ON DEGENERATE TEMPLATES AND THE ADAPTOR HYPOTHESIS

F.H.C. Crick,

Medical Research Council Unit for the Study of
the Molecular Structure of Biological Systems,

Cavendish Laboratory, Cambridge, England.

A Note for the RNA Tie Club.

"Is there anyone so utterly lost as he
that seeks a way where there is no way."

Kai Kā'ūs ibn Iskandar.

early 1955

In this note I propose to put on to paper some of
the ideas which have been under discussion for the last year or
so, if only to subject them to the silent scrutiny of cold print.
It is convenient to start with some criticisms of Gamow's
paper (Dan.Biol.Medd.22, No.3 (1954)) as they lead naturally
to the further points I wish to make.

Some straightforward criticisms first. The list
of amino acids in Table I of the paper clearly needs reconsider-
ation, and this brings us to the very interesting question as to
which amino acids should be on the list, and which should be
regarded as local exceptions. We first remove norvaline which
we now know has never been found in proteins. Nor, as far as
I know, is there at present any evidence for hydroxy glutamic
and cannine. On the other hand asparagine and glutamine
certainly occur, and indeed are probably quite common. We now
come to the "local exceptions". These are:

$\left\{\begin{array}{l}\text{hydroxyproline}\\\text{hydroxylysine}\end{array}\right.$

$\left\{\begin{array}{l}\text{tryosine derivatives, i.e. diiodotyrosine,}\\\text{thryoxine, etc.}\qquad\qquad\text{dibromotyrosine}\end{array}\right.$

diaminopimelic
phosphoserine.

The first two occur only in gelatin. The tyrosine derivatives
are found only in the thyroid (the iodo ones) and in certain
corals (and in other marine organisms?). Diaminopimelic
occurs only in certain algae and bacteria and has not yet been
shown unambiguously to occur in an ordinary protein.
Phosphorous occurs in casein, ovalbumin and pepin, and may be
present as phosphoserine.

There are, in addition, amino acids which occur in
small peptides, such as ornithine, diaminobutiric, etc. - see
Table I of Bricas and Fromageot, Ad.Prot.Chem.(1953) Vol.VIII
for a comprehensive list. Under this heading one should also
include the D isomers of common amino acids, and ethanolamine,
which occurs in gramicidin.

In my view all these special cases can be disregarded for the moment, and moreover proteins in which they occur should not be considered "genuine" proteins without further justification. This applies particularly to collegen, which may turn out to be more a "polymer" than a protein - and I would also discard silk for the same reason. Practically all the small peptides (e.g. the antibiotics) should be ignored, and I myself would be cautious about ocytocin and vasopreosin. I suspect that the tyrosine derivatives and the phosphorous derivatives should be regarded more as modifications to a protein, in the same way as we regard the addition of a prosthetic group. The case of diaminopimelic is more difficult - further evidence is clearly required here. It would be valuable if one of the more biochemical members of the club could write a paper discussing all these points in more detail than I have done here.

There remains the cystine-cysteine problem. It is not unreasonable to discard cystine, and assume that S-S bridges are formed later. I doubt if we have any evidence one way or the other. Thus modified (i.e. with asparagine and glutamine replacing cysteic acid and hydroxyproline) the list comes to 20, as given in the Club tie-pin list.

Application of Gamow's Scheme.

It is well-known that Gamow's scheme does not work for insulin, though the argument given in his paper is not valid because one of the glutamic residues is actually glutamine. I showed some time ago that the B chain could not be coded, but the proof is long and intricate and not worth reproducing. I believe other people have also shown this.

If the insulin data is combined with that for β-corticotropin a very neat proof is possible, as follows. One can list all possible amino acid combinations, using Gamow's code, having the form xyx. It is found that there are ten of these, and that no two of them have the same <u>middle</u>

amino acid.   Now Insulin B has Leu. Tyr. Leu., and β-cortico-
tropin has Ser. Tyr. Ser.   These cannot both be coded by
Gamow's scheme.

Another proof of this type depends on the A chain of
two species of insulin.*   The sequences are identical except
that one (sheep) has Gly. where the other (bovine) has Ser.
The change occurs roughly in the middle of the chain. Both
sequences cannot be coded by a Gamow scheme, since changing
one pair of bases necessarily alters at least two amino acids,
and this cannot be corrected without making further changes in
the base sequence.   The only way to do this efficiently is
to have a sequence of the type

3   4   4   3

3   4   which changes   4   3
                over to

3   4   4   3
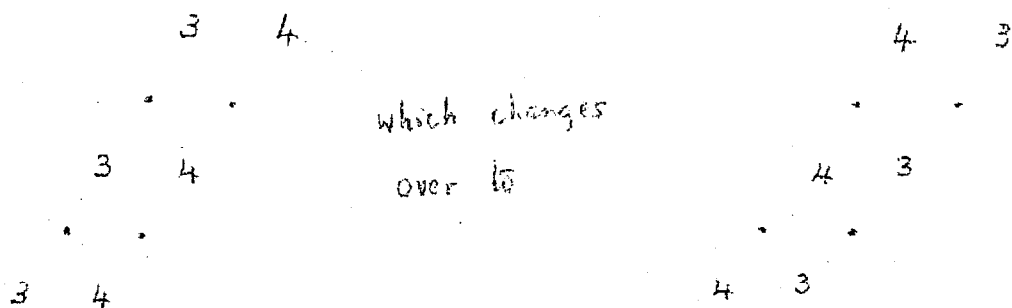
and since there are only two distinct diamonds with (3,4) top
and bottom (r and t), one cannot code more than 5 residues from
the changed amino acid.   Thus to code both species of insulin
A chains is impossible.   A third method to disprove Gamow's
scheme, given sufficient data, is to count neighbours.   This
is particularly useful in a scheme which does not distinguish
between neighbours-on-the-right and neighbours-on-the-left.

Using the data from the two insulin chains and
β-corticotropin one finds 10 amino acids having 8 neighbours
or more.   Gamow's scheme (see his Table III) allows only 8
amino acids to have more than 7 neighbours.   Thus coding
would be impossible.

* F. Sanger. Personal communication and in the press.

-3-

I have used the same method for testing Gamow's scheme assuming it applied to <u>alternate</u> amino acids, i.e. the odd positions form one sequence, and the even ones another. This time the proof is more complicated, since in the above data only 7 amino acids have 8 neighbours or more. However it is easily shown that the association rules of Gamow's Table III cannot be obeyed; as follows

aeio associates with

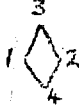(aeio + dghn) + imp + fuv, while dghn associates with
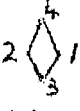
(aeio + dghn) + kst + bcr.

Thus apart from the (aeio + dghn) group, which we have identified (with one exception),the other <u>neighbours</u> of the (aeio + dghn) group should fall into two mutually exclusive classes. This is easily shown not to be the case. Thus Gamow's scheme cannot work.

I have set out these at length, not to flog a dead horse, but to illustrate some of the simplest ways of testing a code. It is surprising how quickly, with a little thought, a scheme can be rejected. It is better to use one's head for a few minutes than a computing machine for a few days!

<u>Gamow's Scheme: Fundamental Objections.</u>

The most fundamental objection to Gamow's scheme is that it does not distinguish between the <u>direction</u> of a sequence; that is, between Thr. Pro. Lys. Ala. and Ala. Lys. Pro. Thr. using the usual convention. There is little doubt that Nature makes this distinction, though it might be claimed that she produces both sequences at random, and that the "wrong" ones - not being able to fold up - are destroyed. This seems to me unlikely.

This difficulty brings us face-to-face with one of the most puzzling features of the DNA structure - the fact that it is non-polar, due to the dyads at the side; or put another way, that one chain runs up while the other runs down. It is true that this only applies to the backbone, and not to the base

sequence, as Delbrück has emphasised to me in correspondence. This may imply that a base sequence read one way makes sense, and read the other way makes nonsense. Another difficulty is that the assumptions made about which diamonds are equivalent are not very plausible. It is not perhaps inplausible that

$\begin{smallmatrix}&3\\1&\diamondsuit&2\\&4\end{smallmatrix}$ should be the same as $\begin{smallmatrix}&4\\2&\diamondsuit&1\\&3\end{smallmatrix}$, (though this

assumption has structural implications), but it has also been

assumed that these are the same in their effect as $\begin{smallmatrix}&4\\1&\diamondsuit&2\\&3\end{smallmatrix}$

and $\begin{smallmatrix}&3\\2&\diamondsuit&1\\&4\end{smallmatrix}$. This would be not unreasonable

if the amino acid could fit on to the template from either side, into cavities which were in a plane, but the structure certainly doesn't look like that. The bonds seem mainly to stick out perpendicular to the axis, and the template is really a surface with knobs on, and presents a radically different aspect on its two sides.

Gamow's argument about the bilateral symmetry of the majority of the amino acids is the wrong way round. Such amino acids would more reasonably be associated with cavities which have this symmetry already - that is, the ones in his list which are not marked with an asterisk.

The Gamow approach.

What, then, are the novel and useful features of Gamow's ideas? It is obviously not the idea of amino acids fitting on to nucleic acids, nor the idea of the bases sequence of the nucleic acids carrying the information. To my mind Gamow has introduced three ideas of importance:

(1) In Gamow's scheme several different base sequences
can code for one amino acid (as just discussed).
This "degeneracy" seems to be a new idea, and, as
discussed later, we can generalise it.

(2) Gamow boldly assumed that code would be of the
overlapping type. That is, if we denote the sequence
of base pairs by 1 2 3 4 5 6 ...., he assumed that
the first amino acid was coded by 1 2 3, and the next by

2 3 4, not by 4 5 6.   Watson and I, thinking mainly
about coding by hypothetical RNA structures rather than
by DNA, did not seriously consider this type of coding.

(3) Gamow's scheme is essentially abstract.   It originally
paid lip-service to structural considerations, but the
position was soon reached when "coding" was looked upon
as a problem in itself, independent as far as possible
of how things might fit together.   As I shall explain
later, such an approach, though at first sight unnecess-
arily abstract, is important.

Finally it is obvious to all of us that without our
President the whole problem would have been neglected and few
of us would have tried to do anything about it.

Structural Considerations.

I want to consider two aspects of the DNA structure.
Firstly its dimensions;   secondly its chemical character.

The dimensional side is soon disposed of.   In the
"paracrystalline" form of DNA (Structure B) we have one base
pair every 3.4 Å in the fibre direction.   A fully extended
polypeptide chain measured about 3.7 Å from one amino acid to
the next.   Therefore it is argued that not more than one base
pair can, on the average, be matched with an amino acid.   If
we go up the outside of the helix the position is worse, since
the distance per base-pair is now greater, perhaps twice as great.

I want to point out that this argument, though powerful,
is not completely water-tight.   To begin with, in crystalline
DNA (Structure A) the distance between base pairs along the film
axis is less than 3.4 Å, being probably about 2.5 Å.   Now
"in solution" one might expect Structure B to prevail, but such
DNA might easily go over to Structure A when amino acids condensed
on it.   Moreover, for all we know, the process of tilting the
bases may perhaps go even further, and there may be a third, semi-
stable, configuration with a base-pair distance even shorter
than 2. 5 Å.

Then, again, we have no evidence to tell us whether the completed part of the polypeptide chain stays on the template. It is just possible that the distance between the growing end of the chain and the next (free) amino acid at the operative moment may be greater than 3.7 Å, though I doubt if it could be much greater. Thus it seems to me just possible, though not very probable, that one amino acid might stretch over two base pairs rather than one. (Notice that this argument is weakened if the polypeptide backbone is put at a distance from the fibre axis, even if the _inside_ of the nucleic acid structure is used for coding, rather than the outside.) It seems highly unlikely on the present DNA structure that one could have three base-pairs per amino acid (RNA may be different of course).

As regards chemical character, I want to consider not only the DNA structure, but also any conceivable form of RNA structure. Now what I find profoundly disturbing is that I cannot conceive of _any_ structure ([for either nucleic acid] RNA or DNA) acting as a direct template for amino acids, or at least as a specific template. In other words, if one considers the physical-chemical nature of the amino acid side chains we do not find complimentary features on the nucleic acid. Where are the knobly hydrophobic surface to distinguish valine from leucine and isoleucine? Where are the charged groups, in specific positions, to go with the acidic and basic amino acids? It is true that a "Teller" scheme, in which the amino acids already condensed act effectively as part of the template, might be a little easier, but a study of sequences from this point of view is not encouraging.

I don't think that anybody looking at DNA or RNA would think of them as templates for amino acids were it not for other, indirect evidence.

What the DNA structure _does_ show (and probably RNA will do the same) is a specific pattern of _hydrogen bonds_, and very little else. It seems to me, therefore, that we should widen our thinking to embrace this obvious fact. Two schemes suggest

-7-

themselves.——In the first small molecules (phospholipides? ions chelated on guanine?) could condense on the nucleic acid and pad it suitably, and the resulting combination would form the template. I shall not discuss this further here. In the second, each amino acid would combine chemically, at a special enzyme, with a small molecule which, having a specific hydrogen-bonding surface, would combine specifically with the nucleic acid template. This combination would also supply the energy necessary for polymerisation. In its simplest form there would be 20 different kinds of adaptor molecule, one for each amino acid, and 20 different enzymes to join the amino acid to their adaptors. Sydney Brenner, with whom I have discussed this idea, calls this the "adaptor hypothesis", since each amino acid is fitted with an adaptor to go on to the template.

The usual argument presented against this latter scheme is that no such small molecules have been found, but this objection cannot stand. For suppose, as is probable, that the small adaptor molecules are in short supply. Then consider the experiment in which all amino acids except one, (say leucine) is supplied to an organism, so that protein synthesis stops. Why do not the intermediaries - the (amino acid + adaptor) molecules - accumulate? Simply because there is very little of them, and no more amino acids can combine with these adaptors until the amino acids, to which they are at that moment attached, have been made into proteins, thus releasing the adaptor molecule. Thus under these conditions free amino acids accumulate, not amino acids-plus-adaptor molecules.

(In passing, it would be interesting to do this experiment with rare amino acids, like tryptophane and isoleucine, to see if proteins without them continued to be synthesised. Perhaps someone has a suitable mutant.)

In any case it seems unlikely that totally free amino acids actually go on to the template, because a free energy supply is necessary, especially when one bears in mind the entropy contribution needed to assemble the amino acids in the correct

-8-

order.   Free energy must be supplied to prevent mistakes in
sequence being made too frequently.

The adaptor hypothesis implies that the actual set of
twenty amino acids found in proteins is due either to a historical
accident or to biological selection at an extremely primitive stage.
This is not impossible, since once the twenty had been fixed it
would be very difficult to make a change without altering every
protein in the organism, a change which would almost certainly
be lethal.   It is perhaps surprising that an occasional virus
has not done this, but even there a number of steps would be
required.   Incidentally the adaptor mechanism may make it easier
to explain some of the local exceptions to the "magic 20"
rule-diaminopimelic should be watched from this point of view,
also thyroxine.

It is also conceivable that there is more than one
adaptor molecule for one amino acid, and the number 20 may be
simply an accident (in any case we need a code for "end chain",
so perhaps 21 would be more reasonable).   Alternatively the same
adaptor molecule might fit on in more than one way (related, say,
by a rotation of θ°.)

Degenerate Templates.

Such a point of view discourages a purely structural
approach to the problem, at least for the moment, and throws us
back on "coding", which, it is important to note, still remains
a problem even with this new approach.   However, we now have
even fewer structural limitations than before, since we can think
of other types of degeneracy, rather than the Gamow type.

To make this clearer, let us consider the Gamow code.
Let us denote the four possible base pairs by A B C D, reserving
the small letters, a, b, c, ... for the amino acids.   Then
in Gamow's code an amino acid is represented by several separate
sequences of three letters.   For example if

1 2	=	A
2 1	=	B
3 4	=	C
4 3	=	D

where 1, 2, 3, 4 are the four bases,

then Gamow's a, which in his notation is

would be

$$
\begin{array}{cccc}
A & A & D & D \\
A\ , & B\ , & A \quad \text{or} & B \\
C & C & B & B
\end{array}
$$
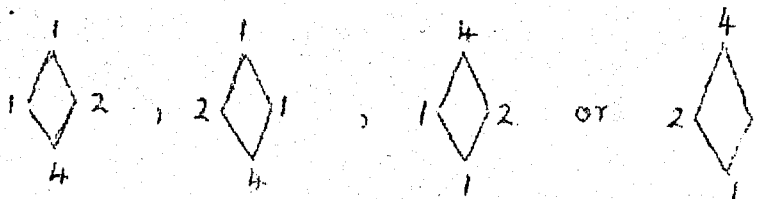
or more conveniently written

CAA, CBA, BAD or BBD

In his code 12 of his amino acids have 4 possible representations and the remaining 8 have 2 representations, making a total of 64 representations in all, this being the number of possible permutations of four types of things taken three at a time.

We can generalise this as follows. We can try to construct a code with the following properties:

(1) Four types of letters: A, B, C, and D.

(2) Each sequence of three consecutive letters has a meaning

(3) Overlapping i.e. DABDC......

means   DAB

then   ABD

then   BDC   etc.

(4) A particular amino acid is represented by one or more sets of three letters, <u>chosen at will</u>.

To illustrate, consider an unlikely code:

The combination code

There are 20 different combinations of four types of thing chosen three at a time (Note that Gamow's 20 comes from twice-ten, where ten is the number of combinations of four types of thing taken two at a time).

Thus one amino acid, say a, would be represented by the permutations: ABC, ACB, BAC, BCA, CAB and CBA.

Another, say b, by BBD, BDB, DBB, and a third, say v, by CCC only.

This code seems structurally unlikely, but it does give the magic number 20, and it does make some letters (amino acids) rather frequent and some rather rare. Note that, like Gamow's code, it has no directional properties.

We can test this very rapidly. It is easy to show that no amino acid could have more than 10 neighbours. The data for insulin and β-corticotropic shows Val. to have 11. Moreover, of its neighbours, not more than three can have more than 7. neighbours, whereas the data show that Glu, Phe, Leu, Ser, Cys, and Pro (all neighbours of Val) have 8 or more. This acts as a double check. Thus the code is impossible.

The Easy-Neighbour Code.

I next tried to see if I could construct a code of this type for which <u>all</u> neighbours and next neighbours were possible. To make things a little simpler to start with, I assumed only 16 amino acids, intending to expand the list later. To my surprise, I found I could do this. I found 6 different and apparently independent solutions (I have not checked this last statement carefully). One of these was

AAA	AAB	AAC	AAD
BAB	BAC	BAD	BAA
CAD	CAA	CAB	CAC
DAC	DAD	DAA	DAB
ABA	ABB	ABC	ABD
BBB	BBC	BBD	BBA
CBD	CBA	CBB	CBC
DBC	DBD	DBA	DBB
ACA	ACB	ACC	ACD
BCB	BCC	BCD	BCA
CCD	CCA	CCB	CCC
DCC	DCD	DCA	DCB
ADA	ADB	ADC	ADD
BDB	BDC	BDD	BDA
CDD	CDA	CDB	CDC
DDC	DDD	DDA	DDB

Each set of four permutations corresponds to an amino acid. It is easy to see that any amino acid (of the 16) can neighbour any other, or near-neighbour any other. Moreover the restrictions

on xyx sequences are not severe, and four types of xxx are

possible.   Thus at first sight it seemed promising.   I was

therefore annoyed to find that it is impossible to code the two

species of insulin A chains with it, as it is impossible to code

two sequences identical except for one amino acid near the middle

of the sequences.   The same applies to my other solutions.
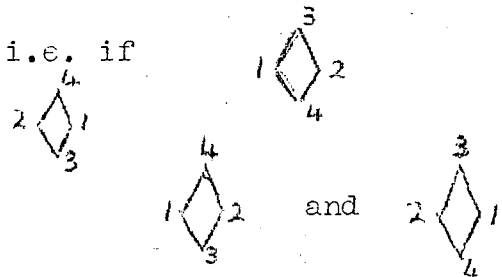
Directional Codes

The above codes are not directional;   that is, no

sequence of letters makes nonsense.   Is it possible to construct

a code which, when read backwards, makes nonsense almost everywhere?

This, again is not very difficult.   Leaving aside symmetrical

sets like AAA, or BAB, one must simply decide for each unsymmetrical

pair (e.g. DBA and ABD), which will mean something and which will

make nonsense.   There are 12 such pairs made of sets having

no two letters the same.   These one can allocate systematically

if one wishes (using a tetrahedron with the four letters at the

vertices).   There are 12 more pairs having two letters of each

set the same.   There seems to be no systematic way of allocating

these between sense and nonsense, so one can do it arbitrarily.

The remaining 16 permutations are symmetrical and we arbitrarily

assume that they represent sense.   Thus one gets 24 permutations

making nonsense, and 40 making sense.   This suggests that one

should systematically degenerate the 40 permutations to 20 pairs

but it is not obvious how to do this.   If it is done (so that

each amino acid is represented by exactly two permutations) then

at the most, on one side, only eight neighbours are possible, and

I am sure that sufficient good data exists to show that more than

eight neighbours, on one side, do occur (following Serine, for

example).   However, it is possible that a logical method of

degenerating exists which would give more than two representations

to some amino acids and less than two (i.e. one) to others.

The latter could only have, at the most, four neighbours on each

side.   I would be interested to know what the known, reliable,

neighbours are for say, Met, Try, Ileu, and Asp (not AspN).

At the moment this scheme looks unpromising and I have not

examined it further.

-12-

Logical Degeneracy.

Although I have argued that there may be no simple relationship between the different triplets of base-pairs representing one amino acid, it is obviously sensible to investigate forms of degeneracy which derive from simple structural ideas, as Gamow's did. To illustrate this, consider a simple example.

Imagine a code based on diamonds like Gamow's, and allow rotational degeneracy, i.e. if [diamond 1,3,2,4] then associate with it [diamond 2,4,1,3] but not the other pair, [diamond 1,4,2,3] and [diamond 2,3,1,4] allowed by Gamow.

This gives too many possibilities. Now argue as follows: Suppose that we consider the NH_2 of adenine as different in its effect from the NH_2 of cytosine, but the $C = O$ of thymine as indistinguishable from that of guanine as far as the top and bottom of the diamonds are concerned. Let us put

$$
\begin{aligned}
\text{Guanine} &= 1 \\
\text{Cytosine} &= 2 \\
\text{Thymine} &= 3 \\
\text{Adenine} &= 4
\end{aligned}
$$

Then, for example, we shall have one amino acid represented by the following diamonds:

[diamonds: 1,3,2,4 ; 1,1,2,4 ; 2,4,1,3 ; 2,4,1,1]
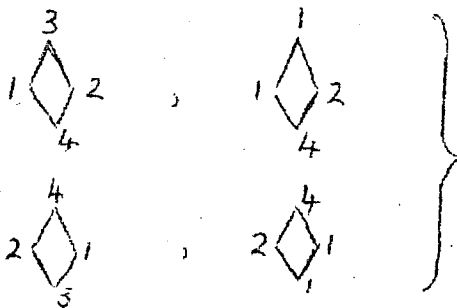
That is, if we have 3 in the top or bottom position, we can also have 1 (and vice versa).

It will be found that there are 18 such sets. Two of them contain eight representations each, eight contain four each, and the remaining eight contain two representations each. This does not quite get us to 20, but one might manage this by relaxing

the degeneracy a little.   This code suffers from the usual defect
of being non-directional, but here again it might be saved by
deleting certain representations;   an end-of-chain mark might be
provided in a similar way.

The "neighbour rules" are not excessively restrictive,
but the code fails to code the two species of insulin A chain.
One cannot code with it two sequences differing, near their
middle, by one amino acid only.

The General Case.

The problem which I have failed to solve is "are all
schemes of this type impossible?"   One test, which can be
applied eventually, is that there cannot be more than 256
different amino-acid pairs (out of a possible 400), since any
sequence of four base-pairs implies a definite pair of amino
acids (though the converse is not true).   My own impression is
that the large number of pairs (i.e. neighbours) now recorded,
and the difficulty of coding the three species of insulin A ,
together with the directional difficulty, make a solution unlikely,
but perhaps someone can produce a proper proof.   It is obviously
not easy since such a large class of codes is involved.

Further Structural Remarks.

If we accept the idea that what matters in DNA are the
hydrogen-bonding sites, it seems plausible to assume that each
"site" will combine with one adaptor and one adaptor only.   That
is, the spare H of the NH_2 on adenine will not combine first with
one adaptor and then another.   This requirement is not essential
but it is likely if adjacent adaptors have to be combined with the
DNA at the same time for polymerisation to occur.   If we restrict
ourselves to the NH and C = O groups this makes anything like
Gamow's scheme unlikely.   It suggests rather schemes of the type
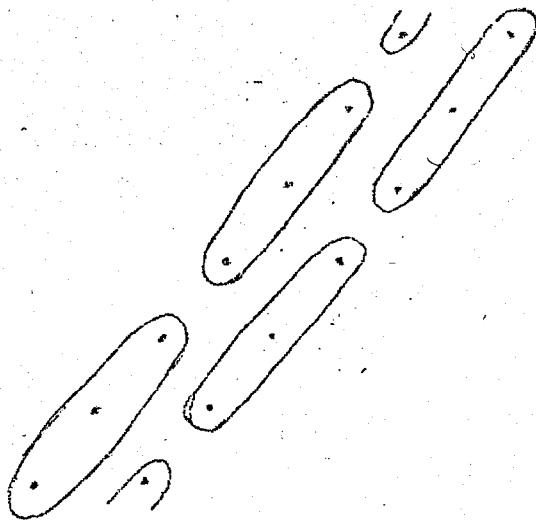
where each dot represents a C = O or NH site on a base, and the bubbles show which sets code for one amino acid. This scheme implies two amino acids every three base pairs, which, as we have seen, is not absolutely impossible on dimensional grounds. I shall not discuss such codes in detail. Obvious modifications and complications suggest themselves, and I may look into it further in the near future. Note that a maximum of 256 amino-acid pairs are possible, where pairs are not all adjacent amino-acids in a sequence, but are split up; i.e. for insulin B, either

Phe. Val.———AspN. GluN.———— His. Leu.—— Cys. Ser.—— etc.

or

-Phe.———— Val.AspN ————— GluN.His.———— Leu. Cys.—— etc.

It is as well to be aware of this sort of possibility while examining the sequence data. Incidentally such a scheme has one minor point to commend it. A fully extended polypeptide chain does not truly repeat after 3.7 $\overset{o}{A}$, but after twice this, the symmetry operation being a screw diad. An association-in-pairs is thus not totally silly.

Our assumption (that a site is only bonded once) does not compel us to a scheme of the above sort, because of the nitrogen in the 7 position of the two purines which could accept a hydrogen bond. This suggests schemes like the following. Represent NH by +, C = O by -, a purine N by X and the corresponding pyrimidine position by O. Thus a schematic view of the sequence

```
    guanine  -  cytosine
   cytosine - guanine
   adenine -  thymine
  thymine - adenine
```
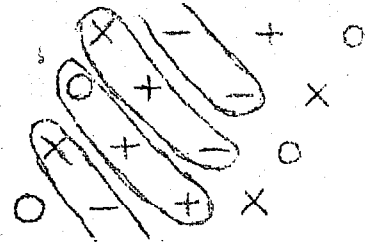
the bubbles representing the groups that decide which amino acids
go in.   (The + and - group will be in slightly different positions
depending upon which base-pair they belong to).

Such a scheme is a special type of our wide class
considered earlier, and since it has not led anywhere I shall
not discuss it further.

General Remarks

The main purpose of this note is to put forward the
adaptor hypothesis for serious consideration and to point out
its implications for degenerate templates.   It can of course be
considered in a wider content.   I have not considered "Teller"
schemes here - by which I mean codes which depend on the previous
amino acid - but the adaptor hypothesis removes even the flimsy
structural justifications put forward for the particular Teller
scheme suggested (and shown to us by Gamow at Woods Hole).
The basic difficulty of Teller schemes is that they are potentially
of enormous variety, and one simply doesn't know how to get down
to them till more sequence data has accumulated.   The fact that
the particular scheme put forward looked implausible should not
mislead anyone into thinking that all schemes of the Teller type
are unlikely.

Leaving aside Teller schemes, the adaptor hypothesis
allows other general types;   for example, depending on a sequence
of four base pairs.   The insulin A chain data make this unlikely,
but it is difficult to disprove rigorously.

I have tacitly dealt with DNA throughout, but the
arguments would carry over to some types of RNA structure.
If it turns out that DNA, in the double-helix form, does not act
directly as a template for protein synthesis, but that RNA does,
many more families of codes are of course possible.   [Incidentally

the protein sequences we use to test our theories - insulin, for example - are probably RNA-made proteins. Perhaps a special class of DNA -made proteins exists, almost always in small quantities (and thus normally overlooked), except perhaps where there are giant chromosomes. In particular base pairing may be absent in RNA or take a radically different form, and there may be more than one base to the asymmetric unit. Without a structure for RNA one can only guess.

Altogether the position is rather discouraging. Whereas on the one hand the adaptor hypothesis allows one to construct, in theory, codes of bewildering variety, which are very difficult to reject in bulk, the actual sequence data, on the other hand, gives us hardly any hint of regularity, or connectedness, and suggests that all, or almost all sequences may be allowed. In the comparative isolation of Cambridge I must confess that there are times when I have no stomach for decoding.