Proc. Natl. Acad. Sci. USA
Vol. 78, No. 3, pp. 1596–1600, March 1981
Biochemistry

Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification

(rhythmic correlations/comma-free code/mutations)

JOHN C. W. SHEPHERD

Biocentre of the University of Basel, Klingelbergstrasse 70, 4056 Basel, Switzerland

Communicated by Manfred Eigen, December 11, 1980

ABSTRACT **The periodic variations obtained by correlating the relative positions of purines and pyrimidines (and of the four bases thymine, cytosine, adenine, and guanine) in a wide variety of genomes of wholly or partly known sequence suggest that there may be enough of an earlier comma-free coding system (i.e., only readable in one frame) still present to permit determination of the reading frame and approximate extent of the present protein coding stretches. The characteristics of these variations support the hypothesis that these primitive messages were formed of coding triplets having the form RNY (R = purine; Y = pyrimidine; and N = purine or pyrimidine). The base sequences and reading frames that have a minimal deviation from such a message are still good predictors of actual coding regions and reading frames in spite of the many mutations that have occurred since such a genetic code was last in use. In fact, the right frame for almost all the proteins in a number of viruses and various prokaryotes and eukaryotes is deduced purely from purine/pyrimidine information and not by using the normal start and stop signals.**

The concept of a comma-free genetic code was first discussed by Crick *et al.* (1) in 1957 when the form of the code was still unknown. A messenger strand was envisaged in which the triplets ("sense" codons) of a message in one reading frame could not be found anywhere among the triplets in the other two frames ("nonsense" codons). It was shown that a maximum of 20 different sense triplets could be used in such a system, but the idea was abandoned when found not to conform with the actual coding system with its start and stop signals and the use of degenerate codons. More recently, however, Crick *et al.* (2) have considered a possible primitive means of protein synthesis in the absence of ribosomes and using only a messenger strand and a few primitive tRNAs, each having two possible conformations (3, 4). With this hypothesis and using data on the sequence regularity in the anti-codon loops of present day tRNAs, they deduced that a primeval message should conform to the pattern of purines (R) and pyrimidines (Y) RRY RRY RRY . . . .

Subsequently, Eigen (5) further considered this concept and found that an RNY (N = purine or pyrimidine) message (i.e., a mixture of RRY and RYY codons) would also fit the tRNA data and the protein synthesis hypothesis and would have several advantages over an RRY message, particularly the better balance between the numbers of R and Y and the symmetry between the plus and the minus strands. Both of these messages are clearly comma-free codes, but now in the simpler sense that they function solely by distinguishing between R and Y rather than among thymine, cytosine, adenine, and guanine as earlier considered (1).

Rhythmic correlations

To understand the evidence for remnants of a comma-free code, it is necessary to describe briefly the main features of the strong R, Y rhythmic correlations with a period of three bases that have been found, with the help of a computer, in the complete genomes of the DNA viruses ϕX174 (6), G4 (7), and fd (8) and of the generally weaker correlations with the same characteristic features found in a DNA virus [simian virus 40 (9)], a plasmid [pBR322 (10)], a RNA virus [MS2 (ref. 11 and references therein)], and various prokaryotic and eukaryotic genes—e.g., a ribosomal protein gene cluster of *Escherichia coli* (12) and the sea urchin histone genes (13). Typical examples of these periodic variations are shown in Fig. 1, where the total occurrences (*c*) of the base pair YR followed by YY with varying numbers (*n*) of separating bases have been plotted, giving a signal with a period of three bases and maxima at *n* = 0, 3, 6, . . . (to be termed phase 0). Other combinations of one, two, or three bases give similar periodic signals but may have maxima at 1, 4, 7, . . . (phase 1) or 2, 5, 8, . . . (phase 2). For the weaker variations, some averaging of successive counts at intervals of three helped to establish the phase; for some combinations, no patterns could be seen.

Further investigations (data to be published elsewhere) have shown that these variations are significant and are analogous to a periodic wave superimposed on a constant background. All the rhythms have a period of three bases and definite characteristics in amplitude and phase. The stronger rhythms extend up to several hundred separating bases with maxima at regular intervals of three before irregularities start occurring as the locally averaged amplitude (say over 40 periods) becomes smaller. One remarkable feature is that, for any one combination of correlated bases giving a clearly marked periodic signal, the phase is the same in all the genomes examined (with few exceptions)—e.g., in Fig. 1 the phase is 0 in all three genomes. Similar rhythms are also found in T, C, A, G counts, but in many cases a large number of apparently random T, C, A, G rhythms add up to a strong composite R, Y variation, suggesting that the R, Y relationships are a fundamental feature of this phenomenon.

Evolutionary background for method

The above variations in correlation counts (*c*) occur more strongly in the protein-coding regions of the genome and it might be thought that the rhythms could have arisen by natural selection for a more efficient phenotype with improved function in its proteins. However, this does not seem to be the case because certain of the ratios in the rhythms (e.g., RNY/RNR, the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: R, any purine; Y, any pyrimidine; N, purine or pyrimidine.

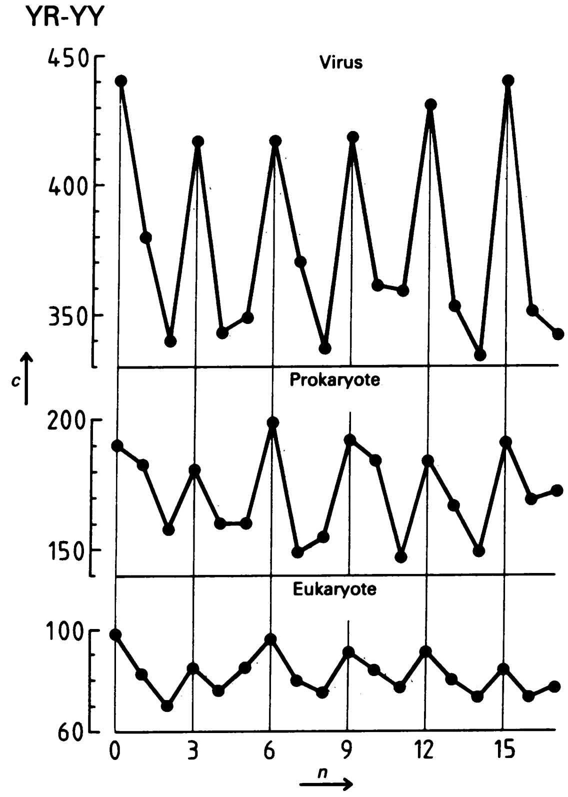

FIG. 1. In the 5′–3′ direction in the coding strand, the total occurrences (c) of a given combination YR followed by YY after n separating bases have been counted [e.g., YR NNNN YY (n = 4)]. Virus is ϕX174; prokaryote is a ribosomal protein gene cluster in *E. coli*; eukaryote is represented by the sea urchin histone genes with all known coding triplets joined together in order.

tendency for a R to be followed by a Y in the second position rather than by a R) are generally most strongly stressed when both forms code for the same amino acid (e.g., in the degenerate codons for glycine, alanine, valine, and threonine in ϕX174, RNY/RNR = 412:130 = 3.17:1, whereas when RNY and RNR code for different amino acids RNY/RNR = 325:292 or 1.11:1).

It is also difficult to account for other aspects of the phenomena, such as the phase and amplitude properties of the rhythms, by such a theory. Computerized simulations, however, lead to the conclusion that the random aggregation of an early messenger strand from roughly equal numbers of the two triplets RRY and RYY (but taking account of the R/Y ratio for the particular genome), followed by a small number of insertions and deletions, together with many point mutations Y ⇄ R, can produce a sequence that will give the characteristic features of these rhythms, an indication of their relative strengths, and almost all the correct phases.

The basic reason for the success of these computations and for the present periodicities can be understood by considering any such RNY messenger strand, say RYY RRY RYY RRY RRY RYY RRY RRY RRY RYY RYY . . . , called the "primeval strand" in the following text. In spite of the extensive mutation that has since occurred, enough of this original R,Y distribution still remains in the present genomes to give the rhythmic counts with phases as now observed.

As a first example, the origin of the rhythmic correlations YR-YY in Fig. 1 may be considered. Counting by hand in the above strand, it is quickly found that YR is only followed by YY if there are n = 0, 3, 6, . . . separating bases. The effect of subsequent mutations on this hypothetical strand will be that some counts for this correlation will occur at intermediate values (n = 1, 2, 4, 5 . . .) in the course of time. For the genomes now investigated, however, maximal counts are still found at n = 0, 3, 6, . . . (phase 0) for YR-YY correlations (see Fig. 1), in full agreement with the phase to be expected from the primeval strand.

In a similar way the phases for other clearly marked rhythms in present-day genomes can be successfully predicted. As a second example, the 64 possible correlations of R,Y triplets with each other (i.e., the eight possible triplets—YYY, YYR, YRY, YRR, RYY, RYR, RRY, and RRR—each followed by itself or by any one of the others) may be considered in the genomes of known sequence (e.g., in ϕX174). Of these 64 correlations, only 36 can be found in the primeval strand (because this contains only six of the above eight triplets). Now the separations found for each of these 36 correlations in such a primeval strand (e.g., n = 2, 5, 8, . . . for YRR followed by YYR) all agree with the phases determined for these correlations in ϕX174 (e.g., phase 2 for this correlation of YRR with YYR). These 36 combinations also give the most significant rhythms in ϕX174. The remaining 28 correlations (e.g., YYY-RRY) cannot be found in the above primeval strand and must have arisen by mutation from this strand. In ϕX174, these 28 have generally weaker patterns or show no rhythmic variation at all. The wide agreement in phases for the strong rhythms in all the present genomes is now understandable in that the rhythms indicate this common origin.

Thus, there is good evidence for the belief that remnants of such a primeval RNY message exist, well preserved in the protein genes (where the number of mutations is thought to have been less than in the intergene regions) and even better in the degenerate codons for the same amino acid. In these codons, the high RNY/RNR ratio is explained because RNY is the primeval form and there may have been little evolutionary advantage from a mutation to RNR (see *Discussion* below). As already mentioned, this is the same type of message as derived by Eigen (5) on quite different grounds.

Method

The above-mentioned genomes were examined to find out whether the present-day genes are still being read in the frames of such early evolutionary comma-free messages. For any given length L (say 60) bases of a genome, the DNA, interpreted as a R,Y sequence in the 5′-3′ direction, is examined in each of the three possible reading frames to determine which R,Y sequence shows the least deviation from a primeval messenger strand formed only from RRY and RYY triplets. Thus, for example, at the start of ϕX174 in Fig. 2*a*, each successive triplet has been considered in the R,Y sequences 1–60 (frame 1), 2–61 (frame 2), and 3–63 (frame 3), and the smallest number of Y ⇄ R mutations required to mutate from either RRY or RYY to the present triplet is noted. Now, after the respective sums (M_1, M_2, M_3) of these minimum mutations for all of the 20 triplets in each of frames 1, 2, and 3 are found, the frame is recorded in which the total mutations are less than in either of the other two frames. This procedure is repeated after moving in steps of S bases (S = 60 for Fig. 2*a*) forward in the 5′-3′ direction along the genome.

In each case the frame determined is plotted against the genome base position N at the midpoint of the 60 bases considered. In the few cases in which equal minimum values occur in the M_1, M_2, and M_3 values, the frame has been taken to be that derived from a slightly longer length of genome still centered at the same position N. The frame plot obtained by this method for ϕX174 is shown in Fig. 2*a*, and it can be compared with the frames and positions of the protein genes already known (6) and shown above the plot.

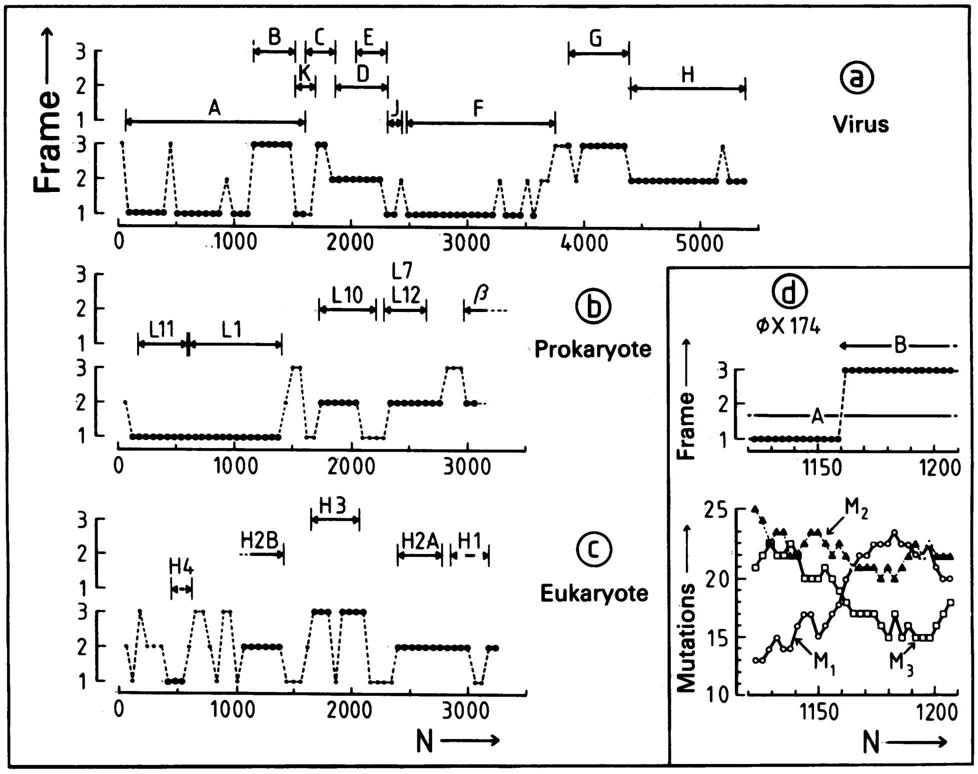

FIG. 2. The frames indicated by the $Y \rightleftarrows R$ mutation counts are plotted beneath the frames and positions of the protein genes as already known (6, 12, 13). (*a*) Counting length $L = 60$ and step forward $S = 60$ in φX174 (5386 bases treated as a loop, with base 1 in this figure corresponding to base 3917 in ref. 6). (*b* and *c*) $L = 120$ and $S = 60$ in a ribosomal gene cluster in *E. coli* (3072 bases) and in the sea urchin histone genes and spacers (taking only the known codons in the incompletely sequenced genes for H4, H2B and H1, and joining over the unsequenced gaps elsewhere to give 3253 bases), respectively. (*d*) More detail near the start of the B gene in φX174 by taking $L = 60$ and $S = 3$. ●, Points considered significant with respect to protein frame determination. In *d*, the mutation counts are shown thus: ○, M_1; □, M_2; ▲, M_3.

A similar procedure has been followed in Fig. 2*b* with the ribosomal gene cluster from *E. coli* (12) and in Fig. 2*c* for the known parts of the sea urchin histone genes and spacers (13). For these genomes, which have been found to have generally weaker periodic correlations than the DNA viruses, the counting length *L* has been taken to be 120 with S = 60. More detail can be obtained in all cases by moving forward in smaller steps. This is too lengthy to show here for a complete genome but an example is given in Fig. 2*d* for the region in φX174 where the protein gene B begins. In this case, the mutations M_1, M_2, and M_3 have also been plotted to illustrate the method.

To evaluate the success of the method, it is first noted that, if *L* is taken approximately equal to the whole known length of any one of the proteins considered, then, except in the case of the short overlapping K and E genes in φX174 [of which E is thought to be a later evolutionary development (14)], the correct frame for every protein in Fig. 2 is deduced. Similar success has been obtained in all the above listed sequences and elsewhere. Second, when shorter lengths *L* of these sequences are used, as in Fig. 2, a strong indication for the presence of a protein gene is given when the frame stays constant over a considerable length of genome, apart from a few irregularities due to mutations (e.g., gene D or H in Fig. 2*a*). Also, in spite of the large number of mutations that must have occurred, there appears to be enough of the original comma-free stretches to give a reasonably good picture of the extent of many of the present protein genes. In some cases, by analyzing with a short step length S, the positions of starts and stops are given with surprising accuracy (e.g., in Fig. 2*d* a gene start is predicted at *N* ≈ 1160 and the ATG start signal of gene B is actually at base 1158).

Other R,Y comma-free codes

As a further test of the present hypothesis, it is interesting to consider if any of the other possible R,Y comma-free codes would fit the present data. To find these codes, a similar argument to that given for the T,C,A,G case (1) can be used. Of the eight possible triplets YYY, YYR, YRY, RYY, YRR, RRY, RYR, and RRR, the triplets YYY and RRR are clearly impossible as sense codons because they could be read in another frame when repeats occur (e.g., YYY YYY). The remaining six triplets divide into two cyclic sets—YYR, YRY, RYY and YRR, RRY, RYR. It is easily seen that only one triplet from each set can be used to satisfy the comma-free conditions (e.g., YYR YRY would not be possible because YRY can be seen in frame 2). This gives the nine possible pairs, YYR with YRR, YYR with RRY, etc. Of these only three—RRY with RYY, RYR with YYR, and YRY with YRR—when randomly mixed in a primeval messenger strand, predict all 36 correct phases for strong correlations of triplets with triplets (as already tested above for the RRY and RYY strand in comparison with φX174). The six other combinations all give incorrect phases (e.g., 18 of these 36 phases wrong by using YYR and YRR). These three successful combinations are the R,Y triplets found in frames 1, 2, and 3, respectively, of the proposed primitive message RRY RYY RRY .... To decide between these three alternatives for the comma-free code (and doubly eliminate the six others), the present protein frame determining method was tried with each of these triplet pairs and a clear-cut result was obtained: only the RRY with RYY combination gives the correct protein frames, and indications of the extent of the protein coding stretches in the genomes tested. It also may be noted that such correct predictions of rhythmic phases and of protein frames could not be achieved by the assumption of only one of these triplets in the primeval message [e.g., RRY as suggested by Crick (2)]. Thus, of all these alternatives the present evidence points clearly to the former use of an RNY primeval strand.

Discussion

A number of alternative hypotheses have been considered to account for the present phenoma but as yet none provides an adequate explanation. For example, with the rhythms from the viral genomes first investigated, it might have been thought that the packing requirements of the DNA in the viral capsid could have some connection with these effects. This cannot be the case, however, because the phenomena have now been observed in a wide variety of other nonviral genomes. Another idea is that the rhythms are merely a reflection of the use of similar proportions of certain common amino acids in the proteins considered. Some simple tests, however, reveal features not explainable by this proposal but accountable in the present theory.

First, a good part of the rhythmic amplitude is due to the uneven use of degenerate codons (as discussed above) and the amplitude is considerably reduced when these are used evenly. Second, if the order of existing coding triplets in a protein gene is altered by random mixing, the rhythms are further reduced (see later explanation of this point). Third, although the amino acid composition of the various proteins tested varies considerably, it becomes evident by a count of the coding triplets used why the protein frame is indicated so well. In this frame, the counts of the proposed original triplets of the type RNY exceed those of the once-mutated ($Y \rightleftarrows R$) types YNY or RNR, and these in turn are greater than the counts for the twice-mutated type YNR (e.g., for the codons used in all the proteins in ϕX174, RNY = 737, YNY = 591, RNR = 422, and YNR = 384 or in fd, RNY = 750, YNY = 521, RNR = 368, and YNR = 314). Mutation away from an original strand RNY gives a satisfactory explanation for these three phenomena and also provides the best possible means yet known to predict the phase and amplitude effects of the rhythmic correlations.

From the computerized simulations, it is found that rhythms of amplitude comparable to those in the ϕX174 or fd genes are given when the number of random point mutations $Y \rightleftarrows R$ applied to an assumed RNY original message is roughly equal to half the number of bases in this message. To simulate the weaker rhythmic effects in other genomes, such as simian virus 40, MS2, or some prokaryotic or eukaryotic genes, even more mutations must be applied. The changes of indicated original reading frame between the genes, as derived by the present method in Fig. 2*a* (e.g., the change from frame 3 to frame 2 between the G and H genes of ϕX174) suggest that insertions or deletions (with lengths not equal to a multiple of three bases) have occurred there or that shorter messages were joined to give this change of frame. Many of the changes of indicated frame within the genes (such as those seen in the A, F, G, and H genes of ϕX174) can also be ascribed to insertions or deletions. The base sequence within such an irregularity still approximates well to an original message and makes good base correlations within itself, but the codons in it are out of phase with the primeval message in the rest of the gene. Thus, if all the coding triplets in the whole gene are randomly mixed, many of these out-of-phase triplets will then be found in an environment of in-phase neighbors and the rhythmic correlation amplitude (for relatively small separations of the correlated bases) decreases, as noted above.

Another point for further discussion is that the original RNY message is best preserved in the degenerate codons for the majority of genes examined. At first sight this may be difficult to understand in the light of some recent observations on the divergence of related genes, in which silent mutations have occurred more frequently than those changing an amino acid. For example, in the nucleotide sequences for trpA of *Salmonella typhimurium* and *E. coli*, 75% of the codon differences are due to synonymous codon changes (15). More evidence comes from the differences between the ϕX174 and G4 genes and the mRNAs of some vertebrate α- and β-globins; a survey has shown that, of the single-base changes in coding triplets, the number of silent mutations is 1.8 to 3.0 times more than the proportion of 25% to be expected statistically (16).

In explaining the present effects, however, it should be remembered that the total $Y \rightleftarrows R$ mutations away from a primeval RNY message in an estimated time of about 3500 million years or more must be taken into account, whereas the above observations are on the comparatively recent divergence of genes that are coding for proteins already largely perfected in their function. On the assumption that the present genetic code was used for the translation of the original message, then RNY could code for a maximum of eight amino acids. Thus, the primeval genes must have suffered many mutations to provide for new amino acids and for the other amino acid changes necessary to develop and improve a protein's efficiency for a particular purpose.

Although generalization is difficult because of the great variety of proteins developed from primitive messages and of likely intermediate changes in the rate of protein evolution due to adaption to new conditions, it would seem that one generally could distinguish two main periods in the evolution of a gene. In the first period, starting from the primeval strand, many mutations giving amino acid changes were advantageous and were selected for, in order to perfect the protein's function. During this time, when considerable protein improvement occurred, organisms experiencing silent nucleotide changes in their genes, which brought little if any functional advantage, would be eliminated together with the other older prototypes in the presence of the considerably more efficient new prototypes developed by nondegenerate codon changes.

In the second period, however, when the protein was already efficient, mutations causing amino acid changes generally decreased the functional efficiency and were selected against. In this second period, one would expect a greater proportion of synonymous changes, as has been observed in the few related genes whose sequences have now been determined (e.g., the vertebrate globins above). Whether these synonymous changes give further functional advantage (e.g., by changing the mRNA structure), now that a sophisticated system for transcription, translation, and control has been developed, does not seem altogether clear as yet. The possibility of neutral changes has also been proposed in the genetic drift theory (17, 18). Corresponding to the features of such a second period, the amino acid replacement rate has become much lower within the last 300–500 million years for vertebrate globins (19), after an intermediate period of rapid change in the first vertebrates as the ancestral gene duplicated and diversified.

Taking all these factors into account and assuming that synonymous changes were relatively few compared to the many mutations causing amino acid changes during the long earlier evolutionary period and later intermediate periods of rapid evolution, a possible explanation for the present minimal mutation away from the primeval RNY message in the degenerate codons is seen.

Conclusion

The method for determining the reading frame of a protein from R,Y information has now been successful for a wide variety of protein genes, in other viruses [the fowl and human influenza virus hemagglutinin genes (20, 21)] in prokaryotes [the trpA genes of *S. typhimurium* and *E. coli* (15), the chloramphenicol transacetylase gene in the transposon Tn*cam* 204 (22), the *E. coli recA* gene (23), and the outer membrane lipoprotein mRNA gene (24)]; in eukaryotes [the bovine corticotropin/lipotropin precursor gene (25) and preproparathyroid hormone gene (26), human mitochondrial cytochrome oxidase subunit II gene (27), and the mRNAs of chicken ovalbumin (28) and of human and rabbit β-globins (29, 30)]. No case has yet been encountered in which an incorrect reading frame has been deduced for the whole gene (apart from overlapping genes or sequences for short peptides), but some irregularities occur within the genes, as already discussed. The method can also be used on the DNA sequences of eukaryotic genes containing introns and exons—e.g., the mouse α- and β-globin genes (31, 32) and the part of the ovalbumin gene now known (33). The correct reading frames are then indicated for the exon stretches.

The present hypothesis still must be treated with caution because so little is known of early life processes. Exceptions to some of the generalizations made are also to be expected. The use of an RNY code as described here, however, it successful in explaining the characteristics of the periodic variations, including their phases and relative strengths and additionally some features of the nonrandom use of codons in various genomes. Once the DNA sequence of a genome is known, the method may be used to determine likely frames or extent of protein genes or even to help in detecting some types of sequence errors (e.g., a single base insertion or deletion). Many aspects of the method are not yet fully explored but a number of possibilities to determine more of the evolutionary history of a genome are evident. As more genomes of known sequence become available, it will be interesting to see if these effects are universal and if comma-free coded messages of the type RNY were the ancestors of most of the present-day genes.

Many thanks are due to Drs. T. Bickle and G. Büldt for helpful discussions and to Professor G. Schwarz for his encouraging support. Acknowledgement is also made of grants given by the Swiss National Science Foundation.

1. Crick, F. H. C., Griffith, J. S. & Orgel, L. E. (1957) *Proc. Natl. Acad. Sci. USA* **43**, 416–421.
2. Crick, F. H. C., Brenner, S., Klug, A. & Pieczenik, G. (1976) *Origins Life* **7**, 389–397.
3. Fuller, W. & Hodgson, A. (1967) *Nature (London)* **215**, 817–821.
4. Woese, C. (1970) *Nature (London)* **226**, 817–820.
5. Eigen, M. (1978) *Naturwissenschaften* **65**, 341–369.
6. Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchison, C. A. III, Slocombe, P. M. & Smith, M. (1978) *J. Mol. Biol.* **125**, 225–246.
7. Godson, G. N., Barrell, B. G., Staden, R. & Fiddes, J. C. (1978) *Nature (London)* **276**, 236–247.
8. Beck, E., Sommer, R., Auerswald, E. A., Kurz, Ch., Zink, B., Osterburg, G., Schaller, H., Sugimoto, K., Sugisaki, H., Okamoto, T. & Takanami, M. (1978) *Nucleic Acids Res.* **5**, 4495–4503.
9. Reddy, V. B., Thimmappaya, B., Dhar, R., Subramanian, K. N., Zain, B. S., Pan, J., Ghosh, P. K., Celma, M. L. & Weissman, S. M. (1978) *Science* **200**, 494–502.
10. Sutcliffe, J. G. & Ausubel, F. M. (1978) in *Genetic Engineering*, ed. Chakrabarty, A. M. (CRC, West Palm Beach, FL), pp. 94–96.
11. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G. & Ysebaert, M. (1976) *Nature (London)* **260**, 500–507.
12. Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. & Dennis, P. P. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1697–1701.
13. Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H. O. & Birnstiel, M. L. (1978) *Cell* **14**, 655–671.
14. Barrell, B. G., Air, G. M. & Hutchison, C. A. III (1976) *Nature (London)* **264**, 34–41.
15. Nichols, B. P. & Yanofsky, C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5244–5248.
16. Jukes, T. H. & King, J. L. (1979) *Nature (London)* **281**, 605–606.
17. Kimura, M. (1968) *Nature (London)* **217**, 624–626.
18. King, J. L. & Jukes, T. H. (1969) *Science* **164**, 788–797.
19. Goodman, M., Moore, G. W. & Matsuda, G. (1975) *Nature (London)* **253**, 603–608.
20. Porter, A. G., Barber, C., Carey, N. H., Hallewell, R. A., Threlfall, G. & Emtage, J. S. (1979) *Nature (London)* **282**, 471–477.
21. Min Jou, W., Verhoeyen, M., Devos, R., Saman, E., Fang, R., Huylebroeck, D. & Fiers, W. (1980) *Cell* **19**, 683–696.
22. Marcoli, R., Iida, S. & Bickle, T. A. (1980) *FEBS Lett.* **110**, 11–14.
23. Horii, T., Ogawa, T. & Ogawa, H. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 313–317.
24. Nakamura, K., Pirtle, R. M., Pirtle, I. L., Takeishi, K. & Inouye, M. (1980) *J. Biol. Chem.* **255**, 210–216.
25. Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A. C. Y., Cohen, S. N. & Numa, S. (1979) *Nature (London)* **278**, 423–427.
26. Kronenberg, H. M., McDevitt, B. E., Majzoub, J. A., Nathans, J., Sharp, P. A., Potts, J. T., Jr. & Rich, A. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4981–4985.
27. Barrell, B. G., Bankier, A. T. & Drouin, J. (1979) *Nature (London)* **282**, 189–194.
28. McReynolds, L., O'Malley, B. W., Nisbet, A. D., Fothergill, J. E., Givol, D., Fields, S., Robertson, M. & Brownlee, G. G. (1978) *Nature (London)* **273**, 723–728.
29. Baralle, F. E. (1977) *Cell* **12**, 1085–1095.
30. Efstratiadis, A., Kafatos, F. C. & Maniatis, T. (1977) *Cell* **10**, 571–586.
31. Nishioka, Y. & Leder, P. (1979) *Cell* **18**, 875–882.
32. Konkel, D. A., Tilghman, S. M. & Leder, P. (1978) *Cell* **15**, 1125–1132.
33. Robertson, M. A., Staden, R., Tanaka, Y., Catterall, J. F., O'Malley, B. W. & Brownlee, G. G. (1979) *Nature (London)* **278**, 370–372.