kvn June 7, 1957

HOW DO AMINO ACIDS READ THE CODE?

by Leo Szilard

(Submitted by Joseph E. Mayer)

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

It is generally believed that proteins are formed alongside of nucleic acid templates. The sequence of purine-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. 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.

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

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 there might be initiated a chemical reaction chain through which adjacent amino acids might be linked 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.

The basic thought underlying this scheme 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. Each amino acid may be tied to a nucleotide through a high energy bond, either a P or PP bond, and such acid anhydrides may release an energy of 12,000 calories or 16,000 calories respectively, when they split.

According to the notions here presented, amino acids can not read themselves the code of the paragene. But the purine and pyrimidine bases of the trinucleotides, which carry the proper amino acids, may attach through the formation of 6 hydrogen bonds to the proper sequence of three purine or pyrimidine bases on the paragene, and thus the amino acids may be lined up in the proper sequence along the paragene. Each amino acid being present in the form of an acid anhydride carries with it the energy that may be released in a chemical reaction chain that links adjacent amino acids through peptide bonds to each other.

Accordingly sequences of three nucleotides along the paragene represent the code words, and the trinucleotides which carry the amino acids would 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 guaninecorresponds to citosineand citosinecorresponds 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.[1] They showed that in a double stranded DNA structure, adenine pairs with thymine (which presumably plays the same role in DNA as does uracil in RNA) and guanine pairs with citosine. The helical structure of DNA permits just such pairing, and hydrogen bonding is possible between adenine and thymine as well as between guanine and citosine.

We may now tentatively adopt 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 on the paragene. When the trinucleotides are lined up in the proper order, then -- since each trinucleotide carries the proper amino acids -- the amino acids are also lined up in the proper order.

We shall now single out for more detailed examination one conceivable model for protein synthesis which might account 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 through an oxygen atom to the (2) carbon atom of the ribose moiety of each nucleotide (ester linkage) and one amino acid is attached to each of these phosphate (or diphosphate) groups. The amino acid anhydrides represent an energy-rich P (or PP) bond. [illegible]

(1) Proc. Roy. Soc., Vol. 223, p. 80, 1954.

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 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 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 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 made available /the energy needed for the formation of peptide bonds between adjacent amino acids.

Clearly, this model imposes certain restrictions on the possible amino acid sequences that paragenes can produce. What are these restrictions?

If we have four letters to choose from, then -- in a code which utilizes three-letter words -- we can construct 64 different words. We might, however, be restricted to the use of 20 out of the 64 words that are available. The reasons for this restriction would be as follows:

If all 64 possible three-letter combinations form in fact a code word, and if the nucleic acid strand assumes at the time of the formation of the polypeptide the helical configuration mentioned above, 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 64-word, three-letter-word, code any 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, assemble alongside the paragene simultaneously, rather than -- from one end/-- consecutively. If we want to have simultaneous assembly of the trimucleotides 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 ~~whith~~ the letters 1, 2, 3 on the template form a code word, and the letters 4, 5, 6 also form/code words/, but /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, that a code which requires no commas may be constructed and that it can accommodate 20 three-letter code words.

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 first letter and once attached to the last letter of the 20 (trimucleotide) 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.

Some of the amino acid sequences that may be found experimentally in sequential analysis of proteins and polypeptides might show that the restrictions imposed by our model on the possible amino acid sequence are too severe, and that the model has to be modified to accommodate established facts. As will be presently seen some modification of our model may be indicated for other reasons also. There is certainly no inherent reason why we should have a pure three-letter code -- as we have assumed above -- and why, for instance, a certain number of four-letter code words should not be utilized also.

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.

So far the number of amino acid residues in polypeptides and proteins have been determined only within one rather special class; all of them represent substances which are secreted by mammalian tissues. The number of amino acid residues found in such polypeptides and proteins, which have been analyzed with adequate accuracy, 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. In a mixed system of three- and four-letter words, one can obviously never draw the conclusion that <u>more than two</u> four-letter words have been included.

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, and that the rate of production of the enzyme goes up almost instantaneously upon adding the 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.

The highest rate at which this enzyme is produced in fully induced wild type bacteria growing in minimal medium is about $2\ 10^{-18}$ grams per cell per second.

Assuming a molecular weight of 100,000 for this enzyme, we would obtain a rate of about 15 enzyme molecules per second per bacterium. Or if we assume a molecular weight of a million (Jacques Monod and Melvin Cohen 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 a single strand of RNA, there might be a few paragenes present per cell rather than just one, and the number of parugenes might be of the order of magnitude of 10.

In these circumstances we are led to believe, on the basis of the figures given above, that for enzyme of high molecular weight - when the enzyme is fully induced and enzyme synthesis proceeds at its maximal rate -- the rate of formation of the enzyme is of the order of magnitude of one per second per paragene.

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 model that we have postulated. For the purposes of this computation, we shall assume that the molecular weight of the enzyme is about 100,000. We then have about 1,000 amino acid residues in the enzyme, and accordingly we would have to assemble alongside the paragene m = 300 trinucleotides, each of which is "loaded" with three amino acids.

In the approximation which we shall use, the minimum time, τ_0, needed for the formation of the polypeptide is composed of two terms, τ_1 and τ_2

$$\tau_0 = \tau_1 + \tau_2$$

After all the amino acids, assembled alongside the template, have been linked into one polypeptide, and if this polypeptide is at once removed, a certain time, τ_1, will elapse until the trinucleotides, which are now demuded of amino acids, evaporate from the template and their place is taken by trinucleotides which are loaded with the proper amino acids. We assume that the concentration of denuded trinucleotides in the cell is very small so that after the demuded trinucleotides evaporate, the loaded trinucleotides do not have to compete with denuded trinucleotides for their legitimate position along the paragene. The time, τ_1, which is necessary to permit evaporation of m denuded trinucleotides and to assemble m loaded trinucleotides in their pla we shall compute here on the assumption that once a loaded trinucleotide has found its position alongside the template, it will not evaporate again. Becaus this assumption is not valid, we must make a correction which is represented by the second term, τ_2.

Computation of τ_2

First we shall now compute this second term, τ_2. This computation is 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 have to deal with a reversible reaction which may be written as follows:

codeword + loadedtrinucleotide $\rightleftharpoons$ codeword—loadedtrinucleotide

The rate at which the reaction proceeds from left to right; i.e. the number of successful hits per unit time is given by

$$\text{hit rate} = A\rho$$

ρ denotes the concentration of the particular kind of loaded trinucleotide, in mol/cc. We shall, for the sake of simplicity, assume that the concentration in the cell of each kind of loaded trimucleotide is the same.

A stands for

$$A = 6 \cdot 10^{23} v \sigma p$$

v is the molecular velocity, $v = \sqrt{\frac{2RT}{\pi M}}$ where M is the molecular weight of the loaded trimucleotide. If $M \approx 1000$, we have $v \approx 5 \times 10^3$ cm/sec.

σ is the target area that must be hit by the loaded trimucleotide if Hydrogen bonding is to take place with the three adjacent nucleotides on the paragene. We may assume for σ a value of $\sigma = 10^{-15}$ cm^2.

p denotes the probability that the loaded trimucleotide, 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 , as a very rough estimate $p = 1/300$; This would give for σp the value $\sigma p = 1/3\ 10^{-17}$.

The rate at which the reaction proceeds from left to right is given by the rate α at which the codeword—loadedtrinucleotide complex dissociates or, as we may also say, the rate α at which the loaded trinucleotide evaporates from the paragene. For this rate we may write

(2) α = rate of evaporation $\approx 10^{13} e^{-\frac{\Delta H}{RT}}$

where ΔH is the binding energy for the loaded trinucleotides.

In equilibrium the hit rate and evaporation rate must be equal. The equilibrium constant, K, of this reversible reaction is defined as the concentration of the loaded trinucleotides at which the code word is covered half of the time. Thus we may write

(3) $$AK \approx \frac{1}{2} 10^{13} e^{-\frac{\Delta H}{RT}} = \frac{1}{2}\alpha$$

$$\alpha = 2AK$$

$\tau_w AK = \frac{1}{\tau_w}$

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

(4) $$f = \frac{1}{1 + \frac{\rho}{K}}$$

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

(5) $$\text{"number of gaps"} = \frac{m}{1 + \frac{\rho}{K}}$$

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

(6) $$\frac{\rho}{K} \gg 1$$

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_L, 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

Ihr Müller Daß zwei Entstörungen
er bezweifelte aber die existence der Erscheinungs

Aus meinem Leben „Wahrheit und Dichtung"

~~Man~~ Man soll auf seine Nase
nicht verlassen kann ja voran
kann man ~~noch~~ man sich den Nase
verlassen?

Ich habe jedoch nicht die Absicht
mich [nicht] wegen der Dichtung zu Entschuldigen ~~keine Wahrheit~~
denn ohne Dichtung gibt es ~~[illegible]~~
~~[illegible]~~ gibt ~~[illegible]~~ es aber keine Wahrheit
Tatsachen [illegible] Schnittmustern von Tatsachen.

so wollen wir hoffen

~~Die~~ Kristalline Form [illegible] gehört haben
Was sagt man [illegible] die Wahrheit in Statu nascendi.—

have to be filled consecutively, and the average time, $\frac{1}{A\rho}$, that it takes to fill one given gap. Thus, for τ_2 we may write

$$\tau_2 = \frac{1}{A\rho}\,\frac{m}{1+\frac{\rho}{K}} \tag{7}$$

Computation of τ_1 and $\tau_0 = \tau_1 + \tau_2$

When the polypeptide is formed and leaves the paragene, the code words are covered with the denuded trinucleotides. We may now compute the average time, , needed for the evaporation of all the denuded trinucleotides and the assembling of all the loaded trinucleotides in their place. We shall, for the sake of simplicity, assume that a denuded trinucleotide evaporates at the same rate, α, as a loaded trinucleotide. The rate, α, at which a loaded trinucleotide evaporates from the template is given by (3) and we may also write this in the form

$$\alpha = A\rho\,\frac{K}{\rho} \tag{8}$$

Since we have assumed $\frac{\rho}{K} \gg 1$, we have

$$A\rho \gg \alpha \tag{9}$$

As may be shown, for very large values of m, ($\ln m \gg 1$) we may write for

$$\tau_1 \approx \frac{1}{A\rho}\,\frac{\rho}{K}\,\ln m = \frac{1}{AK}\ln m \tag{10}$$

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

$$\tau_0 \approx \frac{1}{A\rho}\left\{\frac{m}{1+\frac{\rho}{K}} + \frac{\rho}{K}\ln m\right\} \tag{11}$$

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

$$\frac{\rho}{K} \approx \sqrt{\frac{m}{\ln m}} \tag{12}$$

Substituting this value into (11) gives

(13) $$\tau_0 \approx \frac{2}{A\rho}\sqrt{m \ln m}$$

see back of M.S.

For a polypeptide containing 1,000 amino acid residues, i.e. a paragene containing about 300 code words, we may write m = 300, and thus we obtain from (12) and (13)

(14) $$\frac{\rho}{K} \approx 10$$

and

(15) $$\tau_0 \approx \frac{50}{A\rho}$$

Estimates for the values of A and σ_p

If we estimate $\sigma_p = 1/3\ 10^{-13}$, then we have $A = 10^{10}$. For $\rho = 5.10^{-9}$ mol/cc ($\rho = 5.10^{-6}$ mol/liter) we then obtain $\tau_0 \approx 1$ sec., which means that one polypeptide is formed per paragene per second.

It might well be that σ_p is ten times higher: $\sigma_p = 1/3\ 10^{-12}$ which gives $A = 10^{11}$. In this case $\tau_0 \approx 1$ sec is obtained for $\rho = 5.10^{-10}$ mol/cc (or $\rho = 5.10^{-7}$ mol/liter).

If one were to attempt to find experimentally trinucleotides in bacterial cells, one would have to look for concentrations of this order of magnitude.

Since we have assumed $\frac{\rho}{K} \approx 10$, we must expect to find for K:

(16) $$5.10^{-8} < K \text{ (in mol/liter)} < 5.10^{-7}$$

or

$$5\ 10^{-11} < K \text{ (in mol/cc)} < 5\ 10^{-10}$$

From (3) we obtain for a given pair of values A and K the binding energy ΔH between the loaded trimucleotide and the code word on the paragene. For a value of $A = 10^{10}$ and $K = 10^{-7}$ mol/liter, or $K = 10^{-10}$ mol/cc, we obtain ΔH = 18,000 calories. Since 6 hydrogen bonds are involved, this would mean about 3,000 calories per hydrogen bond.

For the same value of A and a value of K which is ten times larger ΔH would decrease by about 1400 calories.

Conclusion

These considerations show that the theory which we postulated should be able to explain the high rate of enzyme synthesis which one observes in bacteria when the rate of formation of an enzyme is maximally enhanced by the use of an inducer. The basic thought of this theory consists in the assumption that trimucleotides read the code of the paragene and that these trimucleotides carry amino acids. One particular model for protein synthesis which assumed that each trimucleotide carries a sequence of three amino acids was singled out for detailed discussion because this model appeared to be the most plausible. This does not mean, however, that other models need not be considered.

Rather than to assume that each kind of trimucleotide carries a particular sequence of three amino acids, it would be in some ways more appealing to assume that each trimucleotide 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 leading or the trailing nucleotide. Assuming twenty different trimucleotides, 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.

The particular model for protein synthesis here considered cannot be modified simply by saying that each trinucleotide carries one amino acid instead of carrying three amino acids, for adjacent amino acids would not then be at the right distance from each other to be linked into a polypeptide.

Such an alternate model for protein synthesis would, therefore, require additional ideas concerning the configuration -- not necessarily a helical one -- which the paragene might adopt during protein synthesis.

I am grateful for the discussion of a variety of problems arising out of the considerations here presented which I had at the University of Chicago with Prof. Herbert S. Anker, Dr. Nandor L. Balazs, Mr. Hirondo Kuki, Prof. Joseph E. Mayer, and Prof. Leonard J. Savage.

$(x-1)$

$\frac{m}{\left(1+\frac{1}{x}\right)^{x}} + \frac{1}{2} \ln m$

$\frac{m}{x^2} = \ln m$

$x^2 = \frac{m}{\ln m}$

I $\sqrt{m \ln m}$

$x = \sqrt{\frac{m}{\ln m}}$

II $(x-1) \ln m$ $\sqrt{m \ln m}$

$T_0 = 2 \sqrt{m \ln m}$

kan.

June 7, 1957

HOW MAY 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

(communicated by Joseph E. Mayer June 10 1957.)

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~~ some what of a 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 ~~thereafter~~ a chemical reaction chain may link adjacent amino acids through peptide bonds with each other?

It is the purpose of ~~the present~~ this paper to ~~indicate~~ present 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.

~~A.L. Dounce, Enzymologia 15, 251, 1952~~

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

(1) ~~Abstract in proof:~~ by F.H. Crick, J.S. Griffith and L.E. Orgel which appeared in the May issue Proc. Nat. Acad of Sci. Vol 43 p 416 1957 containing a suggestion based on an oral communication by S. Brenner is

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 energ 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/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:

(2)

(2) Note added June 17th. The May issue of the Proc.Nat.Acad.Sci., Vol. 43, p. 416 (1957) which was belatedly received here, contains an article by F.H.Crick, J.S.Griffith, and L.E.Orgel which is in part identical with their previously circulated memorandum referred to in the text. In addition, however, the authors discuss -- in conjunction with a detailed oral communication of S. Brenner to them -- how amino acids might read the nucleotide code. The relevant passage of their paper is quoted in full in Appendix C.

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.(3) They showed that in a double stranded DNA

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

X ~~structure~~ helix, 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 utilise 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 ~~in~~ 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, say 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.(2) 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~~ contain 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 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 (at the (2) carbon atom of the ribose) to the phosphate group, they might well be assembled along the paragene at a 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~~ might 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~~ can 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.

Some of the amino acid sequences that may be found experimentally in sequential analysis of proteins and polypeptides might show that the restrictions imposed by our model on the possible amino acid sequence are too severe, and that the model has to be modified to accommodate established facts. As will be presently seen some modification of our model may be indicated for other reasons also. There is certainly no inherent reason why we should have a pure three-letter code -- as we have assumed above -- and why, for instance, a certain number of four-letter code words should not be utilized also.(4)

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.

So far the number of amino acid residues in non-cyclic polypeptides and proteins have been determined only within one rather special class; all of them represent substances which are secreted by mammalian tissues. The number of amino acid residues found in such polypeptides and proteins, which have been analyzed with adequate accuracy, 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. [~~In a mixed system of three- and four-letter words, one can obviously never draw the conclusion that more than two four-letter words have been included.~~]

(4) See Appendix A

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 coworkers 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. 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 in bacteria -- 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.

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 model that we have postulated. For the purposes of this computation, we shall assume that the molecular weight of the enzyme is about 100,000. We then have about 1,000 amino acid residues in the enzyme, and accordingly we would have to assemble alongside the paragene m = 300 trinucleotides, each of which is "loaded" with three amino acids.

In the approximation which we shall use, the minimum time, τ_0, needed for the formation of the polypeptide is composed of two terms, τ_1 and τ_2.

$$\tau_0 = \tau_1 + \tau_2$$

After all the amino acids, assembled alongside the template, have been linked into one polypeptide, and if this polypeptide is at once removed, a certain time, τ_1, will elapse until the trinucleotides, which are now denuded of amino acids, evaporate from the template and their place is taken by trinucleotides which are loaded with the proper amino acids. We assume that the concentration of denuded trinucleotides in the cell is very small so that after the denuded trinucleotides evaporate, the loaded trinucleotides do not have to compete with denuded trinucleotides for their legitimate positions along the paragene. The time, τ_1, which is necessary to permit evaporation of m denuded trinucleotides and to assemble m loaded trinucleotides in their places we shall compute here on the assumption that once a loaded trinucleotide has found its position alongside the template, it will not evaporate again. Because this assumption is not valid, we must make a correction which is represented by the second term, τ_2.

Computation of τ_2

First we shall now compute this second term, τ_2. This computation is to 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 have to deal with a reversible reaction which may be written as follows:

codeword + loadedtrinucleotide $\rightleftharpoons$ codeword—loadedtrinucleotide

The rate at which the reaction proceeds from left to right; i.e. the number of successful hits per unit time is given by

$$\text{hit rate} = A\rho$$

ρ denotes the concentration of the particular kind of loaded trinucleotide, in mol/cc. We shall, for the sake of simplicity, assume that the concentration in the cell of each kind of loaded trinucleotide is the same.

A stands for

$$A = 6 \cdot 10^{23} v \sigma p \qquad (11)$$

v is the molecular velocity, $v = \sqrt{\frac{2RT}{\pi M}}$ where M is the molecular weight of the loaded trinucleotide. If $M \approx 1000$, we have $v \approx 5 \times 10^3$ cm/sec.

σ is the target area that must be hit by the loaded trinucleotide if Hydrogen bonding is to take place with the three adjacent nucleotides on the paragene. We may assume for σ a value of $\sigma = 10^{-15}$ cm^2.

p 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, as a very rough estimate $p = 1/300$; This would give for σp the value $\sigma p = 1/3\ 10^{-17}$.

The rate at which the reaction proceeds from left to right is given by the rate α at which the codeword—loadedtrinucleotide complex dissociates or, as we may also say, the rate α at which the loaded trinucleotide evaporates from the paragene. For this rate we may write

(2) α = rate of evaporation $\approx 10^{13} e^{-\frac{\Delta H}{RT}}$

where ΔH is the binding energy for the loaded trinucleotides.

In equilibrium the hit rate and evaporation rate must be equal. The equilibrium constant, K, of this reversible reaction is defined as the concentration of the loaded trinucleotides at which the code word is covered half of the time. Thus we may write

(3) $$A K \approx \frac{1}{2} 10^{13} e^{-\frac{\Delta H}{RT}} = \frac{1}{2}\alpha$$

or

$$\alpha = 2AK$$

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

(4) $$f = \frac{1}{1 + \frac{\rho}{K}}$$

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

(5) $$\text{"number of gaps"} = \frac{m}{1 + \frac{\rho}{K}}$$

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

(6) $$\frac{\rho}{K} \gg 1$$

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

* one may verify that the reaction is not diffusion limited and that the concentration of the nucleotides

have to be filled consecutively, and the average time, $\frac{1}{A\rho}$, that it takes to fill one given gap. Thus, for τ_2 we may write

(7) $$\tau_2 = \frac{1}{A\rho}\,\frac{m}{1+\frac{\rho}{K}} \approx \frac{1}{50}\,30 = \frac{1}{2}$$

In the vicinity of the paragene is not appreciably smaller than ρ

Computation of τ_1 and $\tau_0 = \tau_1 + \tau_2$

When the polypeptide is formed and leaves the paragene, the code words are covered with the ~~demuded~~ stripped trinucleotides. We may now compute the average time, τ_1, needed for the evaporation of all the ~~demuded~~ stripped trinucleotides and the assembling of all the loaded trinucleotides in their place. We shall, for the sake of simplicity, assume that a ~~demuded~~ stripped trinucleotide evaporates at the same rate, α, as a loaded trinucleotide. The rate, α, at which a loaded trinucleotide evaporates from the template is given by (3) and we may also write this in the form

(8) $$\alpha = A\rho\,\frac{2K}{\rho} \approx 10$$

Since we have assumed $\frac{\rho}{K} \gg 1$, we have

(9) $$A\rho \gg \alpha$$

As may be shown, (5) for very large values of m, ($\ln m \gg 1$), we may write for τ_1

(10) $$\tau_1 \approx \frac{1}{A\rho}\,\frac{\rho}{2K}\,\ln m = \frac{1}{10}\,5.6$$

$\ln m = 1 + 2 + 2.3$ (5.6)

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

(11) $$\tau_0 = \frac{1}{A\rho}\left\{\frac{m}{1+\frac{\rho}{K}} + \frac{1}{2}\,\frac{\rho}{K}\,\ln m\right\}$$

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

(12) $$\frac{\rho}{K} \approx \sqrt{\frac{2m}{\ln m}}$$

(5) See Appendix B

Substituting this value into (11) gives

(13) $$\tau_0 \approx \frac{\sqrt{2}}{A\rho}\sqrt{m \ln m}$$

For a polypeptide containing 1,000 amino acid residues, i.e. a paragene containing about 300 code words, we may write m = 300, and thus we obtain from (12) and (13)

(14) $$\frac{\rho}{K} \approx 10$$

and

(15) $$\tau_0 \approx \frac{50}{A\rho} \text{ sec}$$ XXX

Estimates for the values of $\sigma\rho$, A, α and K XXX

As we may see from (15) we obtain $\tau_0 = 1$ sec if $A\rho = 50$. This means that for this particular value of $A\rho$ one enzyme molecule is produced per paragene per second. As we have seen before, this is the order of magnitude of the rate at which highly induced bacteria produce the enzyme β-galactosidase per paragene.

We shall, therefore, in the following assume $A\rho = 50$, and compute from it ρ, the concentration at which the different kinds of trinucleotides may be present in the cell.

If we use for $\sigma\rho$ the value of $\sigma\rho = \frac{1}{3}\,10^{-17}\,cm^2$, then we obtain from (1) A; $A = 10^{10}$ and accordingly we have $\rho = 5\cdot10^{-9}$ mol/cc ($\rho = 5\cdot10^{-6}$ mol/liter).

It might well be, however, that $\sigma\rho$ is ten times higher so that we have $\sigma\rho = \frac{1}{3}\,10^{-16}\,cm^2$, and then we obtain from (1) for A; $A = 10^{11}$, so that we have $\rho = 5\cdot10^{-10}$ mol/cc or ($\rho = 5\cdot10^{-7}$ mol/liter).

Thus the concentration, ρ, of the different kinds of trinucleotides in the cell is likely to be between 5×10^{-6} and 5×10^{-7} mol/liter, and one would have to look for concentrations of this order of magnitude in order to obtain experimental confirmation of their presence.

Since we have assumed $\frac{\rho}{K} \approx 10$, and since we believe that we have $A\rho = 50$, it follows that we must have for the rate of evaporation of the trinucleotides from the paragene, α

(16) $\alpha = 2AK = 10/\text{sec.}$ or $1/\alpha = 1/10$ sec.

For values of A between 10^{10} and 10^{11}, we have

(17) $5\ 10^{-11} < K \text{ (in mol/cc)} < 5\ 10^{-10}$

or

$5\ 10^{-8} \leq K \text{ (in mol/liter)} < 5\ 10^{-7}$

From (3) we obtain for a given pair of values A and K the binding energy ΔH between the loaded trinucleotide and the code word on the paragene. For a value of $A = 10^{10}$ and $K = 10^{-10}$ mol/cc($K = 10^{-7}$ mol/liter), we obtain ΔH 18,000 calories. Since 6 hydrogen bonds are involved, this would mean about 3,000 calories per hydrogen bond.

For the same value of A and a value of K which is ten times larger, ΔH would decrease by about 1400 calories.

Conclusion

These considerations show that the theory which we postulated should be able to explain the high rate of enzyme synthesis which one observes in bacteria when the rate of formation of an enzyme is maximally enhanced by the use of an inducer. The basic thought of this theory consists in the assumption that trinucleotides and possibly also tetranucleotides read the code of the paragene, and that these oligonucleotides carry amino acids. One particular model for protein synthesis, which assumed that each trinucleotide or tetranucleotide carries a sequence of three or four amino acids respectively, was singled out for detailed discussion because this model appeared to be the most plausible. This does not mean, however, that other models can not be considered.

For instance, rather than to assume that each kind of trinucleotide carries a particular sequence of three amino acids, one might wish to explore the possibility that each trinucleotide might carry 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 carbon of the 5 carbon sugar of either the leading or the 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.

However, the particular model for protein synthesis here considered cannot be modified by simply saying that each trinucleotide carries one amino acid instead of carrying three amino acids, for adjacent amino acids would not then be at the right distance from each other to be linked into a polypeptide. Such an alternate model for protein synthesis would, therefore, require additional ideas concerning the formation of the polypeptide.

I am grateful for the discussion of a variety of problems arising out of the considerations here presented which I had at the University of Chicago with Prof. Herbert S. Anker, Dr. Nandor L. Balazs, Mr. Hirondo Kuki, Prof. Joseph E. Mayer, and Prof. Leonard J. Savage.

APPENDIX

(Added June 17, 1957)

A.) If, in addition to ribose trinucleotides, carrying three amino acids, we have in the role of anti-code words also ribose tetranucleotides, carrying four amino acids, then we must postulate that only three of the bases of the tetranucleotide pair with complementary bases on the paragene. The fourth base, if it is a purine, must pair with the wrong pyrimidine, and if it is a pyrimidine, it must pair with the wrong purine. The reason for this is as follows:

As formula (17) shows, α, the rate of evaporation of a trinucleotide from the paragene can be estimated to be about 10/sec., corresponding to a binding energy, ΔH, of about 18,000 calories or 3,000 calories per hydrogen bond. If all four bases of the tetranucleotide were paired with the complementary bases on the paragene, we would then have two more hydrogen bonds and presumably an additional binding energy of about 6,000 calories. The equilibrium constant, K, of the tetranucleotide would therefore be lower than the value computed for K of the trinucleotides by a factor of about 10^4, and the rate of evaporation, α, would be lower by the same factor. This would then make, τ_0, the minimal time it takes for a paragene to form a polypeptide too long to be compatible with the observed rate of production of β-galactosidase in highly induced bacteria.

The number of words that may be constructed in a mixed three-letter word and four-letter word code of the kind described above, ~~if~~ where we demand that the code require no commas, has so far not been determined.

B.) The problem of computing τ_1 amounts to the solving of the following problem:

There are m boxes, each of which can hold one white ball or one black ball. Initially, at time, $\lambda = 0$, all of these boxes contain one

white ball -- a stripped trinucleotide. These white balls evaporate at the rate, α, so that at time λ the probability, $W(\lambda)$ of having no white ball in the box is given by

$$W(\lambda) = 1 - e^{-\alpha\lambda} \tag{18}$$

If the rate at which black balls fall into an empty box is designated by β, then the probability, y, that a given box does not contain a black ball at time, t, is given by

$$y(t) = e^{-\alpha t} + \int_{\lambda=0}^{\lambda=t} e^{-\beta(t-\lambda)} \frac{dW}{d\lambda} d\lambda \tag{19}$$

we may write

$$\int_{\lambda=0}^{\lambda=t} e^{-\beta(t-\lambda)} \frac{dW}{d\lambda} d\lambda = \frac{\alpha}{\beta-\alpha} \left[e^{-\alpha t} - e^{-\beta t} \right] \tag{20}$$

so that we have

$$y(t) = \frac{1}{\beta-\alpha} \left(\beta e^{-\alpha t} - \alpha e^{-\beta t} \right) \tag{21}$$

Thus we may write for the probability, x, that a given box contains a black ball

$$x(t) = 1 - y(t) = 1 - \frac{\beta}{\beta-\alpha} e^{-\alpha t} + \frac{\alpha}{\beta-\alpha} e^{-\beta t} \tag{22}$$

From this we obtain for P(t), the probability that all m boxes contain one black ball

$$P(t) = x(t)^m = \left(1 - \frac{\beta}{\beta-\alpha} e^{-\alpha t} + \frac{\alpha}{\beta-\alpha} e^{-\beta t}\right)^m \tag{23}$$

The average time, τ_1, needed for the evaporation of the white balls from all m boxes and the filling of all m boxes with black balls is given by

$$\tau_1 = \int_0^\infty t \frac{dP}{dt}\, dt \tag{24}$$

If $\beta \gg \alpha$, we may write

$$\tau_1 = \int_0^\infty t \frac{d}{dt}\left(1 - e^{-\alpha t}\right)^m dt \tag{25}$$

The expression $\frac{d}{dt}\left(1-e^{-\alpha t}\right)^m$ has a maximum at some value of t; $t = \tau_0$ and -- for large values of m -- it becomes small very rapidly both below and above $t = \tau_0$. Therefore, if m is large we may write

$$\tau_1 \approx \tau_0 \tag{26}$$

We obtain τ_0 by writing

$$\left\{ \frac{d^2}{dt^2}\left(1 - e^{-\alpha t}\right)^m \right\}_{t=\tau_0} = 0 \tag{27}$$

and from this we obtain

$$\tau_0 = \frac{1}{\alpha} \ln m \tag{28}$$

Thus for $\beta \gg \alpha$ and $m \gg 1$ we may write (26)

(29) $$\tau_1 \approx \frac{1}{\alpha} \ln m$$

It can be shown from (23) and (24) that in the next higher approximation we have

(30) $$\tau_1 \approx \frac{1}{\alpha}\{\ln(m+1) + C\} + \Delta \; ; \text{ where } \begin{cases} C = 0.577 \text{ (Euler's const.)} \\ \frac{1}{\beta} < \Delta < \frac{1}{\beta - \alpha} . \end{cases}$$

C.) F. H. Crick, J. S. Griffith and L. E. Orgel write in the May issue of the Proc. Nat. Acad. Sci. (Vol. 43, pp. 419 and 420, 1957):

|| To fix ideas, we shall describe a simple model to illustrate the advantages of such a code. Imagine that a single chain of RNA, held in a regular configuration, is the template. Let the intermediates in protein synthesis be 20 distinct molecules, each consisting of a trinucleotide chemically attached to one amino acid. The bases of each trinucleotide are chosen according to the code given above. Let these intermediate molecules combine, by hydrogen bonding between bases, with the RNA template and there await polymerization. Now imagine that such an amino acid-trinucleotide were to diffuse into an incorrect place on the template, such that two of its bases were hydrogen-bonded, though not the third. We postulate that this incomplete attachment will only retain the intermediate for a very brief time (for example, less than 1 millisecond) before the latter breaks loose and diffuses elsewhere. However, when it eventually diffuses to the correct place, it will be held by hydrogen bonds to all three bases and will thus be retained, on the average, for a much longer time (say, seconds or minutes). Now the code we have described insures that this more lengthy attachment can occur only at the points where the intermediate is needed. If one of the 20 intermediates could

stay for a long time on one of the false positions, it would effectively block the two positions it was straddling and hold up the polymerization process. Our code makes this impossible. This scheme, therefore, allows the intermediates to accumulate at the correct positions on the template without ever blocking the process by settling, except momentarily, in the wrong place. It is the feature which gives it an advantage over schemes in which the intermediates are compelled to combine with the template one after the other in the correct order.

The example given here is only for illustration, but it brings out the physical idea behind the concept of a comma-less code.

In passing, it should be mentioned that while the idea of making three nonoverlapping nucleotides code for one amino acid at first sight entails certain stereochemical difficulties, these are not insuperable if it is assumed that the polypeptide chain, when polymerized, does not remain attached to the template. A detailed scheme along these lines has been described to us by Dr. S. Brenner (personal communication).

A5547 — Aug-PNAS-Biochem-Szilard

June 7, 1957

X30
10 m/12 #8a
Use Webster's dictionary

10/12 caps ital #8c

HOW MAY AMINO ACIDS READ THE NUCLEOTIDE CODE?

Paper

10/12 enc #8a

By Leo Szilard

8/10 ital #8c

Communicated by Joseph E. Mayer

June 10, 1957

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

8/10 DC #8a

transpose

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 is an RNA molecule; and if the template is a DNA molecule, thymine takes the place of uracil.

letters stet only when underscored

It has remained so far somewhat of a 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.

transpose to p. 19

~~A. L. Dounce, Enzymologia 15, p. 251 (1952)~~
~~G. Gamow, Nature, Vol. 173, p. 318 (1954).~~

8/10 #8a #8c +2sp

A 1 V/# A. L. Dounce, Enzymologia, 15 (bf), 251, 1952; G. Gamow, Nature, 173, 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 the code of the paragene. But the trinucleotides, which carry the proper amino acids, may attach with their three bases through the formation of six 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:

transpose to p. 19

Proceedings, 1957

Note added June 17th. The May issue of the Proc. Nat. Acad. Sci., 43, 416, 1957, which was belatedly received here, contains an article by F.H.Crick, J.S.Griffith, and L.E.Orgel which is in part identical with their previously circulated memorandum referred to in the text. In addition, however, the authors discuss — in conjunction with a detailed oral communication of S. Brenner to them — how amino acids might read the nucleotide code. The relevant passage of their paper is quoted in full in Appendix C.

Transpose to f. 17

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.[3] They showed that in a double-stranded DNA 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.

[3] Nature, 173, 318, 1953; Proc. Roy. Soc. London, A, 223, 80, 1954.

If the sequence of bases along one strand of DNA represents a coded message which consists of 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, assembled 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.(2) 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 contain 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 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 distances from each other are smaller than or equal to, but not appreciably larger than, the fundamental repeating distance in a polypeptide chain, which is 7.27A. The fundamental repeating distance in a fully extended polypeptide chain is about 7A. Since before they are linked into a polypeptide, the amino acids can rotate around the chemical bond which ties them to the phosphate group (at the (2) carbon atom of the ribose), they might well be assembled along the paragene at a 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.

Some of the amino acid sequences that may be found experimentally in sequential analysis of proteins and polypeptides might show that the restrictions imposed by our model on the possible amino acid sequence are too severe, and that the model has to be modified to accommodate established facts. As will be presently seen, some modification of our model may be indicated for other reasons also. There is certainly no inherent reason why we should have a pure three-letter code as we have assumed above and why, for instance, a certain number of four-letter code words should not be utilized also. (4)

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.

So far the numbers of amino acid residues in non-cyclic polypeptides and proteins have been determined only within one rather special class; all of them represent substances which are secreted by mammalian tissues. The numbers of amino acid residues found in such polypeptides and proteins, which have been analyzed with adequate accuracy, 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 four-letter word and glucagon at least two four-letter words. ~~In a mixed system of three- and four-letter words, one can obviously never draw the conclusion that more than two four-letter words have been included.~~ (4) See Appendix A.

Transpose to p. 19

A Observed Rate of Enzyme Synthesis. em

According to the notions here adopted, most enzymes are synthetized in growing bacteria at a rather low rate which does not represent the maximum synthetising 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 in learning 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 per cent 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} gm. 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. 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, in bacteria — 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.

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 model that we have postulated. For the purposes of this computation, we shall assume that the molecular weight of the enzyme is about 100,000. We then have about 1,000 amino acid residues in the enzyme, and accordingly we would have to assemble alongside the paragene m = 300 trinucleotides, each of which is "loaded" with three amino acids.

In the approximation which we shall use, the minimum time, τ_0, needed for the formation of the polypeptide is composed of two terms, τ_1 and τ_2, $\tau_0 = \tau_1 + \tau_2$.

After all the amino acids, assembled alongside the template, have been linked into one polypeptide, if this polypeptide is at once removed, a certain time, τ_1, will elapse until the trinucleotides, which are now demuded of amino acids, evaporate from the template and their place is taken by trinucleotides which are loaded with the proper amino acids. We assume that the concentration of demuded trinucleotides in the cell is very small, so that after the demuded trinucleotides evaporate, the louded trinucleotides do not have to compete with demuded trinucleotides for their legitimate positions along the paragene. The time, τ_1, which is necessary to permit evaporation of m demuded trinucleotides and to assemble m loaded trinucleotides in their place we shall compute here on the assumption that once a loaded trinucleotide has found its position alongside the template, it will not evaporate again. Because this assumption is not valid, we must make a correction which is represented by the second term, τ_2.

Computation of τ_2

First we shall compute this second term, τ_2. This computation is based on the fact that (because of reevaporation of the loaded trinucleotides, which reversibly combine with the anti-code words of the paragene) there will always be no matter how long we wait a certain number of code words (sequences of three nucleotides) on the paragene which are not "covered".

We have to deal with a reversible reaction which may be written as follows:

codeword + loaded trinucleotide ⇌ codeword — loaded trinucleotide.

The rate at which the reaction proceeds from left to right, i.e., the number of successful hits per unit time, is given by

$$\text{Hit rate} = A\rho,$$

where ρ denotes the concentration of the particular kind of loaded trinucleotide, in mol/cm^3. We shall, for the sake of simplicity, assume that the concentration in the cell of each kind of loaded trinucleotide is the same.

A is defined by

$$A = 6 \times 10^{23} v \sigma p, \qquad (1)$$

where v is the molecular velocity, $v = \sqrt{\frac{2RT}{\pi M}}$, being the molecular weight of the loaded trinucleotide. If $M \approx 1000$, we have $v \approx 5 \times 10^3$ cm/sec.

σ is the target area that must be hit by the loaded trinucleotide if Hydrogen bonding is to take place with the three adjacent nucleotides on the parugene. We may assume for σ a value of 10^{-15} cm^2.

p 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, as a very rough estimate, 1/300; This would give for σp the value $\frac{1}{3} \times 10^{-17}$.

The rate at which the reaction proceeds from left to right is given by the rate α at which the codeword — loaded trinucleotide complex dissociates or, as we may also say, the rate at which the loaded trinucleotide evaporates from the paragene. For this rate we may write

(2) α = rate of evaporation $\approx 10^{13} e^{-\frac{\Delta H}{RT}}$,

where ΔH is the binding energy for the loaded trinucleotides.

(4) In equilibrium the hit rate and evaporation rate must be equal. The equilibrium constant, K, of this reversible reaction is defined as the concentration of the loaded trinucleotides at which the code word is covered half of the time. Thus we may write

(2) $$AK \approx \frac{1}{2} \times 10^{13} e^{-\frac{\Delta H}{RT}} = \frac{1}{2}\alpha$$

or

$$\alpha = 2AK.$$

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

(4) $$f = \frac{1}{1 + \frac{\rho}{K}}.$$

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

(5) $$\text{"Number of gaps"} = \frac{m}{1 + \frac{\rho}{K}}.$$

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

(6) $$\frac{\rho}{K} \gg 1.$$

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, $\frac{1}{A\rho}$, that it takes to fill one given gap. Thus, for τ_2, we may write

$$\tau_2 = \frac{1}{A\rho}\,\frac{m}{1+\frac{\rho}{K}}. \tag{7}$$

Computation of τ_1 and $\tau_0 = \tau_1 + \tau_2$.

When the polypeptide is formed and leaves the paragene, the code words are covered with the denuded trinucleotides. We may now compute the average time, τ_1, needed for the evaporation of all the denuded trinucleotides and the assembling of all the loaded trinucleotides in their place. We shall, for the sake of simplicity, assume that a denuded trinucleotide evaporates at the same rate, α, as a loaded trinucleotide. The rate, α, at which a loaded trinucleotide evaporates from the template is given by equation (3), and we may also write this in the form

$$\alpha = A\rho\,\frac{2K}{\rho}. \tag{8}$$

Since we have assumed $\frac{\rho}{K} \gg 1$, we have

$$A\rho \gg \alpha. \tag{9}$$

As may be shown, [redacted] for very large values of m, ($\ln m \gg 1$), we may write for τ_1

$$\tau_1 \approx \frac{1}{A\rho}\,\frac{\rho}{2K}\,\ln m. \tag{10}$$

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

$$\tau_0 = \frac{1}{A\rho}\left\{\frac{m}{1+\frac{\rho}{K}} + \frac{1}{2}\,\frac{\rho}{K}\,\ln m\right\}. \tag{11}$$

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

$$\frac{\rho}{K} \approx \sqrt{\frac{2m}{\ln m}}. \tag{12}$$

(5) See Appendix B. Transpose to p. 19

Substituting this value into equation (11) gives

$$\tau_0 \approx \frac{\sqrt{2}}{A\rho}\sqrt{m \ln m}. \tag{13}$$

For a polypeptide containing 1,000 amino acid residues, i.e., a paragene containing about 300 code words, we may write $m = 300$, and thus we obtain from relations (12) and (13)

$$\frac{\rho}{K} \approx 10 \tag{14}$$

and

$$\tau_0 \approx \frac{50}{A\rho}. \tag{15}$$

Estimates for the Values of $\sigma\rho$, A, a, and K.

As we may see from relation (15), we obtain $\tau_0 = 1$ if $A\rho = 50$. This means that for this particular value of $A\rho$ one enzyme molecule is produced per paragene per second. As we have seen before, this is the order of magnitude of the rate at which highly induced bacteria produce the enzyme β-galactosidase per paragene.

We shall, therefore, in the following, assume $A\rho = 50$, and compute from it ρ, the concentration at which the different kinds of trinucleotides may be present in the cell.

If we use for $\sigma\rho$ the value $\frac{1}{3}\times 10^{-17}$ cm^2, then we obtain from equation (1) $A = 10^{10}$, and, accordingly, we have $\rho = 5\times 10^{-9}$ mol/cm^3 ($= 5\times 10^{-6}$ mol/liter).

It might well be, however, that $\sigma\rho$ is ten times higher, so that we have $\sigma\rho = \frac{1}{3}\times 10^{-16}$ cm^2, and then we obtain from equation (1) $A = 10^{11}$, so that we have $\rho = 5\times 10^{-10}$ mol/cm^3, or 5×10^{-7} mol/liter.

Thus the concentration, ρ, of the different kinds of trinucleotides in the cell is likely to be between 5×10^{-6} and 5×10^{-7} mol/liter, and one would have to look for concentrations of this order of magnitude in order to obtain experimental confirmation of their presence.

Since we have assumed $\frac{\rho}{K} \approx 10$, and since we believe that we have $\Delta\rho = 50$, it follows that we must have, for the rate of evaporation of the trinucleotides from the paragene,

(16) $$\alpha = 2AK = 10 \text{ sec}^{-1} \quad \text{or} \quad 1/\alpha = 1/10 \text{ sec.}$$

For values of A between 10^{10} and 10^{11}, we have

(17) $$5\times10^{-11} < K \text{ (in mol/cm}^3\text{)} < 5\times10^{-10}$$

or

$$5\times10^{-8} \leq K \text{ (in mol/liter)} < 5\times10^{-7}.$$

From equation (3) we obtain for a given pair of values A and K the binding energy ΔH between the loaded trinucleotide and the code word on the paragene. For a value of $A = 10^{10}$ and $K = 10^{-10}$ mol/cm^3 ($K = 10^{-7}$ mol/liter), we obtain ΔH 18,000 calories. Since six hydrogen bonds are involved, this would mean about 3,000 calories per hydrogen bond.

For the same value of A and a value of K which is ten times larger, ΔH would decrease by about 1,400 calories.

Conclusion.

These considerations show that the theory which we postulated should be able to explain the high rate of enzyme synthesis which one observes in bacteria when the rate of formation of an enzyme is maximally enhanced by the use of an inducer. The basic thought of this theory consists in the assumption that trinucleotides and possibly also tetranucleotides read the code of the paragene, and that these oligonucleotides carry amino acids. One particular model for protein synthesis, which assumed that each trinucleotide or tetranucleotide carries a sequence of three or four amino acids, respectively, was singled out for detailed discussion because this model appeared to be the most plausible. This does not mean, however, that other models can not be considered.

For instance, rather than assume that each kind of trinucleotide carries a particular sequence of three amino acids, one might wish to explore the possibility that each trinucleotide might carry only one amino acid.(2) 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 carbon of the 5 carbon sugar of either the leading or the trailing nucleotide. Assuming 20 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.

However, the particular model for protein synthesis here considered cannot be modified by simply saying that each trinucleotide carries one amino acid instead of carrying three amino acids, for adjacent amino acids would not then be at the right distance from each other to be linked into a polypeptide. Such an alternate model for protein synthesis would, therefore, require additional ideas concerning the formation of the polypeptide.

12 pr#

I am grateful for the discussion of a variety of problems arising out of the considerations here presented which I had at the University of Chicago with Professor Herbert S. Anker, Dr. Nandor L. Balazs, Mr. Hirondo Kuki, Professor Joseph E. Mayer, and Professor Leonard J. Savage.

Appendix follows

set APPENDIX in 9 pt. type

MDH
6/14/57

To PCo.:

The formulas on
pp. 16-18
could be set up in
10 pt (the same
as regular formulas
in text) if it would
make assembly or
composition easier
(just the formulas -
text should still
be 9/11)

APPENDIXES

(Added June 17, 1957)

APPENDIX A

If, in addition to ribose trinucleotides, carrying three amino acids, we have in the role of anti-code words also ribose tetranucleotides, carrying four amino acids, then we must postulate that only three of the bases of the tetranucleotide pair with complementary bases on the paragene. The fourth base, if it is a purine, must pair with the wrong pyrimidine, and if it is a pyrimidine, it must pair with the wrong purine. The reason for this is as follows:

As formula (17) shows, α, the rate of evaporation of a trinucleotide from the paragene can be estimated to be about 10^{-1} sec^{-1}, corresponding to a binding energy, ΔH, of about 18,000 calories, or 3,000 calories per hydrogen bond. If all four bases of the tetranucleotide were paired with the complementary bases on the paragene, we would then have two more hydrogen bonds and presumably an additional binding energy of about 6,000 calories. The equilibrium constant, K, of the tetranucleotide would therefore be lower than the value computed for K of the trinucleotides by a factor of about 10^4, and the rate of evaporation, α, would be lower by the same factor. This would then make τ_0, the minimal time it takes for a paragene to form a polypeptide, too long to be compatible with the observed rate of production of β-galactosidase in highly induced bacteria.

The number of words that may be constructed in a mixed three-letter-word and four-letter-word code of the kind described above, where we demand that the code require no commas, has so far not been determined.

APPENDIX B

The problem of computing τ_1 amounts to the solving of the following problem:

There are m boxes, each of which can hold one white ball or one black ball. Initially, at time $\lambda = 0$, all these boxes contain one

white ball -- a stripped trinucleotide. These white balls evaporate at the rate, α, so that at time λ the probability, $W(\lambda)$, of having no white ball in the box is given by

$$W(\lambda) = 1 - e^{-\alpha\lambda}. \tag{18}$$

If the rate at which black balls fall into an empty box is designated by β, then the probability, y, that a given box does not contain a black ball at time, t, is given by

$$y(t) = e^{-\alpha t} + \int_{\lambda=0}^{\lambda=t} e^{-\beta(t-\lambda)} \frac{dW}{d\lambda} d\lambda. \tag{19}$$

We may write

$$\int_{\lambda=0}^{\lambda=t} e^{-\beta(t-\lambda)} \frac{dW}{d\lambda} d\lambda = \frac{\alpha}{\beta-\alpha}\left[e^{-\alpha t} - e^{-\beta t}\right], \tag{20}$$

so that we have

$$y(t) = \frac{1}{\beta-\alpha}\left(\beta e^{-\alpha t} - \alpha e^{-\beta t}\right). \tag{21}$$

Thus we may write, for the probability, x, that a given box contains a black ball,

$$x(t) = 1 - y(t) = 1 - \frac{\beta}{\beta-\alpha} e^{-\alpha t} + \frac{\alpha}{\beta-\alpha} e^{-\beta t}. \tag{22}$$

From this we obtain for $P(t)$, the probability that all m boxes contain one black ball,

$$P(t) = x(t)^m = \left(1 - \frac{\beta}{\beta-\alpha} e^{-\alpha t} + \frac{\alpha}{\beta-\alpha} e^{-\beta t}\right)^m. \tag{23}$$

The average time, τ_1, needed for the evaporation of the white balls from all m boxes and the filling of all m boxes with black balls is given by

$$\tau_1 = \int_0^\infty t \frac{dP}{dt}\, dt. \tag{24}$$

If $\beta \gg \alpha$, we may write

$$\tau_1 = \int_0^\infty t \frac{d}{dt}\left(1 - e^{-\alpha t}\right)^m dt. \tag{25}$$

The expression $\frac{d}{dt}(1 - e^{-\alpha t})^m / dt$ has a maximum at some value of t, $t = \tau_0$, and for large values of m it becomes small very rapidly both below and above $t = \tau_0$. Therefore, if m is large, we may write

$$\tau_1 \approx \tau_0. \tag{26}$$

We obtain τ_0 by writing

$$\left\{\frac{d^2}{dt^2}\left(1 - e^{-\alpha t}\right)^m\right\}_{t=\tau_0} = 0, \tag{27}$$

and from this we obtain

$$\tau_0 = \frac{1}{\alpha} \ln m. \tag{28}$$

Thus, for $\beta \gg \alpha$ and $m \gg 1$, we may write equation (26) as

$$\tau_1 \approx \frac{1}{\alpha} \ln m. \tag{29}$$

It can be shown from equations (23) and (24) that in the next higher approximation we have

$$\tau_1 \approx \frac{1}{\alpha}\left\{\ln(m+1) + C\right\} + \Delta \tag{30}$$

where $C = 0.577$ (Euler's const.), $\frac{1}{\beta} < \Delta < \frac{1}{\beta - \alpha}$.

APPENDIX C

F. H. Crick, J. S. Griffith, and L. E. Orgel write in the May issue of the Proceedings of the Nat. Acad. Sci. (43, 416 and 420, 1957):

"To fix ideas, we shall describe a simple model to illustrate the advantages of such a code. Imagine that a single chain of RNA, held in a regular configuration, is the template. Let the intermediates in protein synthesis be 20 distinct molecules, each consisting of a trinucleotide chemically attached to one amino acid. The bases of each trinucleotide are chosen according to the code given above. Let these intermediate molecules combine, by hydrogen bonding between bases, with the RNA template and there await polymerization. Now imagine that such an amino acid-trinucleotide were to diffuse into an incorrect place on the template, such that two of its bases were hydrogen-bonded, though not the third. We postulate that this incomplete attachment will only retain the intermediate for a very brief time (for example, less than 1 millisecond) before the latter breaks loose and diffuses elsewhere. However, when it eventually diffuses to the correct place, it will be held by hydrogen bonds to all three bases and will thus be retained, on the average, for a much longer time (say, seconds or minutes). Now the code we have described insures that this more lengthy attachment can occur only at the points where the intermediate is needed. If one of the 20 intermediates could

stay for a long time on one of the false positions, it would effectively block the two positions it was straddling and hold up the polymerization process. Our code makes this impossible. This scheme, therefore, allows the intermediates to accumulate at the correct positions on the template without ever blocking the process by settling, except momentarily, in the wrong place. It is this feature which gives it an advantage over schemes in which the intermediates are compelled to combine with the template one after the other in the correct order.

The example given here is only for illustration, but it brings out the physical idea behind the concept of a comma-less code.

In passing, it should be mentioned that while the idea of making three nonoverlapping nucleotides code for one amino acid at first sight entails certain stereochemical difficulties, these are not insuperable if it is assumed that the polypeptide chain, when polymerized, does not remain attached to the template. A detailed scheme along these lines has been described to us by Dr. S. Brenner (personal communication)."

closing quote

6 pt #

insert notes from pp. 1, 2, 3, 6, #11

#

June 7, 1957

HOW MAY 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 somewhat of a mystery in just what 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) A. L. Dounce, Enzymologia, 15, 251 (1952)
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[2] 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/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:

(2) Note added June 17th. The May issue of the Proc.Nat.Acad.Sci., Vol. 43, p. 416 (1957) which was belatedly received here, contains an article by F.H.Crick, J.S.Griffith, and L.E.Orgel which is in part identical with their previously circulated memorandum referred to in the text. In addition, however, the authors discuss -- in conjunction with a detailed oral communication of S. Brenner to them -- how amino acids might read the nucleotide code. The relevant passage of their paper is quoted in full in Appendix C.

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, say 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.(2) 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 contain 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 ~~MMM~~ 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 (at the (2) carbon atom of the ribose) 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.

Some of the amino acid sequences that may be found experimentally in sequential analysis of proteins and polypeptides might show that the restrictions imposed by our model on the possible amino acid sequence are too severe, and that the model has to be modified to accommodate established facts. As will be presently seen some modification of our model may be indicated for other reasons also. There is certainly no inherent reason why we should have a pure three-letter code -- as we have assumed above -- and why, for instance, a certain number of four-letter code words should not be utilized also.(4)

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.

So far the number of amino acid residues in non-cyclic polypeptides and proteins have been determined only within one rather special class; all of them represent substances which are secreted by mammalian tissues. The number of amino acid residues found in such polypeptides and proteins, which have been analyzed with adequate accuracy, 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. [illegible]

(4) See Appendix A

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 synthetising 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. 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 in bacteria -- 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.

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 model that we have postulated. For the purposes of this computation, we shall assume that the molecular weight of the enzyme is about 100,000. We then have about 1,000 amino acid residues in the enzyme, and accordingly we would have to assemble alongside the paragene m = 300 trinucleotides, each of which is "loaded" with three amino acids.

In the approximation which we shall use, the minimum time, τ_0, needed for the formation of the polypeptide is composed of two terms, τ_1 and τ_2;

$$\tau_0 = \tau_1 + \tau_2$$

After all the amino acids, assembled alongside the template, have been linked into one polypeptide, and if this polypeptide is at once removed, a certain time, τ_1, will elapse until the trinucleotides, which are now denuded of amino acids, evaporate from the template and their place is taken by trinucleotides which are loaded with the proper amino acids. We shall assume that the concentration of denuded trinucleotides in the cell is very small so that after the denuded trinucleotides evaporate, the loaded trinucleotides do not have to compete with denuded trinucleotides for their legitimate positions along the paragene. The time, τ_1, which is necessary to permit evaporation of m denuded trinucleotides and to assemble m loaded trinucleotides in their place we shall compute here on the assumption that once a loaded trinucleotide has found its position alongside the template, it will not evaporate again. Because this assumption is not valid, we must make a correction which is represented by the second term, τ_2.

Computation of τ_1

First we shall now compute this second term, τ_1. This computation is 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 have to deal with a reversible reaction which may be written as follows:

codeword + loadedtrinucleotide $\rightleftharpoons$ codeword—loadedtrinucleotide

The rate at which the reaction proceeds from left to right; i.e. the number of successful hits per unit time is given by

$$\text{hit rate} = A\rho$$

ρ denotes the concentration of the particular kind of loaded trinucleotide/ in mol/cc. We shall, for the sake of simplicity, assume that the concentration in the cell of each kind of loaded trinucleotide is the same.

A stands for

(1) $$A = 6 \cdot 10^{23} v \sigma p$$

v is the molecular velocity, $v = \sqrt{\frac{2RT}{\pi M}}$ where M is the molecular weight of the loaded trinucleotide. If $M \approx 1000$, we have $v \approx 5 \times 10^3$ cm/sec.

σ is the target area that must be hit by the loaded trinucleotide if Hydrogen bonding is to take place with the three adjacent nucleotides on the parugene. We may assume for σ a value of $\sigma = 10^{-15}$ cm^2.

p 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 , as a very rough estimate $p = 1/300$; This would give for σp the value $\sigma p = 1/3\ 10^{-15}$.

The rate at which the reaction proceeds from left to right is given by the rate α at which the codeword—loadedtrinucleotide complex dissociates or, as we may also say, the rate α at which the loaded trinucleotide evaporates from the paragene. For this rate we may write

$$A = 6\,10^{23}\ 5\,10^{3}\ 10^{-15}\ \frac{1}{300} =$$

$$10^{22}\ 10^{3}\ 10^{-15} = 10^{+10}$$

(2) α = rate of evaporation $\approx 10^{13} e^{-\frac{\Delta H}{RT}}$

where ΔH is the binding energy for the loaded trinucleotides.

In equilibrium the hit rate and evaporation rate must be equal. The equilibrium constant, K, of this reversible reaction is defined as the concentration of the loaded trinucleotides at which the code word is covered half of the time. Thus we may write

(3) $$A K \approx \frac{1}{2} 10^{13} e^{-\frac{\Delta H}{RT}} = \frac{1}{2}\alpha$$

or

$$\alpha = 2AK$$

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

(4) $$f = \frac{1}{1 + \frac{\rho}{K}}$$

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

(5) $$\text{"number of gaps"} = \frac{m}{1 + \frac{\rho}{K}}$$

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

(6) $$\frac{\rho}{K} \gg 1$$

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_L, 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 averuge time, $\frac{1}{A\rho}$, that it takes to fill one given gap. Thus, for τ_2 we may write

$$\tau_2 = \frac{1}{A\rho}\,\frac{m}{1+\frac{\rho}{K}} \tag{7}$$

Computation of τ_1 and $\tau_0 = \tau_1 + \tau_2$

When the polypeptide is formed and leaves the paragene, the code words are covered with the denuded trinucleotides. We may now compute the average time, τ_1, needed for the evaporation of all the denuded trinucleotides and the assembling of all the loaded trinucleotides in their place. We shall, for the sake of simplicity, assume that a denuded trinucleotide evaporates at the same rate, α, as a loaded trinucleotide. The rate, α, at which a loaded trinucleotide evaporates from the template is given by (3) and we may also write this in the form

$$\alpha = A\rho\,\frac{2K}{\rho} \tag{8}$$

Since we have assumed $\frac{\rho}{K} \gg 1$, we have

$$A\rho \gg \alpha \tag{9}$$

As may be shown,(5) for very large values of m, ($\ln m \gg 1$) we may write for τ_1

$$\tau_1 \approx \frac{1}{A\rho}\,\frac{\rho}{2K}\,\ln m \tag{10}$$

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

$$\tau_0 = \frac{1}{A\rho}\left\{\frac{m}{1+\frac{\rho}{K}} + \frac{1}{2}\,\frac{\rho}{K}\,\ln m\right\} \tag{11}$$

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

$$\frac{\rho}{K} \approx \sqrt{\frac{2m}{\ln m}} \tag{12}$$

(5) See Appendix B

Substituting this value into (11) gives

(13) $$\tau_0 \approx \frac{\sqrt{2}}{A\rho}\sqrt{m \ln m}$$

For a polypeptide containing 1,000 amino acid residues, i.e. a paragene containing about 300 code words, we may write m = 300, and thus we obtain from (12) and (13)

(14) $$\frac{\rho}{K} \approx 10$$

and

(15) $$\tau_0 \approx \frac{50}{A\rho}$$

Estimates for the values of $\sigma\rho$, A, α and K

As we may see from (15) we obtain $\tau_0 = 1$ if $A\rho = 50$. This means that for this particular value of $A\rho$ one enzyme molecule is produced per paragene per second. As we have seen before, this is the order of magnitude of the rate at which highly induced bacteria produce the enzyme β-galactosidase per paragene.

We shall, therefore, in the following assume $A\rho = 50$, and compute from it ρ, the concentration at which the different kinds of trinucleotides may be present in the cell.

If we use for $\sigma\rho$ the value of $\sigma\rho = \frac{1}{3}\,10^{-17}\ cm^2$, then we obtain from (1) A; $A = 10^{10}$ and accordingly we have $\rho = 5\cdot10^{-9}$ mol/cc ($\rho = 5\cdot10^{-6}$ mol/liter).

It might well be, however, that $\sigma\rho$ is ten times higher so that we have $\sigma\rho = \frac{1}{3}\,10^{-16}\ cm^2$, and then we obtain from (1) for A; $A = 10^{11}$, so that we have $\rho = 5\cdot10^{-10}$ mol/cc or ($\rho = 5\cdot10^{-7}$ mol/liter).

Thus the concentration, ρ, of the different kinds of trinucleotides in the cell is likely to be between 5×10^{-6} and 5×10^{-7} mol/liter, and one would have to look for concentrations of this order of magnitude in order to obtain experimental confirmation of their presence.

Since we have assumed $\frac{A}{K} \approx 10$, and since we believe that we have $A_j = 50$, it follows that we must have for the rate of evaporation of the trinucleotides from the paragene, α

(16) $\alpha = 2AK = 10/\text{sec.}$ or $1/\alpha = 1/10$ sec.

For values of A between 10^{10} and 10^{11}, we have

(17) $5\ 10^{-11} < K \text{ (in mol/cc)} < 5\ 10^{-10}$

or

$5\ 10^{-8} \leq K \text{ (in mol/liter)} < 5\ 10^{-7}$

From (3) we obtain for a given pair of values A and K the binding energy ΔH between the loaded trinucleotide and the code word on the paragene. For a value of $A = 10^{10}$ and $K = 10^{-10}$ mol/cc ($K = 10^{-7}$ mol/liter), we obtain ΔH 18,000 calories. Since 6 hydrogen bonds are involved, this would mean about 3,000 calories per hydrogen bond.

For the same value of A and a value of K which is ten times larger, ΔH would decrease by about 1400 calories.

Conclusion

These considerations show that the theory which we postulated should be able to explain the high rate of enzyme synthesis which one observes in bacteria when the rate of formation of an enzyme is maximally enhanced by the use of an inducer. The basic thought of this theory consists in the assumption that trinucleotides and possibly also tetranucleotides read the code of the paragene, and that these oligonucleotides carry amino acids. One particular model for protein synthesis, which assumed that each trinucleotide or tetranucleotide carries a sequence of three or four amino acids respectively, was singled out for detailed discussion because this model appeared to be the most plausible. This does not mean, however, that other models can not be considered.

For instance, rather than to assume that each kind of trinucleotide carries a particular sequence of three amino acids, one might wish to explore the possibility that each trinucleotide might carry 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 carbon of the 5 carbon sugar of either the leading or the 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.

However, the particular model for protein synthesis here considered cannot be modified by simply saying that each trinucleotide carries one amino acid instead of carrying three amino acids, for adjacent amino acids would not then be at the right distance from each other to be linked into a polypeptide. Such an alternate model for protein synthesis would, therefore, require additional ideas concerning the formation of the polypeptide.

I am grateful for the discussion of a variety of problems arising out of the considerations here presented which I had at the University of Chicago with Prof. Herbert S. Anker, Dr. Nandor L. Balazs, Mr. Hirondo Kuki, Prof. Joseph E. Mayer, and Prof. Leonard J. Savage.

Leo Szilard

APPENDIX

(Added June 17/57)

A.) If, in addition to ribose trinucleotides, carrying three amino acids, we have in the role of anti-code words also ribose tetranucleotides, carrying four amino acids, then we must postulate that only three of the bases of the tetranucleotide pair with complementary bases on the parugene. The fourth base, if it is a purine, must pair with the wrong pyrimidine, and if it is a pyrimidine, it must pair with the wrong purine. The reason for this is as follows:

As formula (17) shows, α, the rate of evaporation of a trinucleotide from the parugene can be estimated to be about 10/sec., corresponding to a binding energy, ΔH, of about 18,000 calories or 3,000 calories per hydrogen bond. If all four bases of the tetranucleotide were paired with the complementary bases on the parugene, we would then have two more hydrogen bonds and presumably an additional binding energy of about 6,000 calories. The equilibrium constant, K, of the tetranucleotide would therefore be lower than the value computed for K of the trinucleotides by a factor of about 10^4, and the rate of evaporation, α, would be lower by the same factor. This would then make, τ_0, the minimal time it takes for a parugene to form a polypeptide too long to be compatible with the observed rate of production of β-galactosidase in highly induced bacteria.

The number of words that may be constructed in a mixed three-letter word and four-letter word code of the kind described above, WHERE we demand that the code require no commas, has so far not been determined.

B.) The problem of computing τ_1 amounts to the solving of the following problem:

There are m boxes, each of which can hold one white ball or one black ball. Initially, at time, $\lambda = 0$, all of these boxes contain one

white ball -- a stripped trinucleotide. These white balls evaporate at the rate, α, so that at time λ the probability, $W(\lambda)$ of having no white ball in the box is given by

$$W(\lambda) = 1 - e^{-\alpha\lambda} \tag{18}$$

If the rate at which black balls fall into an empty box is designated by β, then the probability, y, that a given box does not contain a black ball at time, t, is given by

$$y(t) = e^{-\alpha t} + \int_{\lambda=0}^{\lambda=t} e^{-\beta(t-\lambda)} \frac{dW}{d\lambda} d\lambda \tag{19}$$

we may write

$$\int_{\lambda=0}^{\lambda=t} e^{-\beta(t-\lambda)} \frac{dW}{d\lambda} d\lambda = \frac{\alpha}{\beta-\alpha} \left[e^{-\alpha t} - e^{-\beta t} \right] \tag{20}$$

so that we have

$$y(t) = \frac{1}{\beta-\alpha} \left(\beta e^{-\alpha t} - \alpha e^{-\beta t} \right) \tag{21}$$

Thus we may write for the probability, x, that a given box contains a black ball

$$x(t) = 1 - y(t) = 1 - \frac{\beta}{\beta-\alpha} e^{-\alpha t} + \frac{\alpha}{\beta-\alpha} e^{-\beta t} \tag{22}$$

From this we obtain for P(t), the probability that all m boxes contain one black ball

$$P(t) = x(t)^m = \left(1 - \frac{\beta}{\beta-\alpha} e^{-\alpha t} + \frac{\alpha}{\beta-\alpha} e^{-\beta t}\right)^m \tag{23}$$

The average time, τ_1, needed for the evaporation of the white balls from all m boxes and the filling of all m boxes with black balls is given by

$$\tau_1 = \int_0^\infty t \frac{dP}{dt}\, dt \tag{24}$$

If $\beta \gg \alpha$, we may write

$$\tau_1 = \int_0^\infty t \frac{d}{dt}\left(1 - e^{-\alpha t}\right)^m dt \tag{25}$$

The expression $\frac{d}{dt}\left(1-e^{-\alpha t}\right)^m$ has a maximum at some value of t; $t = \tau_0$ and -- for large values of m -- it becomes small very rapidly both below and above $t = \tau_0$. Therefore, if m is large we may write

$$\tau_1 \approx \tau_0 \tag{26}$$

We obtain τ_0 by writing

$$\left\{ \frac{d^2}{dt^2}\left(1 - e^{-\alpha t}\right)^m \right\}_{t=\tau_0} = 0 \tag{27}$$

and from this we obtain

$$\tau_0 = \frac{1}{\alpha} \ln m \tag{28}$$

Thus for $\beta \gg \alpha$ and $m \gg 1$ we may write (26)

(29) $$\tau_1 \approx \frac{1}{\alpha} \ln m$$

It can be shown from (25) and (24) that in the next higher approximation we have

(30) $$\tau_1 \approx \frac{1}{\alpha}\{\log(m+1) + C\} + \Delta \; ; \text{ where } \begin{cases} C = 0.577 & \text{(Euler's const.)} \\ \frac{1}{\beta} < \Delta < \frac{1}{\beta - \alpha} . \end{cases}$$

C.) F. H. Crick, J. S. Griffith and L. E. Orgel write in the May issue of the Proc. Nat. Acad. Sci. (Vol. 43, pp. 419 and 420, 1957):

"To fix ideas, we shall describe a simple model to illustrate the advantages of such a code. Imagine that a single chain of RNA, held in a regular configuration, is the template. Let the intermediates in protein synthesis be 20 distinct molecules, each consisting of a trinucleotide chemically attached to one amino acid. The bases of each trinucleotide are chosen according to the code given above. Let these intermediate molecules combine, by hydrogen bonding between bases, with the RNA template and there await polymerization. Now imagine that such an amino acid-trinucleotide were to diffuse into an incorrect place on the template, such that two of its bases were hydrogen-bonded, though not the third. We postulate that this incomplete attachment will only retain the intermediate for a very brief time (for example, less than 1 millisecond) before the latter breaks loose and diffuses elsewhere. However, when it eventually diffuses to the correct place, it will be held by hydrogen bonds to all three bases and will thus be retained, on the average, for a much longer time (say, seconds or minutes). Now the code we have described insures that this more lengthy attachment can occur only at the points where the intermediate is needed. If one of the 20 intermediates could

stay for a long time on one of the false positions, it would effectively block the two positions it was straddling and hold up the polymerization process. Our code makes this impossible. This scheme, therefore, allows the intermediates to accumulate at the correct positions on the template without ever blocking the process by settling, except momentarily, in the wrong place. It is the feature which gives it an advantage over schemes in which the intermediates are compelled to combine with the template one after the other in the correct order.

The example given here is only for illustration, but it brings out the physical idea behind the concept of a comma-less code.

In passing, it should be mentioned that while the idea of making three nonoverlapping nucleotides code for one amino acid at first sight entails certain stereochemical difficulties, these are not insuperable if it is assumed that the polypeptide chain, when polymerized, does not remain attached to the template. A detailed scheme along these lines has been described to us by Dr. S. Brenner (personal communication)."

June 7, 1957

Second printing

~~HOW MAY AMINO ACIDS READ THE NUCLEOTIDE CODE?~~

Problems of Protein Synthesis

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 somewhat of a mystery in just what 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) A. L. Dounce, Enzymologia, 15, 251 (1952)
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 considerations 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 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 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:

(2) Note added June 17th. The May issue of the Proc.Nat.Acad.Sci., Vol. 43, p. 416 (1957) which was belatedly received here, contains an article by F.H.Crick, J.S.Griffith, and L.E.Orgel which is in part identical with their previously circulated memorandum referred to in the text. In addition, however, the authors discuss -- in conjunction with a detailed oral communication of S. Brenner to them -- how amino acids might read the nucleotide code. The relevant passage of their paper is quoted in full in Appendix A.

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 ~~in~~ 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, say 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.(2) 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~~ concentrate 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~~ ~~hydroxyl~~ 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 ~~possess~~ contain 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 ~~MMM~~ 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 (at the (2) carbon atom of the ribose) 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:~~

If we had a 20 ~~letter~~ word code and if 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, then ~~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.

Some of the amino acid sequences that may be found experimentally in sequential analysis of proteins and polypeptides might show that the restrictions imposed by our model on the possible amino acid sequence are too severe, and that the model has to be modified to accommodate established facts. As will be presently seen, some modification of our model may be indicated for other reasons also. There is certainly no inherent reason why we should have a pure three-letter code -- as we have assumed above -- and why, for instance, a certain number of four-letter code words should not be utilized also.(4)

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.

So far the number of amino acid residues in polypeptides and proteins have been determined only within one rather special class; all of them represent substances which are secreted by mammalian tissues. The number of amino acid residues found in such polypeptides and proteins, which have been analyzed with adequate accuracy, 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.

(4) See Appendix A

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 synthetising 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. 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 in bacteria -- 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.

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 model that we have postulated. For the purposes of this computation, we shall assume that the molecular weight of the enzyme is about 100,000. We then have about 1,000 amino acid residues in the enzyme, and accordingly we would have to assemble alongside the paragene m = 300 trinucleotides, each of which is "loaded" with three amino acids.

In the approximation which we shall use, the minimum time, τ_0, needed for the formation of the polypeptide is composed of two terms, τ_1 and τ_2;

$$\tau_0 = \tau_1 + \tau_2$$

After all the amino acids, assembled alongside the template, have been linked into one polypeptide, and if this polypeptide is at once removed, a certain time, τ_1, will elapse until the trinucleotides, which are now denuded of amino acids, evaporate from the template and their place is taken by trinucleotides which are loaded with the proper amino acids. We shall assume that the concentration of denuded trinucleotides in the cell is very small so that after the denuded trinucleotides evaporate, the loaded trinucleotides do not have to compete with denuded trinucleotides for their legitimate positions along the paragene. The time, τ_1, which is necessary to permit evaporation of m denuded trinucleotides and to assemble m loaded trinucleotides in their place we shall compute here on the assumption that once a loaded trinucleotide has found its position alongside the template, it will not evaporate again. Because this assumption is not valid, we must make a correction which is represented by the second term, τ_2.

Computation of τ_2

First we shall now compute this second term, τ_2. This computation is 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 have to deal with a reversible reaction which may be written as follows:

codeword + loadedtrinucleotide $\rightleftharpoons$ codeword—loadedtrinucleotide

The rate at which the reaction proceeds from left to right; i.e. the number of successful hits per unit time is given by

$$\text{hit rate} = A\rho$$

ρ denotes the concentration of the particular kind of loaded trinucleotide, in mol/cc. We shall, for the sake of simplicity, assume that the concentration in the cell of <u>each kind</u> of loaded trinucleotide is the same.

A stands for

$$A = 6 \cdot 10^{23} v \sigma p \qquad (1)$$

v is the molecular velocity, $v = \sqrt{\frac{2RT}{\pi M}}$ where M is the molecular weight of the loaded trinucleotide. If $M \approx 1000$, we have $v \approx 5 \times 10^3$ cm/sec.

σ is the target area that must be hit by the loaded trinucleotide if Hydrogen bonding is to take place with the three adjacent nucleotides on the parugene. We may assume for σ a value of $\sigma = 10^{-15}$ cm^2.

p 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 , as a very rough estimate $p = 1/300$; This would give for σp the value $\sigma p = 1/3\ 10^{-17}$.

The rate at which the reaction proceeds from left to right is given by the rate α at which the codeword—loadedtrinucleotide complex dissociates or, as we may also say, the rate α at which the loaded trinucleotide evaporates from the paragene. For this rate we may write

(2) α = rate of evaporation $\approx 10^{13} e^{-\frac{\Delta H}{RT}}$

where ΔH is the binding energy for the loaded trinucleotides.

In equilibrium the hit rate and evaporation rate must be equal. The equilibrium constant, K, of this reversible reaction is defined as the concentration of the loaded trinucleotides at which the code word is covered half of the time. Thus we may write

(2) $$A K \approx \frac{1}{2} 10^{13} e^{-\frac{\Delta H}{RT}} = \frac{1}{2}\alpha$$

or

$$\alpha = 2AK$$

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

(4) $$f = \frac{1}{1 + \frac{\rho}{K}}$$

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

(5) $$\text{"number of gaps"} = \frac{m}{1 + \frac{\rho}{K}}$$

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

(6) $$\frac{\rho}{K} \gg 1$$

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_1, 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 averuge time, $\frac{1}{A\rho}$, that it takes to fill one given gap. Thus, for τ_2 we may write

$$\tau_2 = \frac{1}{A\rho}\,\frac{m}{1+\frac{\rho}{K}} \tag{7}$$

Computation of τ_1 and $\tau_0 = \tau_1 + \tau_2$

When the polypeptide is formed and leaves the paragene, the code words are covered with the denuded trinucleotides. We may now compute the average time, τ_1, needed for the evaporation of all the denuded trinucleotides and the assembling of all the loaded trinucleotides in their place. We shall, for the sake of simplicity, assume that a denuded trinucleotide evaporates at the same rate, α, as a loaded trinucleotide. The rate, α, at which a loaded trinucleotide evaporates from the template is given by (3) and we may also write this in the form

$$\alpha = A\rho\,\frac{2K}{\rho} \tag{8}$$

Since we have assumed $\frac{\rho}{K} \gg 1$, we have

$$A\rho \gg \alpha \tag{9}$$

As may be shown,[5] for very large values of m, ($\ln m \gg 1$) we may write for τ_1

$$\tau_1 \approx \frac{1}{A\rho}\,\frac{\rho}{2K}\,\ln m \tag{10}$$

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

$$\tau_0 \approx \frac{1}{A\rho}\left\{\frac{m}{1+\frac{\rho}{K}} + \frac{1}{2}\,\frac{\rho}{K}\,\ln m\right\} \tag{11}$$

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

$$\frac{\rho}{K} \approx \sqrt{\frac{2m}{\ln m}} \tag{12}$$

$\sqrt{\frac{2000}{3\times 2.3}} \approx$ 17

(5) See Appendix B

Substituting this value into (11) gives

(13) $$\tau_0 \approx \frac{\sqrt{2}}{A\rho}\sqrt{m \ln m}$$

For a polypeptide containing 1,000 amino acid residues, i.e. a paragene containing about ~~300~~ 1000 code words, we may write m = ~~300~~ 1000, and thus we obtain from (12) and (13)

(14) $$\frac{\rho}{\kappa} \approx$$ ~~10~~ 17

and

(15) $$\tau_0 \approx \frac{50}{A\rho}$$ ~~12~~ $\sqrt{12000} \approx \frac{100}{A\rho}$

Estimates for the values of $\sigma\rho$, A, α and K

As we may see from (15) we obtain $\tau_0 = 1$ if $A\rho =$ ~~100~~ 1000 This means that for this particular value of $A\rho$ one enzyme molecule is produced per paragene per second. As we have seen before, this is the order of magnitude of the rate at which highly induced bacteria produce the enzyme β-galactosidase per paragene.

We shall, therefore, in the following assume $A\rho =$ ~~100~~ 1000 and compute from it ρ, the concentration at which the different kinds of trinucleotides may be present in the cell.

If we use for $\sigma\rho$ the value of $\sigma\rho = \frac{1}{3}\ 10^{-17}\ cm^2$, then we obtain from (1) A; $A = 10^{10}$ and accordingly we have $\rho =$ ~~$5 \cdot 10^{-9}$~~ 10^{-8} mol/cc ($\rho =$ ~~$5 \cdot 10^{-6}$~~ 10^{-5} mol/liter).

It might well be, however, that $\sigma\rho$ is ten times higher so that we have $\sigma\rho = \frac{1}{3}\ 10^{-16}\ cm^2$, and then we obtain from (1) for A; $A = 10^{11}$, so that we have $\rho =$ ~~$5 \cdot 10^{-10}$~~ 10^{-9} mol/cc or ($\rho =$ ~~$5 \cdot 10^{-7}$~~ 10^{-6} mol/liter).

Thus the concentration, ρ, of the different kinds of trinucleotides in the cell is likely to be between ~~5×10^{-6}~~ 10^{-5} and ~~5×10^{-7}~~ 10^{-6} mol/liter, and one would have to look for concentrations of this order of magnitude in order to obtain experimental confirmation of their presence.

Since we have assumed $\frac{\rho}{K} \approx 17$, and since we believe that we have $A\rho = 100$, it follows that we must have for the rate of evaporation of the trinucleotides from the paragene, α

(16) $\alpha = 2AK \sim 10/\text{sec.}$ or $1/\alpha \sim 1/10$ sec.

For values of A between 10^{10} and 10^{11}, we have

(17) $$5\,10^{-11} < K \text{ (in mol/cc)} < 5\,10^{-10}$$

or

$$5\,10^{-8} < K \text{ (in mol/liter)} < 5\,10^{-7}$$

From (5) we obtain for a given pair of values A and K the binding energy ΔH between the loaded trinucleotide and the code word on the paragene. For a value of $A = 10^{10}$ and $K = 10^{-10}$ mol/cc ($K = 10^{-7}$ mol/liter), we obtain ΔH 18,000 calories. Since 6 hydrogen bonds are involved, this would mean about 3,000 calories per hydrogen bond.

For the same value of A and a value of K which is ten times larger, ΔH would decrease by about 1400 calories.

Conclusion

These considerations show that the theory which we postulated should be able to explain the high rate of enzyme synthesis which one observes in bacteria when the rate of formation of an enzyme is maximally enhanced by the use of an inducer. The basic thought of this theory consists in the assumption that trinucleotides and ~~possibly also~~ tetranucleotides (and perhaps higher oligonucleotides) read the code of the paragene, and that these oligonucleotides carry amino acids. ~~One~~ A particular model for protein synthesis, ~~which~~ assumes that each trinucleotide or tetranucleotide carries a sequence of three or four amino acids respectively, was singled out for detailed discussion because this model appeared to be the most plausible. This does not mean, however, that other models can not be considered.

For instance, rather than to assume that each kind of trinucleotide carries a particular sequence of three amino acids, one might wish to explore the possibility that each trinucleotide ~~might carry~~ 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 carbon of the 5 carbon sugar of either the leading or the 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. See Crick et al for commaless code

However, the particular model for protein synthesis here considered cannot be modified by simply saying that each trinucleotide carries one amino acid instead of carrying three amino acids, for adjacent amino acids would not then be at the right distance from each other to be linked into a polypeptide. Such an alternate model for protein synthesis would, therefore, require additional ideas concerning the formation of the polypeptide.(2)

I am grateful for the discussion of a variety of problems arising out of the considerations here presented which I had at the University of Chicago with Prof. Herbert S. Anker, Dr. Nandor L. Balazs, Mr. Hirondo Kuki, Prof. Joseph E. Mayer, and Prof. Leonard J. Savage.

Leo Szilard

APPENDIX

(Added June 17/57)

A.) If, in addition to ribose trinucleotides, carrying three amino acids, we have in the role of anti-code words also ribose tetranucleotides, carrying four amino acids, then we must postulate that only three of the bases of the tetranucleotide pair with complementary bases on the paragene. The fourth base, if it is a purine, must pair with the wrong pyrimidine, and if it is a pyrimidine, it must pair with the wrong purine. The reason for this is as follows:

As formula (17) shows, α, the rate of evaporation of a trinucleotide from the paragene can be estimated to be about 10/sec., corresponding to a binding energy, ΔH, of about 18,000 calories or 3,000 calories per hydrogen bond. If all four bases of the tetranucleotide were paired with the complementary bases on the paragene, we would then have two more hydrogen bonds and presumably an additional binding energy of about 6,000 calories. The equilibrium constant, K, of the tetranucleotide would therefore be lower than the value computed for K of the trinucleotides by a factor of about 10^4, and the rate of evaporation, α, would be lower by the same factor. This would then make, τ_0, the minimal time it takes for a paragene to form a polypeptide too long to be compatible with the observed rate of production of β-galactosidase in highly induced bacteria.

The number of words that may be constructed in a mixed three-letter word and four-letter word code of the kind described above, if we demand that the code require no commas, has so far not been determined.

B.) The problem of computing τ_1 amounts to the solving of the following problem:

There are m boxes, each of which can hold one white ball or one black ball. Initially, at time, $\lambda = 0$, all of these boxes contain one

white ball -- a stripped trimucleotide. These white balls evaporate at the rate, α, so that at time λ the probability, $W(\lambda)$ of having no white ball in the box is given by

$$W(\lambda) = 1 - e^{-\alpha\lambda} \tag{18}$$

If the rate at which black balls fall into an empty box is designated by β, then the probability, y, that a given box does not contain a black ball at time, t, is given by

$$y(t) = e^{-\alpha t} + \int_{\lambda=0}^{\lambda=t} e^{-\beta(t-\lambda)} \frac{dW}{d\lambda} d\lambda \tag{19}$$

we may write

$$\int_{\lambda=0}^{\lambda=t} e^{-\beta(t-\lambda)} \frac{dW}{d\lambda} d\lambda = \frac{\alpha}{\beta-\alpha} \left[e^{-\alpha t} - e^{-\beta t} \right] \tag{20}$$

so that we have

$$y(t) = \frac{1}{\beta-\alpha} \left(\beta e^{-\alpha t} - \alpha e^{-\beta t} \right) \tag{21}$$

Thus we may write for the probability, x, that a given box contains a black ball

$$x(t) = 1 - y(t) = 1 - \frac{\beta}{\beta-\alpha} e^{-\alpha t} + \frac{\alpha}{\beta-\alpha} e^{-\beta t} \tag{22}$$

From this we obtain for P(t), the probability that all m boxes contain one black ball

$$P(t) = x(t)^m = \left(1 - \frac{\beta}{\beta-\alpha} e^{-\alpha t} + \frac{\alpha}{\beta-\alpha} e^{-\beta t}\right)^m \tag{23}$$

The average time, τ_1, needed for the evaporation of the white balls from all m boxes and the filling of all m boxes with black balls is given by

$$\tau_1 = \int_0^\infty t \frac{dP}{dt}\, dt \tag{24}$$

If $\beta \gg \alpha$, we may write

$$\tau_1 = \int_0^\infty t \frac{d}{dt}\left(1 - e^{-\alpha t}\right)^m dt \tag{25}$$

The expression $\frac{d}{dt}\left(1-e^{-\alpha t}\right)^m$ has a maximum at some value of t; $t = \tau_0$ and -- for large values of m -- it becomes small very rapidly both below and above $t = \tau_0$. Therefore, if m is large we may write

$$\tau_1 \approx \tau_0 \tag{26}$$

We obtain τ_0 by writing

$$\left\{ \frac{d^2}{dt^2}\left(1 - e^{-\alpha t}\right)^m \right\}_{t=\tau_0} = 0 \tag{27}$$

and from this we obtain

$$\tau_0 = \frac{1}{\alpha} \ln m \tag{28}$$

Thus for $\beta \gg \alpha$ and $m \gg 1$ we may write (26)

(29) $$\tau_1 \approx \frac{1}{\alpha} \ln m$$

It can be shown from (23) and (24) that in the next higher approximation we have

(30) $$\tau_1 \approx \frac{1}{\alpha} \{ \log(m+1) + C \} + \Delta \; ; \text{ where } \begin{cases} C = 0.577 \text{ (Euler's const.)} \\ \frac{1}{\beta} < \Delta < \frac{1}{\beta - \alpha} . \end{cases}$$

~~6.~~) F. H. Crick, J. S. Griffith and L. E. Orgel write in the May issue of the Proc. Nat. Acad. Sci. (Vol. 43, pp. 419 and 420, 1957):

"To fix ideas, we shall describe a simple model to illustrate the advantages of such a code. Imagine that a single chain of RNA, held in a regular configuration, is the template. Let the intermediates in protein synthesis be 20 distinct molecules, each consisting of a trinucleotide chemically attached to one amino acid. The bases of each trinucleotide are chosen according to the code given above. Let these intermediate molecules combine, by hydrogen bonding between bases, with the RNA template and there await polymerization. Now imagine that such an amino acid-trinucleotide were to diffuse into an incorrect place on the template, such that two of its bases were hydrogen-bonded, though not the third. We postulate that this incomplete attachment will only retain the intermediate for a very brief time (for example, less than 1 millisecond) before the latter breaks loose and diffuses elsewhere. However, when it eventually diffuses to the correct place, it will be held by hydrogen bonds to all three bases and will thus be retained, on the average, for a much longer time (say, seconds or minutes). Now the code we have described insures that this more lengthy attachment can occur only at the points where the intermediate is needed. If one of the 20 intermediates could

stay for a long time on one of the false positions, it would effectively block the two positions it was straddling and hold up the polymerization process. Our code makes this impossible. This scheme, therefore, allows the intermediates to accumulate at the correct positions on the template without ever blocking the process by settling, except momentarily, in the wrong place. It is the feature which gives it an advantage over schemes in which the intermediates are compelled to combine with the template one after the other in the correct order.

The example given here is only for illustration, but it brings out the physical idea behind the concept of a comma-less code.

In passing, it should be mentioned that while the idea of making three nonoverlapping nucleotides code for one amino acid at first sight entails certain stereochemical difficulties, these are not insuperable if it is assumed that the polypeptide chain, when polymerized, does not remain attached to the template. A detailed scheme along these lines has been described to us by Dr. S. Brenner (personal communication)."