May 14, 1957

How Do Amino Acids Read the Code?

By Leo Szilard

The Enrico Fermi Institute for Nuclear Studies
The University of Chicago, Chicago, Illinois

It is generally believed that proteins are synthetized by RNA (ribonucleic acid), and that the sequence of purines and pyrimidines; i.e., adenin, uracil, guanin, and citosin determines the sequence of the amino acids in the polypeptide. It is presumed that these purines and pyrimidines form the basis of a three-letter code. Each of these bases represents one letter of the code, and certain groups of three letters form the word which corresponds to one specific amino acid. For any such code word to which we shall affix the plus sign, we can construct the complementary combination (the anti-code word to which we shall affix a minus sign) by substituting uracil for adenin and adenin for uracil, citosin for guanin and guanin for citosin.

It has remained so far a complete mystery in just what way an amino acid can read such a code. There was no indication up to now what the nature of the chemical affinity might be that may line up the various amino acids - A_i, A_j, A_k, A_l, etc. - in the proper sequence alongside the template that contains the three-letter code.

It is the purpose of the present paper to indicate a conceptually simple way in which this might be accomplished by the living cell.

We shall assume that the cell contains twenty enzymes (or enzyme systems) which couple each amino acid to a trinucleotide that represents its code word. We shall further assume that one strand of RNA, which may be contained in a double stranded helix is composed of a sequence of anti-code words. We shall

It is the purpose of the present paper to indicate a conceptually simple scheme which might be resolved.

We shall assume that the cytoplasm of the cell contains 20 enzymes (or enzyme systems), each of which couples one of the 20 amino acids to the proper trinucleotide (ribose that we shall designate with (+), and that represents the code which is complementary to the code contained in a RNA (1/2 RNA(-) template. To these trinucleotides we shall affix the (+) sign in contradistinction to the trinucleotides contained in a RNA template that carries the code which is specific

C O R R E C T I O N

We shall assume that the cell (presumably the cytoplasm) contains 20 enzymes, E, or enzyme systems which couple each amino acids to a specific nucleotide that represents the anti-code word for the particular amino acid. To these trinucleotides, we shall attach a (+) sign. To the complementary code words; i.e. to the trinucleotides which are contained in the RNA template alongside of which the proteins are formed, we shall attach the (-) sign. We assume that each amino acid is coupled to the phosphate group on the fifth carbon atom of the "leading" nucleotide of the tri-

nucleotide (as an alternative, we shall also consider the possibility that we have in place of the trinucleotide monophosphoamino acid, a trinucleotide diphosphoamino acid). This is a high energy bond representing 12,000 calories.

According to the notions here presented, a specific protein may be formed by a specific ribosenucleic acid template(+) in the following manner:

The trinucleotides ~~which~~ each carrying the proper amino acid will diffuse in the cytoplasm and each trinucleotide will reversibly attach by means of hydrogen bonds to the complementary ribose trinucleotide contained in the 1/2 RNA(+) template. If the protein to be formed has a ~~molecular weight of 100,000 and~~ contains 1,000 amino acid residues, then there will be 1,000 trinucleotides lined up alongside the template -- which we shall designate for our present purposes as a paragene. After a certain time the paragene will be completely covered by the proper trinucleotides(+) and somehow a chemical reaction ~~will be~~ may then be triggered. In this reaction the acid anhydride bonds split and the adjacent amino acids ~~will be~~ linked with each other by peptide bonds. Thus a polypeptide with a specific sequence of amino acids, A_i, A_j, A_k, and A_l, determined by the paragene ~~will~~ may be formed ~~which may be~~ and folded in a manner as yet unknown ~~and will~~ to yield a protein, for instance, ~~the~~ perhaps an enzyme for which this particular paragene is specific.

~~Under revision~~

we must now compute at what rate a single paragene can form the corresponding enzyme. We will do so on the assumption that the binding energy between two trinucleotides, which are not complementary to each other, is somewhat smaller than the binding energy between two nucleotides, which are complementary to each other, and that therefore we can assume that we do not have to take into account the fact that the trinucleotides within the template will be occasionally covered for a short period of time by the wrong trinucleotide; i.e. one which is not complementary to it. We shall assume for the sake of argument that the molar concentration for the ribose-trinucleotides, which are present in the cytoplasm of the cell, is the same for each of the 20 different kinds of trinucleotides, and we shall designate the concentration of these trinucleotides(moles per liter) the rate at which a single trinucleotide within the template is headed by a ribose-trinucleotide(+) in just the right position and with sufficient energy to reversibly combine with it. For this rate we shall write: hit rate

$$\tau_h = \frac{1}{A\rho} \qquad (1)$$

and for the rate at which such trinucleotides will dissociate off from the template, we may write

$$\tau_{AK} = \frac{1}{2AK} \qquad AK = \frac{1}{2}\frac{1}{\tau_{av}}\left(=\frac{1}{2}10^{13}e^{-\frac{Q}{RT}}\right) \qquad (2)$$

where K is the equilibrium constant for the reaction between one ribose-trinucleotide(+) in the solution or between one ribose-trinucleotide(-) on the template. From this we may compute the maximum rate at which a protein molecule can be formed on the assumption that there are m trinucleotides on the template; i.e. corresponding to m amino acid residues contained in the protein for which the template is specific. As may be seen, if we make the binding energy too small; i.e. K too large, we shall have a low value of

$$\frac{2m}{1+X} + \frac{1}{X}$$

$$-\frac{2m}{(1+X)^2} + \frac{1}{X^2}$$

we must now compute at what rate a single parageneon form the corresponding enzyme. We will do so on the assumption that the binding energy between two trinucleotides, which are not complementary to each other, is somewhat smaller than the binding energy between two nucleotides which are complementary to each other, and that therefore we can assume that we do not have to take into account the fact that the trinucleotides within the template will be occasionally covered for a short period of time by the wrong trinucleotide; i.e. one which is not complementary to it. We shall assume for the sake of argument that the molar concentration for the ribose-trinucleotides, which are present in the cytoplasm of the cell, is the same for each of the 20 different kinds of trinucleotides, and we shall designate the concentration of these trinucleotides(moles per liter) the rate at which a single trinucleotide within the template is headed by a ribose-trinucleotide(+) in just the right position and with sufficient energy to reversibly combine with it. For this rate we shall write: hit rate

$$tk = \frac{}{+p}$$ (1)

and for the rate at which such trinucleotides will dissociate off from the template, we may write

$$Ak = \frac{1}{t} \frac{2}{Cm}$$ (2)

where K is the equilibrium constant for the reaction between one ribose-trinucleotide(+) in the solution or between one ribose-trinucleotide(-) on the template. From this we may compute the maximum rate at which a protein molecule can be formed on the assumption that there are m trinucleotides on the template; i.e. corresponding to m amino acid residues contained in the protein for which the template is specific. As may be seen, if we make the binding energy too small; i.e. K too large, we shall have a low value of

$$\bar{T} = \frac{1}{\rho A}\left(1 + \ln m + \frac{m'\rho}{1+\frac{\rho}{K_1}}\right) + \frac{3}{\tau_{ew}}$$ 5-A or

production, and if we make the binding too ~~small~~ large; ie., K too small, we also have a low rate of protein production. /Protein production The maximum on the template is obtained if we have $(K_1 = K_2$

$$\left(\frac{K}{\rho}\right)^2 - 2\frac{K}{\rho} - 1 = 0 \qquad (3)$$

and/this relationship is fulfilled, then we may write if

$$T_{comp} \qquad (4)$$

For the value of A we can write

(5)

In this formula, small p is a frantion of the collisions between the ribose-trinucleotide(+) and the target area Σ of a given ribose-trinucleotide on the template and a given complementary ribose-trinucleotide(-) on the template.

Assuming now that the concentration of each ribose-trinucleotide(+) in the cytoplasm that carries a given amino acid is of the order of magnitude of , and assuming for Σ a value of and for p a value of , we find that a single paragene can produce protein at the rate of 2,000 protein molecules in about 30 minutes. The equilibrium constant, K, we compute from equation (4), which gives a values of This equilibrium constant corresponds to an evaporation rate of nd the corresponding binding energy comes out to be about calories. To each factor of 10 by which the equilibrium constant is increased, there corresponds about 1400 calories by which the binding energy is decreased. As we shall presently see, the explanation of why DNA contains thymin instead of uracil which is contained in RNA might be due to the surmised fact that the binding energy due to the hydrogen bonding between uracil and adenin is

or $$\bar{T} = \frac{1}{A\rho\Sigma}\left\{1 + \ln m + \frac{m}{1+\frac{\rho}{K_1}} + \frac{1}{2\frac{K_2}{\rho}}\right\}$$

RNA ~ DNA

~~higher than between thymin and adenin~~. We may presume that the RNA(-) strand which forms the tem-plates ~~which form~~ for the production of the protein~~s~~ ~~are~~ is produced by ~~the~~ a DNA(+) ~~template~~ strand, ~~the plus~~ and the same DNA(+) template ~~may be presumed to~~ is either on a produce the DNA(-) template. A number of different schemes may be devised to show how strands of RNA may produce strands of DNA, and how strands of DNA may produce strands of RNA. All of these schemes, however, have one thing in common. The DNA template must produce both the DNA template and the RNA template. This raises the problem of a possible mix-up in the formation of a composite template from the ribose-trinucleotides and the desoxyribose-nucleotides. It is tempting to speculate that the difficulty here mentioned is resolved in the following fashion: DNA may be produced on the RNA template from ribose-trinucleotides which carry on the 5-carbon atom of, say, the leading trinucleotide a hydrogen phosphate, while in contrast to this, the DNA strand is formed on the same DNA template from desoxyribose-trinucleotides which carry a hydrogen phosphate bond in the 5-carbon position, not on the leading but on the trailing nucleotide. The RNA strand may then be synthetized on the DNA template as follows: If the ribose-trinucleotide(-) combines with the trinucleotide, which is in a No. 1 position next to the head of the template, it will attach itself by splitting off one phosphate from the head of the template; if it combines with any other position, even though the trinucleotide with which it conbines is complementary to it, it will evaporate within a very short time. We assume here that the equilibrium constants for the combination of a RNA(+) trinucleotide that contains adenin with the complementary desoxyribose-trinucleotide that contains thymin is perhaps 10 ~~xxx~~ times higher than the equilibrium constant for a combination of a trinucleotide and its complementary trinucleotide which contained uracil rather than thymin.

(many schemes)

If a trinucleotide is attached in the No. 1 position, a trinucleotide that reversibly combines with the trinucleotide in the No. 2 position will have a good chance to combine with the trinucleotide attached to the No. 1 position. So the positions will be occupied one by one in the order of their serial number until m (1000) trinucleotides are all aligned and linked to an RNA(-) strand by phosphate bonds. In each case one phosphorus is split off for each trinucleotide attached to the RNA strand that is being formed. The time required for the formation of such an RNA strand is given by

and the time required for its evaporation is given by

If we now assume that the concentration of the ribose-trinucleotides(-), from which the RNA(-) strand is formed, is about the same as the concentration of the ribose-trinucleotides(+) which form the proteins, and if we choose the equilibrium constant, K, to give the shortest possible time, we find that the time it takes for DNA strand(+) to make RNA Strand(-) is about

The optimum K value is higher by a factor of than the optimum K value for a maximum rate of protein formation and the corresponding binding energy is lower by calories.

We have assumed above that the cytoplasm of the cell contains enzymes which form ribosetrinucleotides, and coupled to each such ribose-trinucleotide through a high energy bond the proper amino acid. We shall further assume that these ribosetrinucleotides, which carry amino acid, are condensed on an RNA(-) template, the paragene, which is contained in the cytoplasm.

We shall now discuss in what manner the RNA(-) template may be formed within the nucleus on a DNA(+) strand. It is tempting to postulate that the RNA(-) strand formed inside the DNA(-) strand formed inside the DNA(-) template is formed through the condensation of ribose-trinucleotides(+)

one of the three trinucleotides again as a high energy phosphate bond either on the 3-carbon position or on the 5-carbon position. We shall call this nucleotide the leading nucleotide of the trinucleotides in contradistinction to the trailing nucleotide of the trinucleotide and the center nucleotide of the trinucleotide. For the sake of easier communication, we shall specifically assume without restricting the general validity of our discussion that the DNA template is a head, and that when the ribose-trinucleotides are lined up alongside the DNA(+) template, then counting the letters of the code, starting at the head of the template, the high energy phosphate bonds of the ribose-trinucleotides may be seen attached to the 5-carbon position of the first letter of the code in every one of the trinucleotides. We shall assume that the RNA(-) template is synthetized alongside the DNA(+) template in the following manner: a ribose-trinucleotide which reversibly combines with the complementary code word on the DNA(-) template will, in general, evaporate, and the rate of evaporation is given by

In this expression K_1 designates the equilibrium constant of the desoxyribose-trinucleotide - ribose-trinucleotide complex. As may be seen later, the value

of K_1 may be assumed to be higher than the value of K_o for the binding energy of thymin to adenin due to hydrogen bonding being lower than the corresponding binding energy of the triribonucleotide(+)-triribonucleotide(-) complex that is formed in the case of protein synthesis. Thus the ribose-trinucleotide(+) will dissociate off after a short while from the DNA(+) template unless it is permitted to speak of a phosphate, and use the energies here liberated to form a chemical bond either with the head of the template or with an adjacent -- already chemically bound triribonucleotide. Synthesis of the RNA(-) template will accordingly take place as follows: When the first position next to the head of the template is filled by a triribonucleotide(-), the high energy phosphor group of the hydrogen group will split off, a phosphate and ehcmically bind to the head of the template. When position 2, the adjacent position, is filled with an RNA(-) trinucleotide, the high energy phosphorus will split off from the hydrogen phosphate group of this nucleotide, and the nucleotide will bind to the nucleotide which occupies the No. 1 position on the DNA(+) template. Thus in succession, position after position adjacent to a triribonucleotide that is already chemically bound to its neighbor will be filled with chemically bound

C o rre c t i o n -- phosphate linked

In this manner m ribotrinucleotides will be linked to form an RNA(-) strand. The time required for this process is given by

The RNA strand will in time detach itself from the DNA template on which it was formed. We presume that the bond tying the first triribonucleotide to the head of the template might be enzymatically broken, and that thereafter in succession the ribonucleotides(-), Nos. 1, 2, 3, etc., may dissociate off from the complenentary tridesoxyribonucleotide contained in the

RNA(+) template. Thw time required for this process is given by

The total time needed for formation and dissociation of one strand of RNA is given by

...n with 2 copies)

May 29, 1957

HOW DO AMINO ACIDS READ THE CODE?

by Leo Szilard

The Enrico Fermi Institute for Nuclear Studies
The University of Chicago, Chicago, Illinois

Introduction

It is generally believed that proteins are formed alongside of nucleic acid templates. The sequence of purine-pyrimidine bases in the template is supposed ~~somehow to~~ determine the sequence of the amino acids in the particular polypeptide (protein) that a given template will form. The purine and pyrimidine bases of the template, are adenin, uracil, guanin and citosin if the template ~~is~~ an RNA molecule, and if the template ~~is~~ a DNA molecule, thymin takes the place of ~~citosin~~.

Because ~~the~~ a template which synthetizes protein must carry the same information as the gene but need not necessarily be the gene itself, we shall refer ~~to~~ such a template as a paragene.

It has remained so far a complete mystery in just what conceivable way amino acids could read a code ~~that consists of a sequence of purine-pyrimidine nucleotides~~. In what manner can chemical forces -- of the kind we know to exist -- line up amino acids alongside such a template in the proper sequence and at the proper distance from each other so that there might be initiated a chemical reaction chain through which adjacent amino acids might form a peptide bond with each other, ~~and thus form a polypeptide~~.

It is the purpose of the present paper to indicate a conceptually simple scheme that will -- at least by way of an example -- illustrate

in what manner this might be accomplished in the living cell.

The basic thought underlying this scheme consists in the assumption that there are a number of enzymes (or enzyme systems) in the cell, and that each of these catalyzes the formation of a particular tri-nucleotide which carries a particular amino acid or a particular sequence of three amino acids. The amino acids are tied to the trinucleotide through a high energy bond; either a P or PP bond, and such acid anhydride bonds that when split may release an energy of 12,000 calories or 16,000 calories respectively, when they are split.

according to the notions here presented amino acids can not read the code of the messenger at all. But the trinucleotides can form together purine and pyrimidine bases of the trinucleotides can attach through the formation of a hydrogen bond to the proper locations on the messenger and thus line up the amino acids in the proper sequence. Each trinucleotide acid carries being present in the form of an acid anhydride carries with it the energy that may be released in a chemical reaction when it links adjacent

~~must~~ ~~Each of~~ three ~~nucleotides~~ represents the anti-code words ~~and is complementary to a code word i.e. to certain sequences of three nucleotides~~ ~~the~~ which are complementary to the code words. The code words are certain sequences of three nucleotides which occur on the messenger. We can

Sequences of three nucleotides along the paragene represent the code words, and the trinucleotides which carry the amino acids represent the anti-code words. These anti-code words are complementary to the code words in the sense that where the code-word contains adenine the anti-code word contains uracil (or thymin), and where the code word contains uracil (or thymin), the anti-code word contains adenin, and similarly guanin corresponds to citosin and citosin corresponds to guanin. The rationale for this assumption is as follows:

in what manner this might be accomplished in the living cell. The basic thought underlying this scheme consists in the assumption that a number of enzymes (or enzyme systems) -- not less than 20 and not more than 64 -- are contained in the cell and that each one of them catalyzes the formation of a particular trinucleotide which carries a particular sequence of three amino acids. These trinucleotides contain the five-carbon sugar ribose rather than the five-carbon sugar desoxyribose. Each of these five carbon sugars carries a phosphate group in the two-carbon position and an amino acid is attached to each of these phosphate groups through an acid anhydride bond (either P or PP representing an energy of 12,000 calories or 16,000 calories, respectively.)

(The possibility that the cell utilizes in fact not trinucleotides but tetranucleotides will be discussed later on in passing. And so will be the possibility that each multinucleotide (trinucleotide or tetranucleotide) might carry one amino acid only, tied with an acid anhydride bond to a phosphate group that in turn hangs either on a three- or a five-carbon atom of either the first or the last nucleotide.)

These trinucleotides contain three of the four bases, adenin, uracil, guanin, or citosin, and we assume that these are complementary to the code-words at the paragene. To each code word on the template we can construct a complementary code-word by replacing adenin with uracil, uracil (or thymin) with adenin, guanin with citosin, and citosin with guanin. The rationale for this assumption is as follows:

The concept of code-word and complementary code-word arose originally from the study of the structure of DNA. We know that in a double stranded DNA structure, adenin pairs with thymin (which presumably plays the same role in DNA as does uracil in RNA) and guanin pairs with citosin. Such pairing is required by the

The helical structure of DNA permits just such pairing, and hydrogen bonding is possible between adenin and thymin, as well as between guanin and citosin.

We may now tentatively take the view that during protein synthesis the paragene, whether it be a single DNA or a single RNA strand, assumes a somewhat similar helical configuration. The amino acids carried by the proper trinucleotides (the anti-code words) may then be lined up in the proper sequence along the paragene through the formation of hydrogen bonds between the purine and pyrimidine bases of the trinucleotides and the complementary bases of the paragene. When the trinucleotides are lined up in the proper order then, since each trinucleotide carries the proper amino acid, the amino acids are also lined up in the proper order.

We shall consider here now in greater detail one particular way in which the amino acids may be lined up alongside the paragene in the proper order and at the proper distance from each other.

Each particular ribose trinucleotide (the anti-code word) carries a particular sequence of three amino acids. A phosphate (or diphosphate) group is attached through an oxygen to the (2) carbon atom of the ribose moiety of each nucleotide (ester linkage). To each of these phosphate (or diphosphate) groups there is attached an amino acid. These acid anhydrides represent an energy-rich P (or PP) bond so that each amino acid carries with it the energy which is necessary to form peptide bonds between adjacent amino acids. We assume that during protein synthesis the nucleic acid strand that functions as a template (the paragene) take up a helical configuration resembling the helical configuration of a DNA strand in the double strand DNA helix. The ribose trinucleotides may then line up alongside the helical paragene with their purine and pyrimidine bases paired with the complementary purine and pyrimidine bases of the paragene, and if they are so lined up, then the amino acids carried by the trinucleotides might come to lie at just about the right distance from each other to permit the formation of a peptide bond between the adjacent amino acids. A chemical reaction chain -- starting from the head of a paragene -- might thus move down alongside the paragene, split off one or two phosphates from the number 2 carbon positions of the ribose moieties, and thus supply the energy for the formation of the peptide bonds between adjacent amino acids.

If indeed proteins are formed in this manner, then there are certain restrictions imposed on the possible amino acid sequences that paragenes can produce. As we shall presently see, however, this restriction is not a very serious one. If we have four letters to choose from, we can form 64 different three-letter words.

energy for the formation of the peptide bonds between adjacent amino acids.

If indeed proteins are formed in this manner, then there are certain restrictions imposed on the possible amino acid sequences that paragenes can produce. As we shall presently see, however, this restriction is not a very serious one.

If our code consists in three-letter words, if all 64 possible three-letter combinations form a code word, and if the nucleic acid strand assumes at the time of the formation of the polypeptide the helical configuration discussed above, then it follows that the code on the paragene must be read consecutively from one end -- say, the "head" of the paragene onward. This is so because this helical structure does not provide for commas between the individual code words, and in a 64-word, three-letter word code every three consecutive letters form a word. The letters 1, 2, 3 form a code word which was meant to be conveyed and so do the letters 4, 5, 6, but sequences of three letters which encroach on two adjacent words (such as 2, 3, 4 or 3, 4, 5, for example) form code words which are not meant to be conveyed. In these circumstances, the code would be misread if the trinucleotides, which represent the anti-code words, were to assemble simultaneously rather than consecutively -- alongside of the paragene. If we want simultaneous assembly of the trinucleotides alongside of the comma-less paragene template, then we must be satisfied with only 20 code words instead of the 64 code words that are possible if there are commas between the words.

The notion of such a 20-word code, which needs no commas, was introduced by F.H.C.Crick, J.S.Griffith, and L.E.Orgel in a memorandum circulated in May, 1956 among workers interested in the subject of protein synthesis. They have shown that, if we have four letters at our disposal

From such a code we must demand that while the letters 1, 2, 3 on the template form a code word, and the letters 4, 5, 6 also form ~~a~~ code words, sequences of three letters, which encroach on two adjacent words (such as 2, 3, 4 or 3, 4, 5, for example) form no code word. Crick and his co-workers have shown that this demand can be met, and that a code which requires no commas may be constructed ~~that has available 20 different three-letter words with the letters drawn from a group of four different letters~~ and that it can have code words.

from which we form three-letter words, there can be constructed a code consisting of 20 words which requires no commas. They raised the question of whether the number 20 might represent a more than fortuitous coincidence in view of the fact that there might be just about 20 essentially different amino acids that go into the formation of proteins. On the basis of the notions presented above, this coincidence would have to be regarded as fortuitous. From such a code, which requires no commas, we must demand that while the letters 1, 2, 3 alongside the template form a code word, and the letters 4, 5, 6 also form a code words, sequences of three letters, which encroach on two adjacent words (such as 2, 3, 4 or 3, 4, 5, for example) form no code word. Crick and his co-workers have shown that their 20-letter code meets this demand.

Applying the concept of a 20-letter code, that requires no commas, to our particular system of protein synthesis, we may now say the following:

Each of the 20 amino acids may be carried once as the first letter and once as the last letter of one of the 20 trinucleotide anti-code words. Therefore, among the polypeptides that can be formed, each amino acid may precede any other amino acid, and each amino acid may follow any other amino acid. This does not, of course, mean that any amino acid sequence is possible. Whether some of the amino acid sequences that may be found experimentally in sequential analysis of proteins and polypeptides will prove that the restrictions imposed by this system of polypeptide synthesis on the possible amino acid sequences are too severe, remains to be seen.

There is, however, no inherent reason why we should have a pure three-letter code or why four-letter code words should not be utilized also.

If we had a pure four-letter word code and demanded that this code requires no commas, the number of code words available will be greater than 20.)

(Professor Leonard J. Savage of the University of Chicago informs me that in such a pure four-letter code the number of words is 46 or greater.

If we had a pure three-letter code, we would have to demand that the number of amino acid residues of all polypeptides or proteins synthetized in the manner described above should be a multiple of three.

The number of amino acid residues in the proteins and polypeptides so far analyzed are as follows:

a) Insulin, chain A 21; insulin, chain B 30; corticotropin, B 39; oxytocin, 9; vasopressin, 9; intermedin B 18; all of these would fit a pure three-letter code.

b) Intermedin A 13; glucagon 29 and pancreas ribonuclease 124; these do not fit a pure three-letter code; intermedin and ribonuclease would have to include at least one 4-letter word and glucagon at least two 4-letter words.

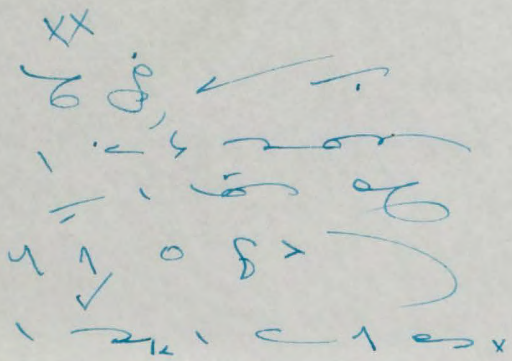
Obviously in a mixed system of three- and four-letter words, one can -- in the case of sufficiently large numbers -- never conclude that more than two 4-letter words must be included.

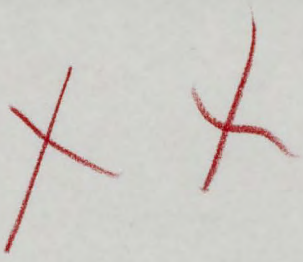
Rate of enzyme synthesis

In general the rate of enzyme synthesis in bacteria is quite low. Most enzymes are synthetized in growing bacteria at a very low rate, which represents and according to the notions here presented these rates do not reflect the maximum synthetic capacity of the corresponding paragenes. We know, however, that the rate of production of a given enzyme may be greatly enhanced if the enzyme is induced. The enzyme, β-galactosidase, for instance, which splits lactose, can be induced -- as Jacques Monod and his co-workers have shown -- by certain chemical analogues of lactose; for instance, thiomethyl galactoside (TMG). The rate of the production of these enzymes goes up almost instantaneously upon adding the inducer to the medium and it may go up by a factor of several thousand to about 5 (and perhaps even 10) enzyme molecules per second.

Because the increase in the production rate of the enzyme is almost instantaneous upon the addition of the inducer, we are inclined to assume that the inducer does not act by increasing the number of templates that make this particular enzyme but rather by increasing the rate at which one particular template makes the corresponding enzyme. It seems, furthermore, reasonable to believe that all templates which make enzymes of about the same molecular weight -- say, of 100,000 -- are potentially capable of synthetizing their enzyme at the same maximum rate. Since the paragenes might consist of RNA and since there may be about five times as much RNA in the bacterial cell as there is DNA, it is conceivable that every enzyme is made by perhaps five paragenes rather than just one paragene. And one might then conclude that a paragene must be capable of synthetizing the enzyme for which it is specific at a rate that is of the order of magnitude of one per second, and the maximum rate might in fact be higher than this.

In these circumstances, we are led to conclude ~~that a lower limit for the order of magnitude for~~ the rate of enzyme synthesis ~~by one~~ paragene is ~~at the limit of the order of~~ one per second.





Computed rate of enzyme synthesis

We shall now attempt to compute at what rate a paragene may be able to synthetize the corresponding enzyme on the basis of the ~~mechanism of protein synthesis~~ that we have postulated. We shall assume that the molecular weight of the enzyme is 100,000, ~~i.e. that~~ we have about 1,000 amino acid residues in the enzyme, and accordingly we would have to assemble m = 300 trinucleotides, each of which is "loaded" with three amino acids. In the approximation which we shall consider, the minimum time, τ, needed for the formation of the polypeptide is composed of two terms, τ_1 and τ_2. $\tau = \tau_1 + \tau_2$

~~If a polypeptide has been assembled and~~ After all the amino acids ~~assembled~~ alongside the template have been ~~joined in a~~ polypeptide ~~chain~~ and assuming that this polypeptide is at once removed, a certain time, τ_1, ~~(to be computed later)~~ will elapse until the trinucleotides, ~~which arrive loaded with amino acids but~~ which are now denuded, evaporate from the template and their place is taken by trinucleotides which are loaded with amino acids. We shall assume that the concentration of denuded trinucleotide in the cell is very small ~~compared to the concentration of trinucleotides which are loaded with the proper amino acids~~ so that after the denuded trinucleotides evaporate, the loaded trinucleotides do not have to compete with ~~the~~ denuded trinucleotides for legitimate positions along the paragene.

The time, τ_1, which is necessary to permit evaporation of ~~all~~ m trinucleotides and to assemble ~~all (i.e. m,)~~ m loaded trinucleotides, we shall compute here on the assumption that once a loaded trinucleotide has found its position alongside the template, it will not evaporate again. ~~This~~ assumption is, of course, not ~~correct~~, and ~~the~~ correction ~~that is needed~~ is represented by the second term, τ_2.

In order to compute this second term, τ_2, we must consider the equilibrium between the loaded trinucleotide in the cell and the complex which such a loaded trinucleotide can form with one complementary sequence of 3 nucleotides on the template.

We shall in the following assume that the concentration ρ_0 (in Mol/l) of the loaded trinucleotide in the cell is the same for each kind of loaded trinucleotide. We shall further assume that the equilibrium constant, K_0, for the equilibrium between one kind of free loaded trinucleotide and one given complementary sequence of three nucleotides is the same for each kind of loaded trinucleotide.

The probability that in equilibrium a given code word on the template is not covered by the proper bonded trinucleotide is given by

$$f = \frac{1}{1 + \frac{\rho}{K}}$$

and the total number of such "holes" alongside the template in equilibrium which contains $3n$ nucleotides is given by

$$\frac{n}{1 + \rho/K}$$

When the formation of the polypeptide between the adjacent amino acids is somehow triggered at the head of the template, then these holes will have to be filled consecutively, and we may write for the time that this will take

$$\tau_2 = \frac{1}{A\rho} \frac{n}{1 + \rho/K}$$

Here β represents the rate at which a loaded trinucleotide combines with a given code word that is complementary to it.

$$\beta = A\rho$$

We may now compute the mean time, $\bar{\tau}_1$, needed for the evaporation of the naked trinucleotides and the ~~condensation~~ assembling of the loaded trinucleotides in their place. The rate, α, at which a loaded trinucleotide would evaporate from the template is given by (*see Appendix)

$$\alpha = 2AK$$

We shall, for the sake of simplicity, assume that a denuded trinucleotide evaporates at the same rate α (*see Appendix)

As ~~will~~ may be shown, we may then write within the approximation here ~~attempted~~ used for the time, $\bar{\tau}_1$,

$$\bar{\tau}_1 = \frac{1}{A\rho}\left\{1 + \ln m - \ln\frac{\beta}{\alpha}\right\}\frac{\rho}{2\alpha}$$

$$\text{when } \frac{\beta}{\alpha} \gg 1$$

For the total time, $\bar{\tau}_0 = \bar{\tau}_1 + \bar{\tau}_2$

we thus obtain

$$\bar{\tau}_0 = \frac{1}{A\rho}\,\frac{m}{1+\frac{\rho}{K}} + \frac{\rho}{2\alpha}\left\{1 + \ln m - \ln\frac{\beta}{\alpha}\right\}$$

If we wish to make this time as small as possible, we have to choose K so as to have

$$\frac{\rho}{K} = \sqrt{\frac{2m}{1+\ln m - \ln\frac{\beta}{\alpha}}} - 1$$

~~From this we obtain~~ or

$$\frac{\rho}{K} \approx 10$$

and, therefore, we obtain for $m \approx 300$

$$\bar{\tau}_0 \approx \frac{50}{A\rho}$$

For paper May 31/57

$A = 6\,10^{23} \times 10^{-3}\, v\, \sigma\, \rho$

$v = \sqrt{\frac{2RT}{\pi M}}$ $\quad \sigma = 10^{-15}$ $\quad \rho = \frac{1}{300}$

$M = 1000$ $\quad \rho = \frac{10}{4\pi(7)^2} \times \frac{1}{5}$

$\underline{A_1 = 10^7}$ for $\rho = 10^{-5}$ Mol/liter

$A_1 \rho = 100$/sec.

~~$A_1 \rho = 100$/sec.~~ $\underline{m = 300}$

rate of evaporation of 1 naked bivalent bridge from ~~a~~ code Word. (assumption

assumption ~~$A_1 \rho = \alpha = 2A_1 K_1$~~ $\alpha = 2A_1 K_1$ low concentration of naked and code

anti code word

T_1 Time for assembly $\sim \frac{1}{A\rho} \frac{\rho}{2K} \left\{ 1 + \frac{m}{m} \ln m - \ln \frac{\beta}{\alpha} \right\}$

$\beta >> \alpha$ $\quad \beta = A\rho$ $\quad \alpha = 2AK$ $\quad \frac{\beta}{\alpha} = \frac{\rho}{2K}$

T_2 time to fill gaps. —

$T_2 = \frac{1}{A\rho} \left\{ m \frac{1}{1 + \frac{\rho}{K}} \right\}$

$T = \frac{1}{A\rho} \left\{ \frac{m}{1 + \frac{\rho}{K}} + \left\{ 1 + \ln m - \ln \frac{\beta}{\alpha} \right) \frac{\rho}{2K} \right\}$

B

$\frac{\rho}{K} = \sqrt{\frac{2m}{B}} - 1$

$\frac{\rho}{K} \cong 9$

$T \cong \frac{50}{A\rho}$ or if $\rho = 10^{-5}$ Mol/l

$T = \frac{1}{2}$ or 2/sec per gene

In order to compute this second term, τ_2, we must consider the equilibrium between the loaded trinucleotide in the cell and the complex which these trinucleotides can form with the complementary nucleotide (the anti-code word) on the template.

The probability that in equilibrium a given code word on the template is not covered by the proper trinucleotide, which is loaded with the proper amino acids, is given by

$$\frac{1}{1+\frac{p}{k}}$$

and the total number of such holes alongside the template in equilibrium is given in equilibrium by

$$\frac{m}{1+\frac{p}{k}}$$

We shall in the following assume that the concentration of each kind of trinucleotide, loaded by the proper amino acids, is about the same. We may then write for the rate, β, at which a free trinucleotide on the template combines with the proper trinucleotide (that is loaded with the proper amino acids)

Here the coefficient, A, stands for

$$A_0 = 6\,10^{20}\, v_0\, \sigma_0\, p_0$$

where v, the molecular velocity

$$v_0 = \sqrt{\frac{2RT}{\pi M}}$$

; For a molecular weight of $M = 1000$ we have $v_0 = 5\,10^3$ cm/sec

if

and σ_0 is the area of the target that must be hit for hydrogen bonding is to take place between the trinucleotide and the complementary code word. We shall assume

$$\sigma_0 \approx 10^{-15}\ \text{cm}^2$$

and finally p denotes the probability that the trinucleotide, when hitting the code, is in just the right position to permit hydrogen bonding to take

place between the three complementary pairs of bases that are involved, ~~which~~ how
~~Taking the molecular weight M for a trinucleotide as about 1,000, we ob-
tain~~

~~and if we~~ ~~very~~ ~~roughly~~ ~~estimate p = 1/300, we obtain for A~~

we may take as a very rough estimate

$$p_0 \approx \frac{1}{300}$$

These estimates give

$$A_0 = 10^7 \text{ per sec}$$

$$6 \, 10^{20} \; 5 \, 10^{3} \; 10^{-15} \; \frac{1}{300} = 10^7 / \text{per sec}$$

~~For a concentration of $\rho \approx 10^{-5}$ mol/sec~~

~~we have a minute~~

If the concentration of each kind of loaded trinucleotides in mol/l in the cell is about

$$\rho = 10^{-5} \text{ mol/l}$$

Then ~~$T \approx \frac{50}{A\rho}$~~ $A\rho = 100$

and T_0 the minimum time needed for one gene to form one protein molecule is obtained from No :

$$T_0 = \frac{50}{100} = 1/2 \text{ sec}$$

thus the maximum rate at which ~~the~~ one gene can make enzyme

is about 2 per second.

~~It is of course possible that the value of σ_P is even~~

Our rough estimate for σ_P might well be off by a factor 10; thus σ_P might be ten times higher than the value we quoted and P might then be 10 times lower i.e. it might be 10^{-6} M/liter assuming $\frac{P}{K} \approx 10$ the corresponding K values range from

$$K = 10^{-7} \text{ to } 10^{-6}$$

~~$K = 10^{-7}$~~ would correspond to a binding energy ΔH of

~~$K \times 10^{7} = 1 \cdot 10^{13} e^{-\frac{Q}{RT}}$~~

We can compute from K the binding energy ΔH between the promoter site and the template proper ~~site~~ creating at the template by writing:—

$$2AK = 10^{13} e^{-\frac{Q}{RT}}$$

for $K = 10^{-7}$

$$2 = 10^{13} e^{-\frac{Q}{RT}} \qquad 2.3 \quad 6 \quad 12$$

$$\frac{Q}{RT} + 1 = 13 \times 2.3 \qquad \frac{Q}{RT} \approx 30 \qquad \frac{23}{7} = 30$$

$Q = 18{,}000$ Cal. or 3000 Cal/H bound and 1,400 cal less for $K = 10^{-6}$

June 7, 1957

HOW MIGHT AMINO ACIDS READ THE NUCLEOTIDE CODE?

by Leo Szilard

(Submitted by Joseph E. Mayer)

The Enrico Fermi Institute for Nuclear Studies
The University of Chicago, Chicago, Illinois

It is now generally believed that proteins are formed alongside nucleic acid templates. The sequence of purine and pyrimidine bases in the template is supposed to represent a code that may somehow determine the sequence of the amino acids in the particular polypeptide (protein) that a given template will form.(1) The purine and pyrimidine bases of the template, the letters of the code, are adenine, uracil, guanine and cytosine, if the template be an RNA molecule; and if the template be a DNA molecule, thymine takes the place of uracil.

It has remained so far a complete mystery in just what conceivable way amino acids could read such a code. In what manner can chemical forces -- of the kind we know to exist -- line up amino acids alongside such a template in the proper sequence and at the proper distance from each other, so that a chemical reaction chain may link adjacent amino acids through peptide bonds with each other?

It is the purpose of the present paper to indicate a conceptually simple scheme that will -- at least by way of an example -- illustrate in what manner this might take place in the living cell.

---

(1) G. Gamow, Nature, Vol. 173, p. 318 (1954).

Because a template which synthetizes protein need not necessarily be the gene itself but must carry the same information as the corresponding gene, we shall here refer to such a template for the sake of brevity as a paragene.

The basic thought underlying the scheme here presented consists in the assumption that there are a number of enzymes (or enzyme systems) -- perhaps twenty altogether -- in the cell, and that each of these catalyzes the formation of a particular trinucleotide which carries either one particular amino acid or, more likely perhaps, a particular sequence of three amino acids. If the amino acid is carried by the nucleotide on a phosphate or pyrophosphate group as an acid anhydride -- which is a high energy compound -- then the energy needed for the formation of the peptide bonds will become free when the amino acid is split off. In this sense one can say that each amino acid may carry the energy needed for forming its peptide bond.

According to the notions here presented, amino acids can not themselves read ~~themselves~~ the code of the paragene. But the trinucleotides, which carry the proper amino acids, may attach with their three bases through the formation of 6 hydrogen bonds to the proper sequence of three bases on the paragene, and thus the amino acids may be lined up in the proper sequence along the paragene.

Accordingly, sequences of three nucleotides along the paragene represent the code words, and the trinucleotides which carry the amino acids represent the anti-code words. We assume that these anti-code words are complementary to the code words in the sense that, where the code word contains adenine the anti-code word contains uracil (or thymine), where the code word contains uracil (or thymine) the anti-code word contains adenine; and similarly guanine corresponds to cytosine and cytosine corresponds to guanine. The rationale for this assumption is as follows:

The concept of code-letter and complementary code-letter arose originally from the interpretation of the structure of DNA given by J.D. Watson and F.H.C.Crick.(2) They showed that in a double stranded DNA

(2) Nature, Vol. 173, p. 318 1953.
Proc. Roy. Soc., Vol. 223, p. 80, 1954.

structure, adenine of one strand pairs with thymine of the other strand (which presumably plays the same role in DNA as does uracil in RNA) and similarly guanine pairs with cytosine. The helical structure of DNA permits just such pairing, and hydrogen bonding is possible between adenine and thymine, as well as between guanine and cytosine.

If the sequence of bases along one strand of DNA represents a coded message which consists in three letter-words, then, because we have four letters to choose from, such a message could utilize 64 different words. We might, however, be limited to the use of 20 out of the 64 words that are available. The reasons for this limitation would be as follows:

If all 64 possible three-letter combinations form a code word, and if the paragene assumes at the time of the formation of the polypeptide a helical configuration similar to the helical configuration of DNA, then it follows that the code on the paragene must be read consecutively from one end -- say, from the "head" of the paragene downward. This is so because such a helical structure does not provide for commas between the individual code words, and in a code containing 64 words any three consecutive letters form a word. If we number the letters along the paragene, from the head of the paragene downward, then the first three letters, the letters 1, 2, 3, form a word which was meant to be conveyed and so do the next three letters, the letters 4, 5, 6. But sequences of three letters which encroach on two adjacent words (such as 2, 3, 4 or 3, 4, 5, for example) form code words which are not meant to be conveyed.

In these circumstances, the code would be misread if the trinucleotides, which represent the anti-code words, assemble alongside the paragene simultaneously, rather than -- from one end on -- consecutively. If we want to have simultaneous assembly of the trinucleotides alongside the comma-less paragene, then we are restricted to 20 code words.

The notion of such a 20-word code, which needs no commas, was introduced by F.H.C.Crick, J.S.Griffith, and L.E.Orgel of the Medical Research Council Unit at the Cavendish Laboratory, Cambridge, in a memorandum circulated in May, 1956 among workers interested in the subject of protein synthesis. From such a code we must demand that the letters 1, 2, 3 on the template form a code word, and the letters 4, 5, 6 also form a code word, but sequences of three letters, which encroach on two adjacent words (such as 2, 3, 3 or 3, 4, 5, for example) form no code word. Crick and his co-workers have shown that this demand can be met, that a code which requires no commas may be constructed and that it can accommodate 20, three-letter, code words.

We shall now single out for more detailed examination one conceivable model for protein synthesis which might provide for the lining up of the amino acids alongside the paragene, both in the proper order and at the proper distance from each other. This particular model is based on the following assumptions:

The trinucleotides which form the anti-code words contain the sugar ribose rather than the sugar desoxyribose. Each particular ribose trinucleotide (the anti-code word) carries a particular sequence of three amino acids. A phosphate (or diphosphate) group is attached to the (2) carbon atom of the ribose moiety of each nucleotide and an amino acid is attached to each of these phosphate (or diphosphate) groups. The amino acid anhydrides represent an energy-rich P, or PP, bond which, when split, may release 12,000 or 16,000 calories, respectively.

During protein synthesis the nucleic acid strand that functions as a template (the paragene) may take up -- so we here assume -- a helical configuration resembling the helical configuration of a ~~EXN~~ DNA strand in the double stranded DNA helix. The trinucleotides may then line up alongside the helical paragene with their purine and pyrimidine bases paired with the complementary bases of the paragene, and if they are so lined up, then the amino acids carried by the trinucleotides may come to lie at just about the right distance from each other to permit the formation of a peptide bond between adjacent amino acids. A chemical reaction chain -- starting perhaps from the head of a paragene -- may then move down along the paragene, split the acid anhydrides, and thus free the amino acids as well as make available the energy needed for the formation of peptide bonds between adjacent amino acids.

Adjacent amino acids can be linked only if the distance from each other is smaller or equal but not appreciably larger than the fundamental repeating distance in a polypeptide chain which is 7.27Å. The fundamental repeating distance in a fully extended polypeptide chain is about 7Å. Since before they are linked into a polypeptide, the amino acids can rotate around the chemical bond which ties them to the phosphate group, they might well be assembled along the paragene at somewhat smaller distance from each other than the fundamental repeating distance of the polypeptide chain.

Applying the concept of a 20-letter code that requires no commas to our particular model of protein synthesis, we may now say the following:

Each of the 20 amino acids may appear once attached to the leading letter and once attached to the trailing letter of the 20 (trinucleotide) anti-code words. Therefore, among the polypeptides that can be formed, each amino acid may precede any other amino acid, and each amino acid may follow any other amino acid. This does not, of course, mean that any amino acid sequence is possible.

Observed rate of enzyme synthesis

According to the notions here adopted most enzymes are synthetized in growing bacteria at a rather low rate which does not represent the maximum synthetizing capacity of the corresponding paragenes. The rate of production of a given enzyme may, however, be greatly enhanced if the enzyme is induced, and what we are interested to learn is the maximal rate at which a paragene may be able to form the corresponding enzyme.

One of the most studied cases of enzyme induction is the induction of the enzyme β-galactosidase which splits lactose. Jacques Monod and his co-workers have shown that the production rate of this enzyme in bacteria can be greatly enhanced by certain chemical analogues of lactose, which act as inducers, and that the rate of production of the enzyme goes up almost instantaneously upon adding such an inducer to the medium. We are thus led to believe that the inducer may act by increasing the rate at which one template produces the enzyme rather than by increasing the number of templates that produce the enzyme at an unchanged rate.

In fully induced wild type bacteria growing in minimal medium this enzyme is contained in the amount of about 8 mgm. per 10^{12} bacteria and thus amounts to about 8% of the total proteins. We obtain the rate at which this enzyme is produced in minimal medium per bacterium by dividing the amount contained in one bacterium by 1.44 times the doubling time (40 minutes) of the bacterium. We thus find for the rate, at which this enzyme is produced in fully induced wild type bacteria growing in minimal medium, about $2\ 10^{-18}$ grams per cell per second.

If we assume a molecular weight of a million (Jacques Monod and Melvin Cohn estimate the molecular weight of this enzyme at about 800,000), we obtain a rate of 1.5 enzyme molecules per cell per second. The number of paragenes per cell is not known. ~~If the paragene is RNA~~ there might be a few paragenes present per ~~cell~~ rather than just one, and the number of paragenes might be of the order of magnitude of 10. On the other hand, smaller enzyme molecules might be synthetized somewhat faster than larger enzyme molecules.

On the basis of the figure given above, we are thus led to believe that when an enzyme is fully induced and enzyme synthesis proceeds at its maximal rate -- the rate of formation of the enzyme may be of the order of magnitude of one per second per paragene.

Computation of τ_2

First we shall now compute this second term, τ_2. This computation will be based on the fact that (because of reevaporation of the loaded trinucleotides, which are reversibly combined with the anti-code words of the paragene) there will be - no matter how long we wait - always a certain number of code words (sequences of three nucleotides) on the paragene which are not "covered". We shall assume that the equilibrium constant, K_o, for the complexing of one code word by the proper loaded trinucleotide, is the same for the different kinds of trinucleotides, and that each code word complexes only with the proper trinucleotide. Similarly we shall assume that the different kinds of loaded trinucleotides are all present at the same concentration, ρ_o, in the cell.

In equilibrium the probability, f, for a given code word on the paragene not being covered by the proper loaded trinucleotide is given by

(1) $$f = \frac{1}{1 + \frac{\rho_o}{K_o}}$$

Accordingly, in equilibrium, the total number of such gaps along the paragene which contains 3 m nucleotides is given by

(2) $$\text{"number of gaps"} = \frac{m}{1 + \frac{\rho_o}{K_o}}$$

We shall assume that most code words are "covered" in equilibrium and this means that

(3) $$\frac{\rho_o}{K_o} >> 1$$

We presume that after such equilibrium is established a chemical reaction chain is somehow triggered, and, moving down along the paragene, links adjacent amino acids into a polypeptide. The average time, τ_2, needed for the formation of the polypeptide from the amino acids assembled along

the paragene is given by the product of the "number of gaps", that have to be filled consecutively, and the average time, ~~1/β~~ $\frac{1}{A\rho}$, that it takes to fill one given gap.

Thus, for $\bar{\tau}_2$ we may write

$$\bar{\tau}_2 = \frac{1}{A\rho} \frac{m}{1 + \frac{\rho_0}{K_0}} \tag{4}$$

In this expression $1/\beta$ is the average time that it takes for a given gap to be filled by the proper trinucleotide, and we may write for β

$$\beta = A_0 \rho_0$$

where ρ_0 is the concentration ~~(in moles per liter~~ of the loaded trinucleotides of any one kind (which we took to be equal for all ~~for each kind of trinucleotide)~~, and A_0 is a constant for which a value is given further below.

Computation of $\bar{\tau}_1$ and $\bar{\tau}_0$

We may now compute the average time, $\bar{\tau}_1$, needed for the evaporation of the naked trinucleotides and the assembling of the loaded trinucleotides in their place. The rate, α, at which a loaded trinucleotide evaporates from the template is given (see appendix) by

$$\alpha = 2 A_0 K_0 = 2 \frac{K_0}{\rho_0} \beta \tag{5}$$

(6) and for $\frac{\rho_0}{K} >> 1$ we have $\beta >> \alpha$

We shall, for the sake of simplicity, assume that a denuded trinucleotide evaporates at the same rate, α.

As may be shown (see appendix), we may then write ~~within the approximation here used for~~ for very large values of m ($m >>> 1$) for the time, $\bar{\tau}_1$,

$$\bar{\tau}_1 \approx \frac{1}{A\beta} \frac{\rho_0}{2K_0} \ln m \tag{7}$$

For the total time, $\tau_0 = \tau_1 + \tau_2$ we thus obtain

(8) $$\tau_0 \approx \frac{1}{AB}\left\{\frac{m}{1+\frac{\rho_0}{K_0}} + \frac{1}{2}\frac{\rho_0}{K_0}\ln m\right\}$$

If we wish to make this time as small as possible, we have to choose K_0 so as to have for $\frac{\rho_0}{K_0}$

(9) $$\frac{\rho_0}{K_0} \approx \sqrt{\frac{2m}{\ln m}}$$

~~Substituting this value into (8) and writing $A_0\rho_0$ for β, we obtain~~

(10) $$\tau_0 \approx \frac{\sqrt{2}}{A_0\rho_0}\sqrt{m \ln m}$$

For a polypeptide containing 1,000 amino acid residues ~~or~~ i.e. a paragene containing about 300 code words, we may write m = 300, and thus we obtain from (8) and (10)

(11) $$\frac{\rho_0}{K_0} \approx 10$$

(12) $$\tau_0 \approx \frac{50}{A_0\rho_0}$$

Estimate for the value of A_ρ and ρ

If ρ_0 is expressed in mol/liter, then the coefficient A_ρ is given by

(13) $$A_0 = 6\,10^{20}\,\nu_0\,\sigma_0\,p_0$$

v_α is the molecular velocity

(14) $$v_\alpha = \sqrt{\frac{2RT}{\pi M}}$$

and for a molecular weight of $M \approx 1000$, we have $v_\alpha \approx 5 \times 10^3$ cm/sec.

σ is the target area that must be hit if hydrogen bonding is to take place between three adjacent nucleotides on the paragene and the complementary, loaded trinucleotides that move about freely within the cell. We assume for σ_0 the value of $\sigma_0 = 10^{-15}$ cm^2.

p_0 denotes the probability that the loaded trinucleotide, when hitting the code-word, is in just the right geometrical position to permit hydrogen bonding to take place between the three complementary pairs of bases that are involved. We may take for p_0 as a very rough estimate $p_0 = \frac{1}{300}$.

Thus we obtain $\sigma_0 p_0 = 1/3 \ 10^{-13}$.

With the above quoted values we obtain: $A_\alpha = 10^7$/sec.

If the concentration ρ_α of each kind of loaded trinucleotide in the cell is $\rho_0 = 10^{-5}$ mol/liter, then we have

(17) $$A_\alpha \rho_0 = 100 \text{ hits/sec.}$$

Substituting this value for $A_\alpha \rho_0$ into (12) gives for the average time, τ_0, that it takes for one paragene to form one polypeptide: $\tau_0 = 0.53$ sec.

This means that the paragene may form about 2 polypeptide ~~molecules~~ per second.

The value of τ_0 depends only on the product of A_α and ρ_0. It might well be that the value which we estimated for $\sigma_0 p_0$ is too low by a

factor of 10 and consequently the value which we obtained for A_o is also low by the same factor. If this were the case, then the value we chose for ρ_o is ten times too high and the ~~correct~~ true value would rather be $\rho_o = 10^{-6}$ mol/liter.

It is not possible for the present to estimate these values any closer. In these circumstances our assumption for the equilibrium constant, K, $\left(\frac{\rho_o}{K_o} \approx 10\right)$ gives only a rough estimate

(18) $$10^{-7} < K_o \text{ (in mol/liter)} < 10^{-6}$$

From the value of K_o we may estimate the binding energy, ΔH, for the combination of the trinucleotide with the paragene ~~by writing~~ from

(19) $$2 A_o K_o = 10^{13} e^{-\frac{\Delta H}{RT}}$$

This gives for $K = 10^{-7}$ mol/liter, $\Delta H \approx 18{,}000$ calories, or since six hydrogen bonds are involved, about 3,000 calories per hydrogen bond.

If K is ten times larger, then ΔH is about 1400 calories lower.

Conclusion

These considerations show that the theory which we postulated would be able to explain the high rate of enzyme synthesis which one observes in bacteria when the rate of formation of an enzyme is enhanced by the use of an inducer. The basic thought of the theory here given consists in the assumption that trinucleotides read the code of the paragene and that these trinucleotides carry amino acids. We worked out the details in the case of a particular model for protein synthesis which assumed

that each trinucleotide carries a sequence of three amino acids. Naturally I chose for the detailed theory the model which appeared to me to be the most ~~likely to be correct~~ plausible. This does not mean, however, that other models need not be considered also.

Rather than to assume that each kind of trinucleotide carries a particular sequence of three amino acids, it would be in some ways more appealing to assume that each trinucleotide carries only one amino acid. In this case the amino acid might be carried by a phosphate group linked by an oxygen atom (ester linkage), either to the third or the fifth of the (5') carbon sugar of either the ~~first~~ leading or the ~~third~~ trailing nucleotide. Assuming twenty different trinucleotides, each carrying one particular amino acid, we could have a code that requires no commas, with no restrictions imposed on the possible amino acid sequences of the proteins formed by the paragenes.

If we make this assumption, however, we can not assume at the same time that during protein synthesis the paragene takes up the helical configuration which we postulated above. In the case of such a helical configuration the trinucleotides lined up along the paragene would not place the amino acids at the right distance from each other to permit adjacent amino acids to be linked to a polypeptide.

It is conceivable that the paragene might assume during protein synthesis some entirely different configuration which meets the requirement of bringing adjacent amino acids at the right distance from each other when the trinucleotides, which carry one amino acid each, line up alongside the paragene. But unless it is possible to indicate a plausible, non-helical, and yet regular configuration that meets this requirement, it does not appear useful to carry the discuss of such an alternate model for protein synthesis any further.

These trinucleotides(+) may take up positions face to face with the corresponding trinucleotides(-) contained in the 1/2 RNA(-) strand. We assume that the amino acids are linked to the trinucleotide through a hydrogen bond, presumably an acid anhydride bond, between the phosphate group of the trinucleotide and the carboxyl group of the amino acid (estimated at 12,000 calories). One of several things might now happen:

(1) The adjacent acids from the peptide bonds join with each other so that there arises a polypeptide.

The adjacent ribonucleotides may link up with each other and form a strand of RNA which we may designate as 1/2 RNA(+).

The polypeptide can detach itself from the RNA strand and form a protein, and the 1/2 RNA(+) strand can detach itself from the 1/2 RNA(-) template, or it can form a two stranded RNA molecule designated by RNA(±).

To each polynucleotide one may assign an arror pointing from left to right or from right to left, which indicates the manner in which the phosphoric acid group is linked to the OH groups of the 5-carbon sugar. If we assign to the trinucleotides that carry the 20 different amino acids an arrow pointing from left to right, then the 1/2 RNA(+) strand which would arise in the process described above would also carry an arrow pointing from left to right.

Using this terminology, the 20 enzymes that catalyzed the formation of the postulated amino acid trinucleotides may be designated as . The basic thought here presented permits devising a number of different schemes for the synthesis of RNA and DNA. What schemes one might prefer depend to some extent on whether one believes that it might be possible to line up in the groove left free by one- and two-stranded helix formed by 1/2 RNA(+) and 1/2 RNA(-) strand of RNA. One may expect to have lined up

How can we account for the existence of the 1/2 RNA(-) templates in the cytoplasm? We assume that this template is synthetized within the nucleus. There are within the nucleus, so we shall assume, 20 enzymes which form twenty ribose-trinucleotides(-) which represent the anti-code word for the 20 amino acids. These enzymes couple each of these ribose-trinucleotides(-) to certain amino acids, perhaps predominantly arginine, histidine and lysin. We shall designate these enzymes with: E(AA, ribo-tri-nucleotide -)

There is, so we assume, within the nucleus a strand of DNA, which we shall designate as 1/2 DNA(+), which contains desoxyribose-trinucleotides (the code words) in the sequence which corresponds to the amino acid sequence, A_i, A_j, A_k, of the specific protein. In analogy to what happened in the cytoplasm, the ribose-trinucleotides(-) will come face to face with the corresponding desoxyribose-trinucleotides(+) contained in the 1/2 DNA(+) template, and with the help of the energy supplied by the acid anhydride bonds of the ribose-trinucleotides will link up to form a strand of RNA - 1/2 RNA(-). Some of the molecules thus formed in the nucleus will diffuse out into the cytoplasm where they will serve as a template for protein formation as described above.

The molecules remaining within the nucleus serve as a template for the synthesis of a strand of DNA -- 1/2 DNA(+). This strand of DNA is synthetized from desoxyribose-trinucleotides(+) which are formed and coupled to certain amino acids, perhaps predominantly arginine, histidine and lysin, by enzymes which are present within the nucleus. These enzymes may be designated by

The acid anhydride bonds of the ribose-trinucleotides that are lined up inside the template supply the energy for forming peptide bonds between the adjacent amino acids that are carried by the ribose-trinucleotides and to link the adjacent ribose-trinucleotides with each other through ester bonds. Thus the polypeptide and the RNA strand, 1/2 RNA(-), are formed but in contrast what happens in the formation of protein and polypeptide and the polynucleotide need not separate so that we have a ribonucleic protein which we may designate with AA - 1/2 RNA(-). Some of these will diffuse into the cytoplasm and may serve there as a template for the formation of proteins as discussed above.

Insert - New paper

~~1-B~~ 3

The ~~ribose~~ trinucleotides carrying the various amino acids will freely diffuse around in the cytoplasm of the cell. Assuming that the concentration which an amino acid which is thus coupled to the corresponding trinucleotide is of the order of a milligram per liter, we may expect that the right trinucleotide will come to lie face to face with the corresponding code word of the 1/2 RNA(-) template. Assuming an activation energy of 0, the code word will complex with the anti-code word at the rate of about 100 times per second. If the complex remains undissociated for a sufficiently long period of time and if the template synthetizes an enzyme which contains 1000 amino acid residues, then the time which it takes on an average to have all the code words of the 1/2 RNA(-) template thus complexed by the anti-code word will be proportionate to the natural logarithm of M, and hus in the case of the synthesis of an enzyme which is between 100 and 100,000 amino acid residues, it may take about 1/10th of a second to assemble all the anti-code words on the template. If the binding energy of the word-anti-word complex is about 18,000 calories, then the anti-code words stick to the template long enough to permit the completion of the whole amino acid sequence. It is presumed that the six high energy bonds that can be formed between the two complementary trinucleotides will supply binding energy in excess of this value. We may assume that the free phosphate group of the trinucleotides is linked to the amino acid which the trinucleotide carries by an oxide-anhydride bond representing about 12,000 calories. After all the ~~ribose~~ trinucleotides(+) are lined up on the 1/2 RNA(-) template, something might trigger a chemical reaction in which ~~a phosphate is split off in which the amino acids carried by the adjacent trinucleotides form a peptide linkage and the adjacent trinucleotides are linked by a phosphate ester bond and at the same time one phosphate is split off. In this way we would obtain simultaneously a poly-~~

peptide and 1/2 RNA(+) strand. The polypeptide has an amino acid sequence which is determined by the 1/2 RNA(-) template and the RNA strand, 1/2 RNA(+), that is newly formed is complementary to the 1/2 RNA(-) template.

DNA OK missing in second pos.

Since it is presumed that protein synthesis can occur without accompanying net RNA synthesis, we have to consider several possibilities:

a) That when the amino acids that are lined up are linked to a polypeptide, contrary to what we said above, the ribose trinucleotides remain unlinked and are returned to the free ribose trinucleotide pool.

b) That the 1/2 RNA(+) strand is formed but that such naked RNA strands are hydrolized to mononucleotides.

c) That the 1/2 RNA(-) template was part of a double stranded structure where the other strand is the 1/2 RNA(+) strand. These two complementary strands of RNA might form a helix and in the groove left free there may conceivably fit in the ribose trinucleotides(+) which carry the amino acids. When the polypeptide is formed simultaneously a new 1/2 RNA(+) strand is formed. One of the two 1/2 RNA(+) strands, either the original strand or the newly made strand, may remain united with the 1/2 RNA(-) template, and the other strand which is now naked might be hydrolized. If protein synthesis by the RNA template in the cytoplasm follows this pattern, we would then expect a considerable turn-over of RNA to accompany protein synthesis.

1/2 RNA(-) template, we presume, is formed at 1/2 DNA(+) strand in the nucleus. It is tempting to speculate that the RNA and DNA strands are formed in the nucleus on the basis of a principle very similar to the one described above in connection with the synthesis of specific proteins, with the following difference: The RNA and DNA strands must be synthetized from ribose trinucleotides which have a free pyrophosphate
from trinucleotide phosphates from trinucleotiphosphate amino acids In the former case the synthesis may result in ribose or desoxyribonucleic acid strands, whereas in the latter case ribose or desoxy ribonucleoprotein may be formed.

This means that there would have to be in the nucleus enzymes which catalyze the formation of ribotrinucleotide diphosphates and desoxyribonucleotide diphosphates, or in the case of the second alternative mentioned ribose trinucleotide amino acids, diphospho-amino acids and desoxyribose and trinucleotide amino acids. Among the amino acids thus coupled to trinucleotides, we presume that lysin, arginine, and histidine were predominant.