ENGINEERING AND SCIENCE
April 1962, Volume XXV, No. 7

GENETIC CODING

A mathematician considers the problem of how genetic information is encoded for transmission from parent to offspring.

by Solomon W. Golomb

How is genetic information encoded for transmission from parent to offspring? It has been known for many years that the vast amount of information required to specify a complete organism is somehow embodied in the chromosomes of each of the cells of that organism. The occurrence of such probabilities as ½, ¼, and ⅛ in Mendelian genetics already suggests an underlying *discrete* genetic mechanism. However, it has only been during the past dozen years that significant progress has been made toward explaining how this information is stored. Specifically, it has been demonstrated that it is not the protein matter in the chromosomes, but rather the nucleic acid, a different type of compound, which bears the genetic information.

The nucleic acid in the chromosomes is a type called *desoxyribonucleic acid,* or *DNA* for short. The DNA occurs in long strands which are in fact known to be paired helixes. Each strand may be regarded conceptually as a long segment of punched tape, in which four types of notches are punched to constitute a message. *Chemically, these "notches" are four distinct* side-groups, called *nucleotides,* which are attached linearly to the DNA stalk, at regular intervals. Thus the DNA strand is a message written in a four-symbol alphabet, where typically there are several thousand symbols per strand of DNA, and several thousand strands of DNA in the various chromosomes which make up the complete genetic blueprint of the organism.

Mathematically, the four symbols in the genetic code may be designated A, C, G, and T, the initials of *adenine, cytosine, guanine,* and *thymine,* the names of the *bases* corresponding to the four nucleotides. In some organisms the only nucleic acid used is single-stranded RNA (ribonucleic acid), but in most cases, double-stranded DNA contains the genetic information. In the paired strands of DNA, one member of the pair is clearly redundant, since A in one strand is always opposite T in the other, while C in one strand is always opposite G in the other, and conversely. It is believed that each strand serves as a template for the manufacture of the other, and that every time the cell divides, the paired DNA strands all separate, and replicate by the simple expedient of attracting the "complementary" nucleotides needed for the second strand.

To recapitulate, genetic information is stored on an organic tape called DNA, with the data inscribed using the four-symbol alphabet of A, C, G, T. By pairing a "positive" with a "negative" copy of the tape, when cell division occurs, the positive makes a new negative, and the negative a new positive, thus allowing replication to continue indefinitely.

Many years ago the mathematician John von Neumann described the design of a computer-like machine which would be capable of making duplicates of itself. It contained a punched tape with full instructions for building just such a machine, and the final instruction was to duplicate the tape. Von Neumann didn't know it at the time, but nature was already *using precisely this technique.*

Knowing that there is a coded message, the next question is: What is the content of the information which has thus been encoded; or, in operational terms, how is the tape "read," and what is built on the basis of the blueprint? By and large, the activities in which a cell engages consist mostly of the manufacture of proteins out of the basic sub-protein building blocks known as *amino acids.* There are 20 or more distinct amino acids which may be used for this purpose. It has been widely conjectured that there is a direct interpretation, or decoding, whereby several consecutive nucleotides of the DNA uniquely specify the occurrence, when decoded, of one particular amino acid.

(Actually, the nucleic acid RNA plays an important intermediate role. It appears that the true sequence of events is that the DNA message is first replicated onto "template RNA," a sort of temporary storage, and then shorter strands of "soluble RNA" perform the task of locating the amino acids, and of aligning them along the RNA template. These matters will be brushed aside rather high-handedly as "mechanical details" during the remainder of this discussion, with the RNA alphabet of A, C, G, U being treated as equivalent to the DNA alphabet of A, C, G, T.)

A basic tenet of "orthodoxy" is that the code used is the same for all terrestrial organisms. This tenet is supported by the fact that all known earthly creatures confine themselves to the same 20-odd amino acids as basic building blocks for protein, whereas if the code were evolving along with the life forms using it, other amino acids which are just as simple chemically as many of those used would be expected to come into the picture.

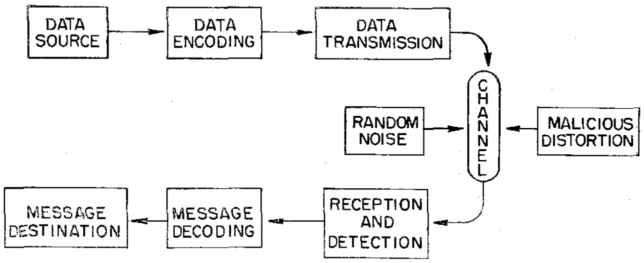
The viewpoint of communication theory

Any discussion of communication theory begins with a diagram of a general "information system" (below).

In the case of genetic information, the "data source" may be regarded as any conceptually complete description of the organism in question. The "data encoding" is the representation of this description in the form of DNA tape. The "data transmission," the "channel," and the "reception and detection" then refer to the sequence of steps whereby the DNA message is transferred to template RNA, and thence to soluble RNA, up to the point where the correspondence between nucleotides and amino acids is established. The "message decoding" is the formation of the protein molecules specified by the original DNA sequence, and the "message destination" is the new organism which results. If desired, one may "close the loop" by considering successive generations of this process, and in so doing, a "natural selection filter" could be inserted to incorporate Darwinian evolution into the model.

Although the diagram below shows "random noise" entering only into the "channel," as in radio communication systems, it is more realistic to recognize that errors can occur at every stage of reading information out of one box and into the next, and even when information is sitting quietly in storage. The "malicious distortion" (or "jamming") box can be interpreted as virus activity, where the intent of the organism to replicate itself is subverted by some operation on the coded sequence, into making many replicas of the virus instead.

General Diagram of an Information System



By examining this diagram of an information system, one may hope to determine both necessary and optional features of the genetic code, which could then be tested experimentally. A partial list of "desirable coding features" would include:

1. Efficient Use of the Channel
2. Self-Synchronization
3. Error Correction
4. Anti-Jamming
5. Convenient Decodability

It must be possible to recover the encoded information completely, unambiguously, and by means of the available chemical machinery. Another essential feature is *synchronization.* There must be some reliable method of deciding when the code group for one amino acid has ended and the next begun. On the other hand, *efficient use of the channel* is an optional feature. The information theory methods of Claude Shannon, Robert Fano, and David Huffman of MIT would suggest short code words for the frequent amino acids and longer code words for the infrequent ones. However, the decision as to relative frequencies would have been made so early in evolutionary history that it might bear little or no relation to the amino acid frequencies in organisms currently available. Moreover, the problems of data handling created by non-uniform word length could easily outweigh the advantages of a shorter average message length. The possibility that the genetic code incorporates features to combat random mistakes *(error correction)* or to combat malicious distortion *(anti-jamming)* is certainly mathematically intriguing, but far from being an essential requirement, a priori, of the genetic code.

If there is a valid function for the mathematician in such a field as this, and I firmly believe there is, it is to propose mathematical models for the biological situation, to deduce consequences and properties of these models, and to submit the consequences and properties for verification or refutation by the experimentalist. Then, by retaining those properties verified, and discarding those refuted, more and more precise models can be constructed, until the process finally converges to a mathematical model which is a completely faithful replica of the experimental situation. As a matter of fact, the history of the genetic coding problem actually exhibits such an interplay between models proposed and crucial experiments to reinforce or discard them.

Historical background

It was the cosmologist George Gamow who first proposed that it might be reasonable to formulate hypotheses concerning the nature of the DNA-amino acid coding based primarily on mathematical considerations of what the code is expected to do (i.e. its function), even in the absence of extensive experimental data about the physical *structure* of the nucleotide sequences. The first assumption Gamow introduced, back in 1954, was "uniform block length," viz. that the same number of nucleotides should be used to code for each of the amino acids. Since with only 2 nucleotides there are only 16 possibilities, which falls short of the 20 or more amino acids actually involved, a minimum *block length* of 3 was suggested. However, there are 64 arrangements of 3 nucleotides, so that an additional constraint seems needed to get the number of possibilities back down again. Gamow suggested that 3 nucleotides were indeed used to code for each amino acid, but that the order of their occurrence did not alter the amino acid they produced. Thus, AAC, ACA, and CAA would all code for the same amino acid. (A code in which more than one code word stands for the same object is called a *degenerate* code.) Miraculously, or so it seemed, this leads to exactly 20 *distinct* amino acids that could be coded for – a highly plausible number. (The fact that it is all too easy to go from 64 words to 20 classes by imposing almost any arbitrary extra constraint has been a persistent curse in the history of this problem.)

Gamow's Code – 1954

1.	AAA
2.	CCC
3.	GGG
4.	TTT
5.	AAC,ACA,CAA
6.	AAG,AGA,GAA
7.	AAT,ATA,TAA
8.	CCA,CAC,ACC
9.	CCG,CGC,GCC
10.	CCT,CTC,TCC
11.	GGA,GAG,AGG
12.	GGC,GCG,CGG
13.	GGT,GTG,TGG
14.	TTA,TAT,ATT
15.	TTC,TCT,CTT
16.	TTG,TGT,GTT
17.	ACG,AGC,CAG,CGA,GAC,GCA
18.	ACT,ATC,CAT,CTA,TAC,TCA
19.	AGT,ATG,GAT,GTA,TAG,TGA
20.	CGT,CTG,GCT,GTC,TCG,TGC

Properties:

a. Uniform length (triplet code)

b. Overlapping

$$\text{(ATGCT)} \ldots = \begin{matrix}\text{ATG} & & \\ & \text{TGC} & \\ & & \text{GCT}\end{matrix}$$

c. Totally decipherable (no nonsense)

d. No error detection or correction

Experimental Fate:

Nature does not use an overlapping code.

Gamow's degenerate triplet code was also an *overlapping* code. That is, after using the first, second, and third nucleotides to describe *one* amino acid, the *next* amino acid is described by the second, third, and fourth nucleotides. An overlapping code of this sort implies strong constraints on the types of transitions which can occur from one amino acid to the next, and it was possible to prove by experiment that enough different transitions do occur that no overlapping triplet code could possibly be involved. This was in fact proved by Caltech Research Fellow Sydney Brenner. The main contribution by Gamow was, thus, not that he solved the problem, but that he recognized and stated it.

It is interesting to note that Gamow's code was *totally decipherable*. That is, no matter what sequence of nucleotides is written down, it will always have an interpretation. It is not possible to write a "nonsense" message. Consequently, also, if an error occurs, the erroneous word will be interpreted without any possibility of error correction.

The main reason that Gamow suggested an *overlapping* triplet code was to avoid the synchronization problem which arises with a non-overlapping code. That is, if ATA and CTG are code words for two consecutive amino acids, there is the danger that, when juxtaposed, . . . ATACTG . . . also contains the triplets TAC and ACT, which might code for *other* amino acids, and if these amino acids happened to form *first*, by whatever chemical process is involved, then the sense of the genetic message would be lost.

Another solution to the synchronization problem was offered by F. H. C. Crick in 1956, using a non-overlapping triplet code with what Crick called the *comma-free* property. A code is *comma-free* if when a b c and d e f are two words of the code (distinct or not), then none of the "overlap" words which appear when the comma is dropped from a b c, d e f (such as the words b c d and c d e) are words of the dictionary. For example, if the words BAT and END were in a comma-free dictionary, the words ATE and TEN could not also be in the dictionary (because of bATEnd and baTENd). Also, no word of the type XXX could be in the dictionary, because of the sychronization problem created by . . . XXXXXX . . . .

Using three-letter words formed from a four-letter alphabet, Crick showed that the maximum number of words in a comma-free dictionary is 20, and exhibited examples of such dictionaries.

Crick's Code – 1956

1.	ACA
2.	ACC
3.	AGA
4.	AGC
5.	AGG
6.	CGA
7.	CGC
8.	CGG
9.	ATA
10.	ATC
11.	ATG
12.	ATT
13.	CTA
14.	CTC
15.	CTG
16.	CTT
17.	GTA
18.	GTC
19.	GTG
20.	GTT

Properties:

a. Uniform length (triplet code)

b. Comma-free (bat, end excludes ate and ten)

c. Non-degenerate (nonsense exists)

d. Detects numerous errors

Experimental Fate:

Since XXXX and even XXXXX occur in nature, the comma-free triplet hypothesis is false.

In 1956, at the instigation of Max Delbrück, Caltech professor of biology, Basil Gordon, Lloyd Welch and I obtained some general results of a mathematical nature about comma-free codes using k-letter words from an n-letter alphabet. Welch and I then applied these methods to the biological situation in greater detail in 1957, finding all possible comma-free dictionaries of 20 words with $k = 3$ and $n = 4$.

It was shown that in any message written from any such dictionary, the same symbol could not be repeated consecutively more than three times. When subsequent experimental data showed that nucleotides *were* repeated four and even five times consecutively in the DNA, it was clear that the comma-free triplet hypothesis was not valid – at least not in the form originally envisioned by Crick. (Mathematically, the "no four in a row" is rather profound while "no five in a row" is quite trivial. It is ironic that, experimentally, it was not much harder to observe the fives than the fours.)

On the basis of the best estimates available in 1960, I proposed a type of code dictionary consisting of 24 code words, each six symbols long, where the dictionary is both comma-free and maximally error-correcting. However, the recent experimental breakthrough appears to rule out most of the features of this code from further consideration.

Golomb's Code – 1960

1. TTTTTG
2. GCACTA
3. GGATGT
4. TACTCC
5. GATGGA
6. GCTCAT
7. GTGGCC
8. TTCAGC
9. TGGCAA
10. TCAGAG
11. GGCACG
12. TAGATT
13. AAAAAC
14. CGTGAT
15. CCTACA
16. ATGAGG
17. CTACCT
18. CGAGTA
19. CACCGG
20. AAGTCG
21. ACCGTT
22. AGTCTC
23. CCGTGC
24. ATCTAA

Properties:

a. Uniform length (sextuplet code)

b. Comma-free and orthogonal

c. Non-degenerate (nonsense exists)

d. Error-detecting and correcting

Experimental Fate:

Appears to conflict with experiments of Nirenberg and Ochoa, 1961.

Before turning our attention to the recent revolutionary developments in this field, it is appropriate to mention that many other models, some straightforward and some quite bizarre, had also been proposed for the genetic code during the past eight years. Even where there may have been no influence on the course of biological events, these studies have significantly enriched the literatures of information theory and of combinatorial analysis.

One particularly unusual hypothesis, advanced by a biologist, was that each code word should have the property of coding for any specified amino acid after at most a *simple mutation* – i.e., a change in only one of the symbols of the word. While this postulate has proved to be of no particular merit in genetics, it leads to the notion of "error-distributing codes," which are the precise opposite of the "single error-correcting codes" of information theory, and add considerable elegance to the entire subject.

Recent developments

On November 20, 1961, when I visited the Institute for Genetics in Cologne, I was given two items to read which had just arrived in the day's mail. One was a manuscript by Crick purporting to establish the triplet nature of the genetic code. The other was a glimpse at some partial results of Severo Ochoa of the New York School of Medicine, concerning the nucleotides which seem to be contained in the code words for certain of the amino acids (at least in *E. coli*). It is with such speed that the recent developments in this area have occurred.

Crick had experimented with one of the bacteriophages (bacterium-eating viruses) of the bacterium *E. coli*, and established, at least to his own satisfaction, that all the code words have length three (or, much less likely, some multiple of three), and that synchronization is achieved by starting at one end of the genetic tape, and reading off three symbols at a time. Ochoa, on the other hand, had been extending the work of Marshall Nirenberg of the National Institutes of Health, who had reported several months earlier that the RNA sequence UUUUUU . . . inserted into *E. coli* produced the protein whose amino acid sequence was phenylalanine-phenylalanine-phenylalanine . . . . Ochoa has taken *random* mixtures of various nucleotides, and observed what different amino acids were produced, without learning the exact lengths of the code words, or the specific order of the nucleotides within each code word.

For example, a random mixture of U and C was "decoded" by the *E. coli* machinery, and found to contain not only phenylalanine (presumably from UUU), but also proline, leucine, and serine. Similarly, while AAAAA ... was found to produce *nothing*, a random mixture of U and A yielded tyrosine and isoleucine as well as phenylalanine. A very similar set of experiments by Nirenberg and Matthaei gave largely the same results. Assuming a triplet code, it was even possible to deduce that proline is produced by one U and two C's (in the proper order), while serine is coded by two U's and one C. Some amino acids, notably leucine, definitely seemed to have more than one corresponding code word. The experimental evidence thus leans toward a partially degenerate, triplet code. It is remarkable that Gamow's original guess (page 11) had so many of the correct properties!

A summary of the amino acids and some of the nucleotide combinations which seem to produce them (based on the Ochoa and the Nirenberg-Matthaei data) is given in the following table;

Apparent Correspondence Between RNA Bases and Amino Acids

Amino Acid	*RNA Bases*
Alanine	UCG
Arginine	UCG
Aspartic Acid	UAG
Asparagine	UAA,UAC
Cysteine	UUG
Glutamic Acid	UAG
Glutamine	UCG
Glycine	UGG
Histidine	UAC
Isoleucine	UUA
Leucine	UUC,UUG,UUA
Lysine	UAA
Methionine	UAG
Phenylalanine	UUU
Proline	UCC
Serine	UUC
Threonine	UAC,UCC
Tryptophan	UGG
Tyrosine	UUA
Valine	UUG

For several important reasons, this is not yet the final answer to the coding problem. For one thing, several of the entries will probably change before the list will be accepted as authoritative. More important, the *order* of the nucleotides in each code word remains to be determined. Finally, of the 64 possible triplet code words, a large proportion have not yet been properly tested to see what, if any, amino acid they incorporate.

There is another important kind of data available on the codeword structure; this was obtained when Heinz-Günter Wittman of the Max Planck Institute in Tübingen, Germany, experimented with tobacco mosaic virus (TMV). Mutations were included in the RNA of TMV using nitrous acid, which has the effect of changing C into U and A into G. Then the changes in the amino acid sequence of the "coat protein" of TMV were observed. The mutational transitions are summarized at the right, where the *numbers* on the arrows indicate the frequency with which the particular transition was observed. Note especially how well the part involving proline, leucine, serine, and phenylalanine agrees with the table above, which finds proline, leucine and serine produced from mixtures of C and U, and phenylalanine from poly-U. The rest of Wittman's results can also be readily reconciled with the table.

Proposed experiments

a. One of the crucial features which would distinguish between possible coding systems is the amount of nonsense present. A quantitative comparison of the *amount* of protein produced by either poly-U or poly-C as opposed to random poly-UC (say in equal proportions of U and C) would give very different answers for different coding schemes. The *lengths* of the poly-peptides formed would be informative *if* it is assumed that a nonsense word breaks the chain. The total quantity of protein made would be informative even if the amino acid chain "closes ranks" where nonsense has occurred.

b. Arthur Kornberg, biochemist at the Stanford Medical School, has made the DNA sequence ATATAT ...., which can be used to make the RNA sequence UAUAUA ... . In any triplet code, this groups off as (UAU) (AUA) (UAU) (AUA) ...., and the protein formed may consist of alternating tyrosine and isoleucine. However, it is conceivable that only one amino acid, or perhaps nothing at all, would be produced. There are certain technical difficulties associated with this experiment, but if performed, its outcome would certainly shed considerable light on the entire problem.

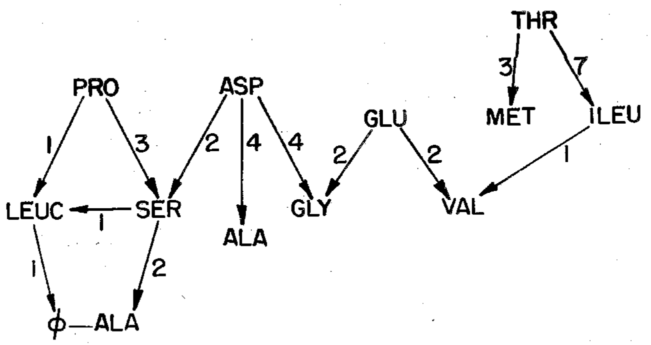
c. If poly-G could be made, it would produce "nonsense" in certain coding schemes but not in others. *This experiment would definitely reduce the* number of possibilities. It might be possible to attach a short poly-G chain to a long poly-U chain, and then observe what prefix (if any) is attached to the resultant poly-phenylalanine stalk.

d. All the code words in the table at the left contain at least one U. It is important to determine whether any of the 27 triplets *not* containing U code for amino acids.

Outlook for the future

If we make a few modest assumptions, the present state of knowledge (or ignorance) concerning the genetic code can be assessed quantitatively. Accepting Crick's conclusion that every code word has three letters, and agreeing that there is only one genetic code in general use, the problem may be formulated thus: Each of the 64 possible *trinucleotides* (triples of RNA symbols, such as AGA and UAC) must be identified either with one of the 20 or so amino acids,

Wittman's Results



or with "nothing." Thus, for each of 64 possible code words, a decision must be made into which of 21 categories to assign it.

Each such assignment represents $\log_2 21 = 4.392$ bits of information, and the entire problem therefore involves the specification of $64 \log_2 21 = 281.1$ bits. Combining the results of Ochoa, Nirenberg, and Wittman, approximately 80 of these bits are already specified. (Knowing that UUU makes phenylalanine immediately supplies 4.392 bits, as does knowing that AAA makes nonsense. Obviously *less* than 4.392 bits is obtained from the fact that at least one of the six triplets UUA, UAU, AUU, UAA, AUA, AAU makes isoleucine.)

At present, then, there are some 200 bits of uncertainty remaining about the genetic code, or 2^{200} ways of filling in the code dictionary consistent with the experimental data at hand. Ultimately, it is hoped that it may be possible to test all 64 triples "directly," for example by attaching each triple in turn as a prefix to a long poly-U chain, and seeing what amino acid (if any) occurs at the start of the resulting polyphenylalanine chain.

Until then, however, it seems worthwhile to make various assumptions regarding the economy, simplicity, optimality, or extremalness of nature's strategy, and explore the consequences of such assumptions insofar as completing the code dictionary by interpolation or extrapolation from the existing data is concerned. Simple experimental tests should then be devised to confirm or refute the assumptions. It would probably be advantageous to write a digital computer program to store and collate all the experimental data as thev become available, thereby reducing the clerical effort required to determine whether or not new models for the code dictionary are consistent with the data of various kinds already established.

A useful geometric model for the coding problem is a 4x4x4 cube, containing 64 cells corresponding to the 64 triplet code words.

The Four Layers of the 4x4x4 Cubic Model of the Genetic Code

UAU	CAU	GAU	AAU
UGU	CGU	GGU	AGU
UCU	CCU	GCU	ACU
UUU	CUU	GUU	AUU

UAC	CAC	GAC	AAC
UGC	CGC	GGC	AGC
UCC	CCC	GCC	ACC
UUC	CUC	GUC	AUC

UAG	CAG	GAG	AAG
UGG	CGG	GGG	AGG
UCG	CCG	GCG	ACG
UUG	CUG	GUG	AUG

UAA	CAA	GAA	AAA
UGA	CGA	GGA	AGA
UCA	CCA	GCA	ACA
UUA	CUA	GUA	AUA

Fortunately, 4x4x4 models of this type are available in toy stores as "boards" for three-dimensional tic-tac-toe! It is clear that the biological problem is to decide which amino acid to put in each of the 64 cells. (Thus, "phenylalanine" should be written into the cell indexed UUU.)

The code words not containing U form a 3x3x3 sub-cube. The cube above can be partitioned into eight 2x2x2 sub-cubes, each containing permitted paths for Wittman's nitrous-acid-induced mutations. Considering the 4x4x4 cube as a three-dimensional chess board, two cells differ by a "single mutation" if and only if a *rook* can go from one cell to the other in a single move. In many ways, this geometric model significantly simplifies the problems of thinking about the genetic code.

The solution of the coding problem will not be the end of genetics research, any more than learning to read is the end of education. In fact, the deciphering of all the code words for all the amino acids represents somewhat less than learning to read, in that it does not include the "punctuation." That is, in addition to code words for the amino acids themselves, there must be encoded instructions to "start protein formation," "stop protein formation," and so forth.

In the terminology of the digital computer field, the storage in the nucleic acid includes *program* as well as *memory*. Wittman's study of tobacco mosaic virus indicates that significantly less than half of the nucleotides in TMV are involved in coding for the "coat protein." There is strong evidence that the nucleotide sequence has significant *regulatory* functions, so that it determines not only what proteins can be produced, but how much of each protein to make, and when. In particular, when certain nutrients are absent from the medium, a bacterium produces the necessary enzymes (proteins) to synthesize these nutrients; but the presence of the nutrients inhibits the production of the enzymes. A complete understanding of the "digital control system" involved in protein synthesis is still many years off, and is sure to engage the serious attention of an ever increasing number of microbiologists.

Looking far into the future, I envision a keyboard with the symbols A, C, G, and U. An operator will type out any sequence of his liking, feed it into a "tape-reader" for processing, and out will crawl the newly designed organism. Later, a more advanced model keyboard, using a "compiler language," will enable the operator to type directly in terms of amino acids and "punctuation marks." I must admit that biologists who understand the technical difficulties are loath to share this vision, but an unconcern for details of implementation is one of the chief advantages of being a mathematician.

As recently as a year ago, the possibility of putting borrowed or synthetic RNA into a convenient organism to obtain the corresponding protein seemed remote indeed. Yet Nirenberg put the RNA strand poly-U into *E. coli*, and got out poly-phenylalanine, a protein almost certainly never made *in vivo* before. Recently, experiments have shown that RNA strands borrowed from many organisms can be inserted into *E. coli* to produce their usual enzymes. This wasn't quite how Dr. Frankenstein went about it, but I believe the implications are even more remarkable.