there are discontinuities which are opportunistically utilized in the exploitation of a new environment or of an altered physiological function. In general, the trend in macroscopic organic evolution is toward increased complexity. We can trace the development of increasingly subtle physiological functions and greater control leading to a progressive extension of the organism's ability to utilize the environment and maximize reproductive ability.

It is reasonable to ask whether a similar principle applies on the molecular level. Thus, has there been an increase in the complexity of molecular organization and function during the course of evolution? It is perhaps not unreasonable to believe that this may have occurred and accordingly we can inquire about the various ways in which this may have developed, especially in relation to the biological information transfer system.

In the previous section we described a primitive, self-replicating nucleic acid system which has started to develop a feed-back mechanism in which a modification of the replicating ability of the nucleic acids is associated with the polymerization of a polypeptide chain, which can then be liberated to form a catalytically active primitive enzyme. Once this system is able to form the activating enzymes it would then tend to become more directional in its development. Some classes of nucleic acids which, through chance, have developed a nucleotide sequence which coded for a set of functional proteins, would tend to become more numerous. Perhaps an initial reason for this is simply one of concentration. If a given nucleic acid started producing a nucleic acid polymerase enzyme in addition to its activating enzymes, larger amounts of this type of nucleic acid would be produced because of the higher concentration of the enzyme in its vicinity.

There are several parts of this molecular system in which increasing complexity could manifest itself. We can list these and discuss possible historical alterations which may have occurred in the

system. Such a listing may be of some value in that it may help to interpret data on biochemical evolution as it is uncovered. In addition, we are on the threshold of space exploration and we may have an opportunity to experimentally study more primitive life forms.

VII. Changes in the Composition of the Nucleic Acids

We may ask why the nucleic acids have four units in them at the present time (for this discussion we will equate thymine and uracil, and ignore the difference between sugars). In order to contain information, it is obvious that two bases would be enough; for example, simply adenine and thymine. A primitive organism whose nucleic acid contained only two complementary bases could still develop a similar type of biochemical system for information transfer. The sequence of the adenine and thymine residues along one chain carries information in many ways analogous to the dot-dash system used in the Morse code. In such a hypothetical organism, if one had to code for twenty amino acids, it would require five bases in order to obtain a minimum of twenty combinations. Nonetheless, in principle, a system of this type would be functional.

Is it possible that there may have existed an early form of life which used nucleic acids with only two bases? For example, it has been shown by Kornberg and his collaborators that the DNA polymerase enzyme, when left without a primer in the presence of the nucleotide triphosphates, will make an AT copolymer. This is a DNA-like molecule with a regular repeating sequence of adenine and thymine. One is tempted to ask whether this could be an atavistic form of a nucleic acid. It is also worth noting that the crab contains a DNA molecule which is almost entirely an AT copolymer (Sueoka, 1961). In this regard it will be of great interest to study the base composition of nucleic acids as a function of phylogenetic evolution to see whether there is any significant trend. Preliminary steps have already been taken in this direction by Sueoka (1961).

One can also ask why has Nature used nucleic acids built with four bases instead of six, for example. In a more general way, one can ask whether it is possible for nucleic acids to be built out of additional complementary purine-pyrimidine pairs which would still maintain the same geometry and specificity in the hydrogen bonds such as is now seen in DNA. The answer is not particularly clear since a great deal depends upon the tautomeric form of the bases. However, if we restrict ourselves to amino rather than imino

groups and keto rather than enol groups, another possible type of complementary pair is that between isoguanine and isocytosine, as illustrated in Fig. 2c. If this is the stable tautomeric form of these two molecules, then they could also be incorporated into a nucleic acid molecule and would have the requisite specificity. That is, they would hydrogen bond only with each other but not with any of the other four bases. It is important to note that this requires that the hydrogen atom in isoguanine be located on N-1 of the ring rather than on N-3. It is not clear whether this condition is fulfilled in isoguanine. However, it is sufficient to note that there are *very few* additional complementary pairs which one can fit into the DNA structure which would have the necessary specificity. It is quite clear, however, that a nucleic acid built out of six bases using the same principle of complementarity could function suitably in a biological system. To code for twenty amino acids in such a system, it would only be necessary to have two bases to contain the necessary information, rather than three.

VIII. Has the Number of Amino Acids Increased During Evolution?

Due to the recent discovery by Nirenberg (Nirenberg and Matthaei, 1961) that we can use synthetic polyribonucleotides as a synthetic form of Messenger RNA to direct the synthesis of polypeptides, it seems quite likely that we will have the entire amino acid code deciphered very soon. The recent work by Crick and coworkers (1961) on a series of acridine orange mutants makes it quite likely that the code is a triplet, and it also points out the possibility that the code is degenerate. Thus there may be more than one triplet of nucleotides which specifies a given amino acid.

It is usually stated that there are twenty amino acids; however, this ignores the fact that several additional amino acids are found in nature but their distribution is limited. Examples are the hydroxyproline or hydroxylysine amino acids which are found in collagen and related proteins. It should be relatively easy to determine whether or not these minor amino acids have separate code letters, i.e., groups of nucleotides assigned to them. Thus, for example, if one found that hydroxyproline had a particular soluble RNA molecule to which it was attached by means of a special activating enzyme, then it would be quite likely that there is a special code letter which is unique for hydroxyproline. Similar investigations

can be extended to include the other minor amino acids. The results of these investigations may show that there are only twenty amino acids which can be incorporated into proteins through the activating enzymes. But perhaps a more likely finding is that the number twenty is not unique, and that there are code letters attached to some of the other so-called trace amino acids. This could represent one of the ways in which increasing complexity has been built into biochemical systems. Thus the number of amino acids used by biological systems to build up proteins may be a time dependent function. In a very primitive biochemical system, only a small number of amino acids may have been used and gradually, as the system evolved, finer degrees of functional control were developed through utilizing other amino acids with slightly different chemical side chains attached to them. Such a hypothesis can be examined experimentally.

It should be noted that a degenerate coding system for amino acids is well adapted for this type of evolutionary change. For example, suppose leucine were represented by four sets of nucleotide triplets. A mutational error might assign one of these triplets to a new amino acid, say hydroxyproline, but leucine would continue to be used by the organism. Accordingly the system would tend to move toward a state of less degeneracy and a wider distribution of amino acids. It is quite clear that considerable selective advantage would accrue with the development of a coding system which allows for a gradual increase in the number of amino acids which can be utilized.

It is important to recognize that the imminent decoding of the relation between the Messenger RNA and the polypeptide chain which it helps to synthesize will throw considerable light on the possibilities outlined above.

IX. An Evolutionary Increase in Nucleic Acid Content

If we survey the DNA content of a variety of organisms, we note that the mass of DNA increases with the complexity of the organisms. In a similar fashion it is not unreasonable to believe that the amount of genetic information contained by individual primitive cells may have increased through the very early stages of biochemical evolution. For example, there may have existed at one time a primitive organism containing enough nucleic acid to code for five or ten proteins. This could be possible in a specialized en-

vironment in which the concentration of many of foodstuff molecules were built up by natural processes as discussed earlier. In this system, the nucleic acid replicates in a manner similar to the replication of DNA as we now understand it. Thus there is a building up of complementary chains and the separation of the two original nucleic acid strands of the parent molecule.

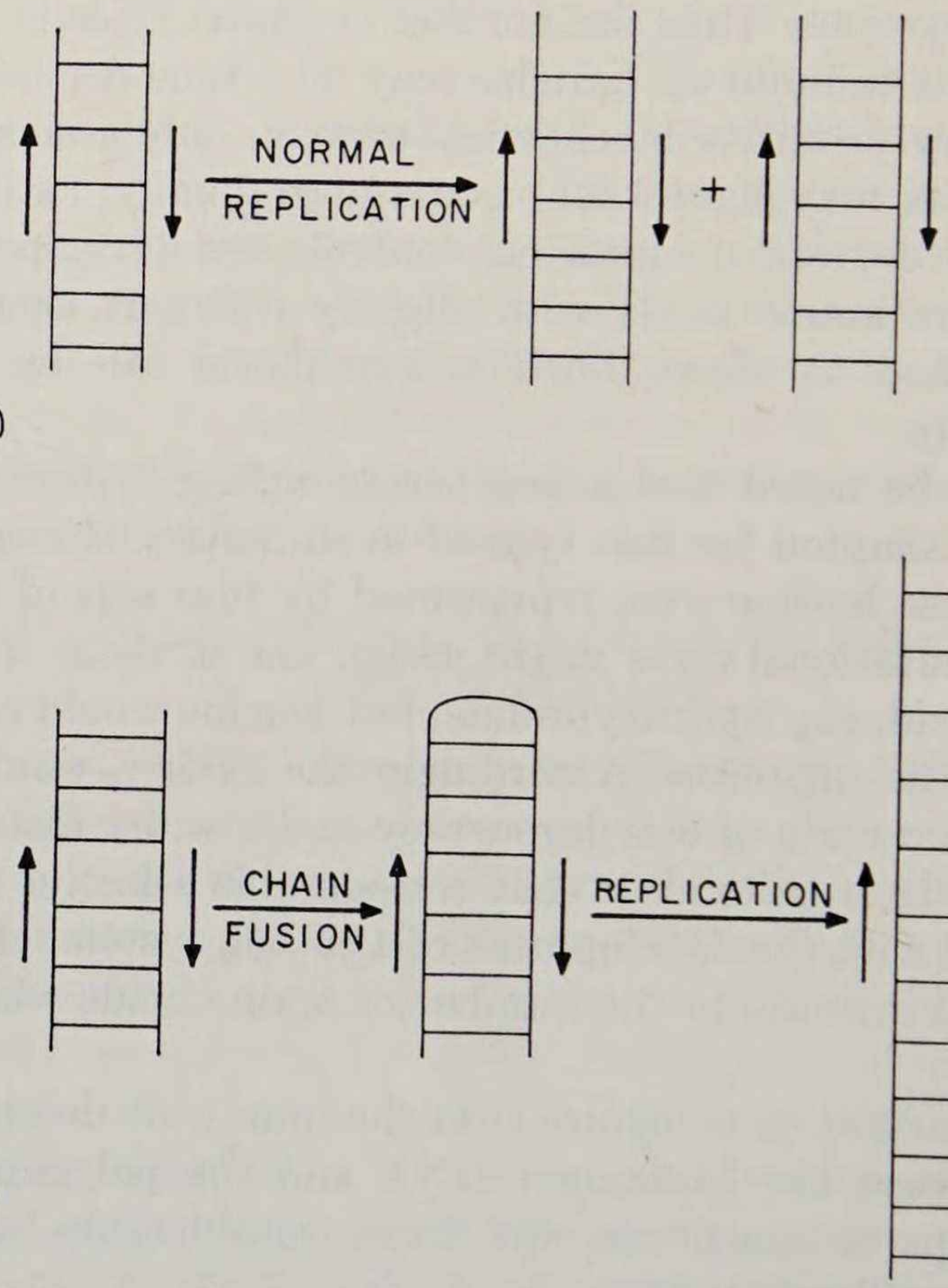

FIG. 3. (a) The replication of one nucleic acid molecule to produce two identical daughter molecules. (b) The effect on replication of joining the ends of the parent molecule.

We might ask what would happen in such a system if, under the action of a cosmic ray or other radiation, the ends of the two-stranded DNA molecule were joined together with a normal covalent bond. There would result, as shown in Fig. 3, a doubling in length of the original nucleic acid molecule. During the next cycle of replication the DNA would have twice the original length. It has the same informational content as the original nucleic acid but there

are two copies present. An accident of this type may seem to offer no selective advantage. However, during the process of mutation, alterations in these bases may occur in one of these copies but not in the second one which codes for the same protein. This means that a mutation which might be lethal for the organism would not be lethal in this duplex state. Accordingly an anomaly of this type allows for the development of a variety of new protein molecules which can, in a sense, explore the environment more readily than the original organism with only a single copy of its genetic information. Here the luxury of surplus, redundant information appears to provide a selective advantage in evolution. If this process goes on several times, eventually the organism ends up with many genetic copies of an original prototype protein molecule. These copies might then evolve along somewhat separate evolutionary lines and give rise to classes of molecules which, though similar, are in fact different in many ways. It is possible that this may be the origin of certain classes of related proteins such as myoglobin and the α- and β-chains of hemoglobin, in which there are many similarities in both amino acid sequence and secondary structure.

It might be possible to carry out experiments to determine whether or not the process outlined in Fig. 3b is possible. Thus, a homogeneous solution of DNA could be subjected to radiation, and an attempt could be made to detect a small number of longer molecules after treating with the DNA-polymerase system.

X. Why Are There Two Nucleic Acids?

It is quite remarkable that contemporary biochemical systems have two nucleic acids, DNA and RNA, which differ only by a systematic hydroxyl group and an occasional methyl group. Despite the great chemical similarities, the molecules nevertheless have quite different functions in the cell. DNA acts as the major carrier of genetic information, while the RNA molecule is used to convert this genetic information into actual protein molecules. Because of the close chemical similarities, we are tempted to ask whether they could have originated historically from a common stem nucleic acid molecule which then specialized in the course of evolution to produce the two different classes of nucleic acids which we see today. To discuss this further we should note that the RNA molecule is also able to carry genetic information, as, for example, in the RNA-containing viruses. Thus, it may be reasonable to speculate that the

hypothetical stem or parent polynucleotide molecule was initially an RNA-like polymer which was able to convey genetic information as well as organize the amino acids into a specific sequence to make proteins. This implies that the RNA polynucleotide strand had the ability to replicate itself and produce a complement in a manner somewhat similar to that which is found in DNA. It is possible that an enzyme of this type may have been observed already. By this view, DNA may be regarded as a derivative molecule which has evolved in a form such that it only carries out part of the primitive nucleic acid function. It specialized in the molecular replicating cycle that is part of the mechanism for transmitting genetic information. DNA is metabolically less reactive than RNA, perhaps because of the absence of the hydroxyl group on carbon 2. The loss of this hydroxyl group may have made it impossible for the DNA molecule to have attached to it the amino acids which are used in protein synthesis. However, considerable selective advantage may be derived from the development of two different classes of nucleic acids, one of which is less active metabolically and specializes in self-replication. In a sense, this tends to preserve the primary copy of the genetic information. It will be of considerable interest to study the available simple life forms to see whether some of them may exist with only one type of nucleic acid rather than two types. It is possible that the RNA containing viruses may be regarded as present-day examples which may have degenerated evolutionarily from such a primitive life form.

It should be noted that the changes which are discussed in many of the sections above may only have occurred during the very early stages of evolution. What we see today is the net result of a complex competition among many different types of systems.

XI. Extraterrestrial Life

We can make a few generalizations from our description of the molecular basis of life as we know it on this planet. The most characteristic feature of a living system is the fact of its enormous complexity on the molecular level. There are a large number of chemical reactions going on which involve the manufacture of many specific protein catalysts which are used as regulators for facilitating these reactions. However, the information necessary to make these large molecule catalysts is stored in polymer chains in a much more compact form than the mass of catalyst molecules themselves. Indeed

one might say that a characteristic of life is the storage of information in polymer molecules and its subsequent expression in terms of building other polymer molecules which have specific functions, such as making catalysts, cell walls, and other specialized structures.

We are now on the verge of being able to carry out an exploration of space to look into the possibility of life existing on other planets or, eventually, other stellar systems. If life exists in these other sites, it is possible that it may exist in quite a different form from our own. Thus, for example, we could imagine that a living system evolving in a somewhat different environment, may have selected different classes of molecules to be built into polymers both for storing genetic information as well as for carrying out the large number of necessary chemical reactions. Indeed the molecular basis of other life forms may be entirely distinct from ours. However, it is quite likely that there would be one major feature in common with terrestrial life, namely, the use of polymer molecules to store the large amount of information which is needed to adequately define the complexity of events which go on in a system which we call "living." However, the polymers might be quite different and the molecular composition of other living systems might bear no further resemblance to terrestrial life. In particular, the nucleic acids which play such a central role in transmitting information in our form of life, may not be found at all in other forms of life but may be replaced by a different type of polymer.

XII. Conclusions

In this essay the essentials of biochemical information transfer have been outlined briefly. An attempt has been made to discuss various problems which arise in trying to understand how this information transferring system may have evolved. Since we have an almost complete lack of factual information concerning such events, an enterprise such as this must, of necessity, be speculative. Nonetheless some of these reflections may serve as a stimulus for additional experimental approaches which will enable us to place biochemical evolution on a firmer foundation.

References

Bessman, M. J., Lehman, I. R., Simms, E. S., and Kornberg, A. (1958). *J. Biol. Chem.* **233,** 171.

Brenner, S., Jacob, F., and Meselson, M. (1961). *Nature* **190,** 576.

Crick, F. H. C., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). *Nature* **192,** 1227.
Fox, S. W., and Harada, K. (1961). *Science* **133,** 1923.
Gros, F., Hiatt, H., Gilbert, W., Kurland, C. G., Rissebrough, R. W., and Watson, J. D. (1961). *Nature* **190,** 581.
Hurwitz, J., Bresler, A., and Dringer, R. (1960). *Biochem. Biophys. Research Communs.* **3,** 15.
Meselson, M., and Stahl, F. W. (1958). *Proc. Natl. Acad. Sci. U.S.* **44,** 671.
Miller, S. L., and Urey, H. C. (1959). *Science* **130,** 245.
Nirenberg, M. W., and Matthaei, J. H. (1961). *Proc. Natl. Acad. Sci. U.S.* **47,** 1588.
Rich, A. (1959). *Rev. Modern Phys.* **31,** 191.
Schram, G., Grötsch, H., and Pollmann, W. (1961). *Angew. Chem.* **73,** 619.
Sueoka, N. (1961). *J. Mol. Biol.* **3,** 31.
Tener, G. M., Khorana, H. G., Markham, R., and Pol, E. H. (1958). *J. Am. Chem. Soc.* **80,** 6223.
Weiss, S. B. (1960). *Proc. Natl. Acad. Sci. U.S.* **46,** 1020.