J Mol Evol (1981) 17:94–102

Journal of Molecular Evolution



# Periodic Correlations in DNA Sequences and Evidence Suggesting Their Evolutionary Origin in a Comma-less Genetic Code

John C.W. Shepherd

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

**Summary.** Strong rhythms with a period of three bases have been seen while correlating the relative positions of purines and pyrimidines and of the four individual bases in the complete DNA sequence of the viruses $\phi$X174, G4 and fd. Generally weaker variations of the same type have been found in the DNA virus SV40, the plasmid pBR322, the RNA virus MS2, and elsewhere in procaryotes and eucaryotes (e.g. in a ribosomal protein gene cluster of *E. coli* and the sea urchin histone genes). From the interrelation of four-base with purine-pyrimidine rhythms it seems that the purine-pyrimidine relationships have a basic significance. An explanation is proposed in terms of the former use of a comma-less genetic code (i.e. readable only in one frame) of the general form RNY (R = purine, Y = pyrimidine and N = purine or pyrimidine). In spite of subsequent mutation, there appears to be still enough of the primitive messages remaining to produce these periodic variations with their characteristic properties in phase and amplitude. Particularly good evidence for this hypothesis is provided by the fact that the phases for the stronger rhythms are the same in all the genomes tested and can be successfully predicted by a simple consideration of the original RNY pattern. With regard to amplitude it can be similarly foreseen which variations will be more clearly marked than others. The observed behaviour of the amplitude as the separation between correlated bases increases is also explained by the insertions, deletions and point mutations which have occurred. Additionally it is possible to account for some notable features of the non-random use of codons for the same amino acid by this theory.

**Key words:** DNA purine-pyrimidine rhythms – Codon usage – Primeval message – Mutations

## Introduction

An investigation of the base correlations in the recently determined complete DNA sequences of viruses $\phi$X174 (Sanger et al. 1978), G4 (Godson et al. 1978), fd (Beck et al. 1978), SV40 (Reddy et al. 1978; Fiers et al. 1978) and plasmid pBR322 (Sutcliffe 1979) was first undertaken. This report will describe the main features of the purine-pyrimidine and T,C,A,G rhythms which have been found. The correlation counts of certain base combinations at varying separations give a signal variation analogous to a wave imposed on a constant background. This wave has a period of three bases, and very definite characteristics in phase and amplitude. Subsequent work on a number of other genomes including procaryotic and eucaryotic genes (to be summarized in conclusion) gives evidence that these rhythms may be a widespread phenomenon. A survey of these effects together with the results obtained from some evolutionary simulations will suggest a relatively simple hypothesis to explain these observations.

## Methods and Results

### *Purine-Pyrimidine Correlations*

With the aid of a computer all occurrences of a given sequence, e.g. YRY of pyrimidine (Y) and purine (R) were found in one strand of the DNA, as observed in the 5′- 3′ direction and considered for each of the above first five genomes. Then, looking in this direction forward from each such sequence, a count c was made of all the cases in which a second sequence, e.g. YYR, is found separated from the first sequence by n bases of any type, e.g. (YRY)NNNNN(YYR). Care was taken that no such pair was counted twice in the loop. This correlation will be denoted by YRY.YYR and a similar notation used throughout, while determining how the count c varies with increasing n. Some of the results for $\phi$X174 are illustrated in Figs. 1a, b and c for cases

0022–2844/81/0017/0094/$01.80

in which the two sequences contain one, two or three bases, respectively. In Fig. 1a all possible correlations between single bases are shown and it was found that maxima occur regularly every three bases (up to much higher n values than illustrated) with a few exceptions (see e.g. an irregularity at n = 13 in the Y.R counts). The minima counts are up to ~10% less than the maxima. At any separation n, the counts for R.Y can be easily seen to be equal to those for Y.R in any loop (from a consideration of the correlations round the loop at fixed intervals until the starting base is again reached – more than once round the loop if necessary to include all the bases). It can also be shown that Y.Y. – R.R (even for a random loop) for any value of n is equal to the difference Y – R between the total numbers of pyrimidines and purines in the complete genome (by subtracting the two equations. Y.Y + Y.R = Y and R.R + R.Y = R, which are obviously true at any n). Hence Y.Y. will follow the same pattern as R.R but at a higher level of counts c (for $\phi$X174, Y.Y. – R.R = Y – R = 2841 – 2145 = 296). No such regular patterns of counts could be obtained from any randomly generated series with the same relative numbers of Y and R as $\phi$X174. This is illustrated for the combination Y.R in Fig. 1a. For the two and three base combinations, Fig. 1b and c show a typical selection from the possible permutations with a random control. For pairs (e.g. YR.YR) the differences between maximum and minimum counts (to be termed amplitudes) range up to ~35% of the maximum values, for triplets (e.g. YYR.YRY) up to ~50%. In $\phi$X174 such regular rhythms are seen by this simple correlation method for about 11 of the 16 possible pairs and for about 25 of the 64 combinations of triplets.

*Phases*

A comparison with the random controls shows clearly that the rhythms are a significant phenomenon and they occur in three different phases, 0,3,6,..., 1,4,7,..., and 2,5,8,... (to be termed phases 0,1 and 2), the n values at which the maxima are found. Examining the computer output for all five genomes and for all possible correlations of single bases, pairs and triplets, the three rhythm is found to have more successful combinations for $\phi$X174, G4 and fd than for SV40 and pBR322, where the amplitudes are generally less (a comparison for YY.YR is shown in Fig. 1e). In all cases, however, where the rhythms, for any one combination are clearly marked in different genomes, then (with a few exceptions to be discussed later) they have the same phase, e.g. phase 2 in Fig. 1e, thus indicating the general nature of the phenomenon. (The sequence for bacteriophage G4, which more recently became available, was found to give patterns very similar to those of the closely related $\phi$X 74, despite an overall 39% base sequence mismatch.) For the weaker rhythms, local averaging of a number of successive counts at intervals of three bases helped to raise the signal above the noise and determine the phase.

*T,C,A,G Rhythms*

Some of the T,C,A,G rhythms similarly found are also well marked and have the same phase in different genomes. It is also interesting to see how the patterns from certain Y,R combinations correspond to the composite T,C,A,G correlations. One example of this is found in Fig. 1g, where the four composite correlations C.T (phase O), T.C (phase 1), C.C and T.T (phase 2) add up to the Y.Y correlation (phase 2) given in Fig. 1a. From the 16 combinations, two with two, adding up to the overall correlation YR.YR (phase 1) seen in Fig. 1b, three are shown in Fig. 1h, one with phase 0, two with phase 1 and one with phase 2. Many of the others showed no pattern, although one of them, TG.TG (phase 1) has a particularly strong rhythm and is shown for n = 0 to 59 in Fig. 1f (no phase error at all until n = 234).

A notable feature is that seemingly random composite T,C,A,G correlations add up to a significant Y,R total rhythm. This can be illustrated by a typical case like YRY.YRY, where of the 64 composite T,C,A,G correlations, 61 are apparently random and only three (TAT.TGC, TGT.TGC and TGC.TGC) appear to have a marked rhythm. Nevertheless, even after subtracting these, 3 the 61 remaining counts still add up to a very significant Y,R rhythm. Thus the Y,R relations seem to have a basic importance in these phenomena.

*Variation of Amplitude*

Whereas the mean value of counts c (the background) stay reasonably constant as the separation n increases (although being different for the various groups correlated and for different genomes), the amplitude of the rhythms shows considerable variation. If the differences are taken between successive maxima and minima, rather irregular individual values of amplitude are obtained, but a general tendency to decrease at first is clearly seen in $\phi$X174 and fd. Averaging each forty successive amplitudes for the correlation YR.YR in these two genomes to smooth out the local variations and disregarding any phase errors at larger n values (the pattern stays perfect in phase only up to n = 343 and 375, respectively) the mean amplitude A is shown plotted in Fig. 2a for n = 0 to 2501. The two genomes have already shown themselves very similar in the Y,R and T,C,A,G patterns observed and now it is seen that both graphs start with approximately the same downward slope, followed by more irregular undulations. (The amplitude graph for G4 followed much the same form as that for $\phi$X174).

## Discussion

*Preservation in Codons*

One of the first thoughts that comes to mind is to see if the patterns might be a reflection of protein structure, i.e. that they might have been developed by natural selection for a phenotype with more efficient proteins. Hence, for example, the tendency for a purine to be followed by a pyrimidine in second position rather than a purine was examined in $\phi$X174. For all the triplets in the whole genome RNY:RNR = 1388:1157 or 1.20:1 from the data for Fig. 1a. For the triplets coding for amino acids, however, RNY:RNR = 737:422 or 1.75:1, as seen from Table 1. This table also shows that this higher ratio is due mainly to the much higher value of RNY:RNR (412:130 or 3.17:1) in the degenerate codons for glycine, alanine, valine and threonine. When the two forms code for different amino acids, RNY:RNR = 325:292 = 1.11:1. (The case of isoleucine which has ATG as one of the RNR forms has been included in this last category although its other codons much favour the RNY form). For G4 and fd, similar results are obtained, the ratio RNY:RNR again being highest in the degenerate codons (439:191 = 2.30:1 and 422:121 = 3.49:1, respectively) compared with the lower ratios (1398:1222 = 1.14:1 and 1647:1256 = 1.31:1) for the whole genomes.

Thus the fact that this RNY to RNR preference is stronger in the protein codons than elsewhere does

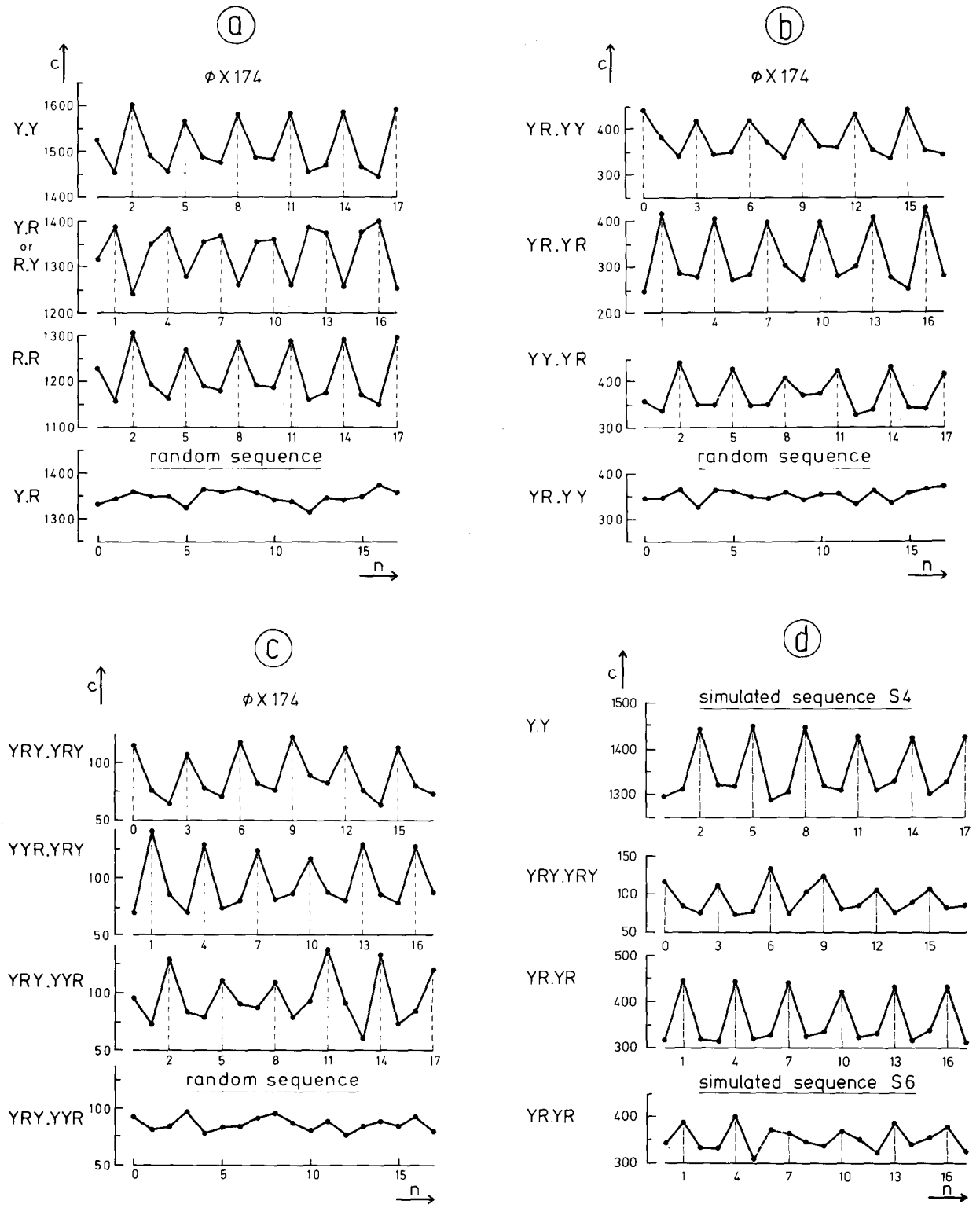
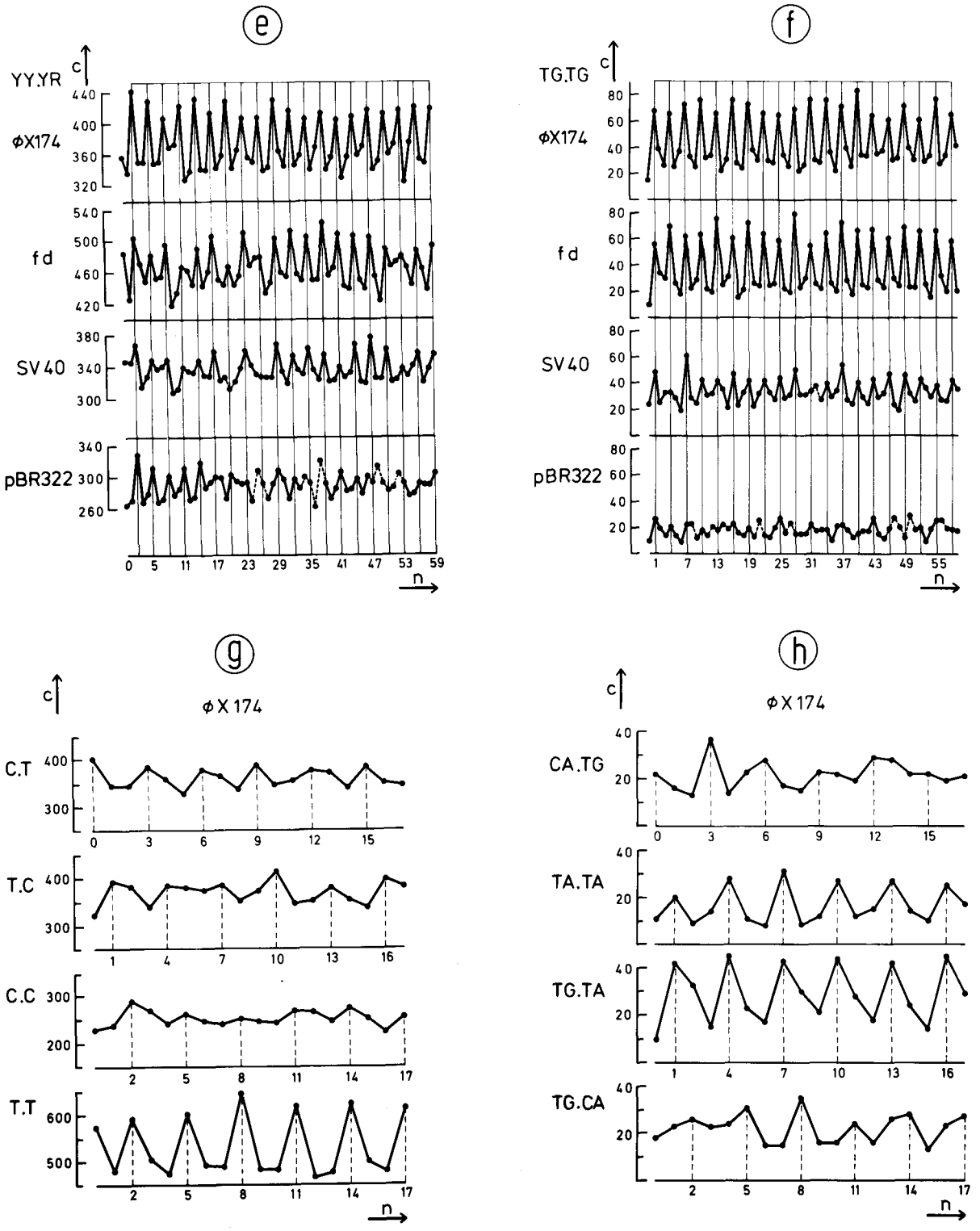



**Fig. 1a-h.** Correlation counts c plotted against separating base count n, e.g. YY.YR indicates the correlation of YY with YR. **(a)**, **(b)**, (c) typical purine-pyrimidine rhythms for $\phi$X174 together with comparable random sequence counts. **(d)** pattern for simululated sequence S4, obtained by random aggregation of RRY and RYY triplets and subsequent mutations (4 insertions, 3 deletions and 3250 Y $\rightleftarrows$ R point mutations). S6 shows the effect of a further 1500 point mutations on S4.
(e), (f) examples of purine-pyrimidine and four-base ryhthms maintained in all genomes with the same phase (2 and 1, respectively, for the cases shown). (g) the 4 composite T,C,A,G rhythms adding up to the Y.Y pattern seen in 1a. **(h)** 4 of the 16 possible T,C,A,G combinations, adding up to YR.YR, shown in 1b


e
YY.YR
c
440
400
360
320
φX174
540
500
460
420
fd
380
340
300
SV40
340
300
260
pBR322
0 5 11 17 23 29 35 41 47 53 59
n
f
TG.TG
c
80
60
40
20
φX174
fd
SV40
pBR322
1 7 13 19 25 31 37 43 49 55
n
g
c
φX174
C.T
400
300
0 3 6 9 12 15
T.C
400
300
1 4 7 10 13 16
C.C
300
200
2 5 8 11 14 17
T.T
600
500
2 5 8 11 14 17
n
h
c
φX174
CA.TG
40
20
0 3 6 9 12 15
TA.TA
TG.TA
1 4 7 10 13 16
TG.CA
2 5 8 11 14 17
n

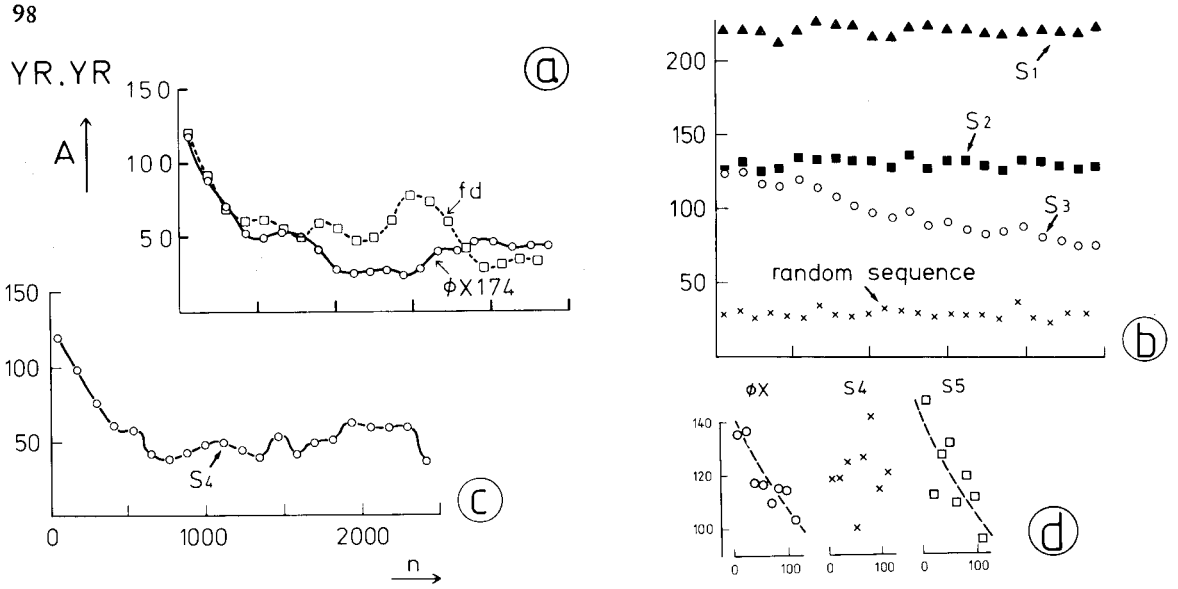



**Fig. 2a-d.** Mean amplitude A of rhythmic variations plotted against number n of separating bases for correlation YR.YR in: (a) $\phi$X174 and fd; (b) simulated genomes S1,S2,S3 and random sequence; (c) simulated genome S4; (d) $\phi$X174 and simulated genomes S4, S5. Amplitudes have been averaged in 40's for (*a*), (*b*), (*c*) and in 5's (*d*). Notation (*a*) ○ $\phi$X174, □ fd; (*b*) ▲ S1, ■ S2, ○ S3, x random sequence; (*c*) ○ S4, (*d*) ○ $\phi$X174, x S4, □ S5. Horizontal scale is the same in (*a*), (*b*) and (*c*)

**Table 1.** Use of RNY and RNR triplets in $\phi$X174

| RNY | | | | RNR | | | |
|---|---|---|---|---|---|---|---|
| | | Whole genome | Protein codons | | | Whole genome | Protein codons |
| *Same amino acid* | | | | | | | |
| Gly | GGT | 100 | 63 | Gly | GGA | 63 | 19 |
| | GGC | 72 | 41 | | GGG | 20 | 4 |
| Ala | GCT | 149 | 103 | Ala | GCA | 60 | 16 |
| | GCC | 71 | 29 | | GCG | 67 | 21 |
| Val | GTT | 140 | 74 | Val | GTA | 54 | 12 |
| | GTC | 65 | 17 | | GTG | 66 | 16 |
| Thr | ACT | 89 | 61 | Thr | ACA | 47 | 14 |
| | ACC | 62 | 24 | | ACG | 63 | 28 |
| Totals | | 748 | 412 | | | 440 | 130 |
| *Different amino acids* | | | | | | | |
| Ile | ATT | 127 | 69 | Ile | ATA | 60 | 5 |
| | ATC | 56 | 17 | Met | ATG | 140 | 60 |
| Asn | AAT | 102 | 61 | Lys | AAA | 133 | 74 |
| | AAC | 67 | 36 | | AAG | 93 | 47 |
| Ser | AGT | 52 | 13 | Arg | AGA | 73 | 11 |
| | AGC | 53 | 7 | | AGG | 74 | 2 |
| Asp | GAT | 101 | 70 | Glu | GAA | 70 | 35 |
| | GAC | 82 | 52 | | GAG | 74 | 58 |
| Totals | | 640 | 325 | | | 717 | 292 |
| Grand totals | | 1388 | 737 | | | 1157 | 422 |

The codons for the proteins A – H, K and J (see Sanger et al., 1978) have been counted. Those for the protein A* (formed by a later start in the protein A gene; see Linney and Hayashi, 1974) are not included, since they have already been counted for protein A

not seem to be a reflection of protein structure since it is strongly maintained even if both combinations code for the same amino acid. Indeed it is especially emphasized in this case, rather than when RNY and RNR code for different amino acids. This could well have a logical explanation. Supposing an originally stronger occurrence of RNY than RNR in the whole genome, a mutation from RNY to RNR in the protein coding part would have no effect on the protein when both forms code for the same amino acid. In the case when the two forms code for different amino acids, however, there is a chance that a new more efficient protein is created and eventually new organisms with the RNR form at this genome position could supersede the older ones. Hence this finding that the present rhythms are generally stronger in protein coding stretches than in inter-gene regions (also for the genomes later considered; see conclusion) seems to suggest that there have been less mutations in protein genes than elsewhere and less still in the degenerate codons for the same amino acids. Here the sum total mutations in a very long time span are being considered. (Further discussion on this point is given below in the mutation section).

*Evolutionary Simulations*

With this indication that the effects could be due to a remnant of some early evolutionary base sequence, and keeping in mind the underlying Y,R significance, a number of different computer simulations were tried to see if a messenger strand could be synthesized with similar rhythmic properties to the present genome. To obtain the phase and amplitude characteristics required, the best success was eventually achieved by the random aggregation of RYY and RRY triplets, followed by a suitable number of insertions, deletions and point mutations. One such computation will be described in detail in order to illustrate the general method and the effects of mutations on the patterns.

A sequence S1 of length 5385 bases is first formed by the random aggregation of RRY and RYY triplets (taking equal probabilities for either type occurring at each successive position) and then randomly applying 2650 Y $\rightleftharpoons$ R point mutations (transversions). This sequence shows rhythms very similar to those observed for the 5386 long $\phi$X174 genome in Fig. 1a,b and c, correct in their phases, but considerably too large in amplitude. Also, for any particular rhythm the amplitudes A (averaged as before in 40's) when plotted against separation n give a plot (see S1 in Fig. 2b) quite unlike the actual plots for $\phi$X174 and fd in Fig. 2a. For S1 the amplitude stays approximately constant with increasing separation n between the correlated base pairs, as could have been anticipated from such an aggregation method and randomly distributed point mutations.

The general level of mean amplitudes may be lowered by applying a further 600 such point mutations to produce the sequence S2 (amplitude plot in Fig. 2b), and the effect of inserting one base at position 5386 to form a sequence S3 of the same length as $\phi$X174 is now tried. The amplitude plot produced shows A decreasing with increasing n (see Fig. 2b). The cumulative effect of this insertion is easily understood, since it puts the triplets following it 'out of phase' with those before. Evidently the amplitude variation is very sensitive to insertions or deletions, and by three times randomly applying a single base insertion and a single base deletion to S3, the resultant sequence S4 gives an amplitude plot (see Fig. 2c) approximating much better to that for $\phi$X174 in Fig. 2a. (A short length of DNA for insertion or deletion, if not a multiple of three, would have a very similar effect to a single base insertion or deletion.) Undulations as in the amplitude graphs of $\phi$X174 and fd after their initial steep fall are seen in the simulation and can now be understood (apart from the smaller random variations as seen in S1, S2 and S3) as the effect of insertions and deletions, taking into account the locations at which they occur in the genome and the chance that some sections of genome may return to an 'in phase' condition after several such random phase shifts. Three typical correlation patterns (Y.Y, YRY.YRY and YR.YR) obtained from S4 are shown in Fig. 1d. They are clearly similar to the same correlations in $\phi$X174 and agree with them in phase. The effect of an additional 1500 point mutations on S4 may also be seen in the lowest graph of Fig. 1d (sequence S6). The amplitude of the variations becomes much smaller and one phase fault is produced in the pattern for YR.YR (a maximum is given at n = 6 instead of n = 7) due to the randomizing effect of these mutations. These rhythms now more closely resemble those found in SV40 and pBR322 (Fig. 1e), which may well have experienced more insertions, deletions and point mutations than $\phi$X174 and fd (pBR322 has also man-made rearrangements).

One possibility for further improvement in the simulation is seen by examining the first 40 amplitudes for $\phi$X174 and S4. These have been averaged to give the initial points in the mean amplitude graphs of Fig. 2a and c (A = 120 for $\phi$X174 and A = 118 for S4). To see more detail, these 40 amplitudes are averaged in 5's and plotted against n as shown in Fig. 2d. The $\phi$X174 values then show a tendency to decrease while those for S4 appear randomly distributed about their mean. Further investigation shows that the only way to obtain such an effect without appreciably altering the initial slope of the (averaged in 40's) amplitude graph in Fig. 2c is to have short 'out of phase' stretches within the genome. The initial slope in Fig. 2c is then little altered for, at the larger separations n, the majority of the correlations span these short stretches and stay the same as before. A particular case is illustrated in Fig. 2d for the sequence S5, which has been formed

from an original RNY strand by applying 2950 Y ⇌ R point mutations plus 4 insertions and 3 deletions allowed to fall randomly anywhere in the genome as before, and, additionally, 10 insertions randomly applied but each followed by a randomly positioned deletion within a distance of not more than 75 bases. The required fall in the mean amplitudes, averaged in 5's can now be obtained at small n values (see Fig. 2d) and yet the long range amplitude graph remains much the same as in Fig. 2c.

Other refinements could be built into the simulations. For example, the Y/R ratio in a given genome (e.g. Y/R is 2841/2545 for $\phi$X174) could be approximated after such a very large number of point mutations if the relative probabilities for Y → R and R → Y mutations during the whole evolutionary time were known and suitable starting proportions of RRY and RYY in the primeval strand taken in the simulation. (The Y → R and R → Y probabilities have been assumed equal above.) The present computations have merely indicated the possible derivation of these rhythmic effects but have avoided for the moment any such further assumptions required to obtain a more detailed matching to a given genome.

Further simulations have also suggested an explanation for the relation of the purine pyrimidine rhythms with those of the four bases T,C,A,G. By randomly interpreting the Y,R triplets in terms of the various possible T,C,A,G combinations, similar T,C,A,G rhythms can be found to those in $\phi$X174. The addition of in and out of phase components (see e.g., Fig. 1g,h) to Y,R rhythms and the fact that some T,C,A,G correlations are particularly strong (see e.g., Fig. 1f) can derive from the types and proportional numbers of the component T,C,A,G triplets in the RNY message.

### Comma-less Code and Phase Prediction

For reasons unconnected with the present findings, an original messenger strand with the codon pattern RRY has been earlier suggested by Crick et al. (1976) and one with the above pattern RNY by Eigen (1978). These two specific sequence proposals derive from considering the sequence regularity in the anti-codon loops of a variety of present tRNA's. Ways in which such messenger strands could have come into existence are discussed in both these papers and a method of protein synthesis envisaged using a primitive tRNA molecule with two possible conformations (Fuller and Hodgson 1967, Woese 1970). A number of advantages for the RNY form are given by Eigen (1978), particularly stressing those arising from the symmetry between the forward and reverse strands and the more even proportions of purines and pyrimidines possible.

Such an original RNY messenger strand provides a (purine-pyrimidine) comma-less code (cf. Crick et al. 1957), in which the sense codons of the form RRY and RYY can only be found in one reading frame. In spite of the great extent of the mutations which have occurred, this pattern of purines and pyrimidines could be essentially the feature from which the present characteristics in period, phase and amplitude derive and also be the basic reason for the importance of the Y, R rhythms. In the original form of such a message, for example:

RRY RYY RYY RRY RYY RRY RRY RYY ...,

counts of certain correlated base combinations are only given at separation intervals of three bases and are zero at all other separations (i.e. the correlation signal amplitude is a maximum and the background is zero). This phenomenon has been modified by mutations, but still a maximum count is given at intervals of three (i.e. amplitude less and background increased). In fact the phases as generally observed in $\phi$X174 and other genomes still fully agree with those to be expected from such an original messenger strand. For example, counts of the correlation YR.YY are seen in the above sequence to be given at n = 0, 3, 6, ... separating bases and to be zero at other n values. The phases for 9 of the 16 possible correlations of Y, R bases, two with two, can thus be immediately written down and agree with those found in $\phi$X174 (e.g. the phases in Fig. 1b). 6 of the remaining combinations represent correlations such as YR.RY, which can be seen in the above strand with two alternative separations (0 and 2 in this case), and RY.RY appears with three separations (0, 1 and 2). For these 7, the phases depend on the relative proportions of the two triplets and on the order in which they occur, as can be shown by computer simulation. Such cases also satisfactorily account for a few phase disagreements which have been noted between some of the above genomes (e.g. YR.RY has phase 0 in fd but phase 2 in $\phi$X174). With Y, R correlations of three with three, 36 of the 64 possible combinations can be found in any such primeval RNY message as above, and each of these has only one set of possible separations. All of these 36 are in fact found with separations which give full agreement with the phases found in $\phi$X174 and elsewhere (e.g. phases 0, 1 and 2 for the correlations in Fig. 1c). Of the remaining 28 combinations, one or other of the triplets (or both) must have arisen in the present genomes by mutations from the original strand. The relative amplitudes of the periodic variations, e.g in $\phi$X174, also support this theory. The 36 correlations of non-mutated triplets are clearly marked and of the remaining 28, those (24) with one mutated triplet are generally weaker or have no rhythm, and all those (4) with two mutated triplets have no rhythm.

Thus the phases of the present rhythms have been successfully predicted from this simple comma-less message. The symmetry of this message in both strands

could also account for the fact that rhythms with correct phases have been seen in SV40 and pBR322, although these genomes code for proteins in both strands.

*Mutation*

The rhythmic effects are by no means only due to the high RNY/RNR ratio in the degenerate codons. It is also found by a count of all the codons in the proteins of $\phi$X174 that RNY = 737, YNY = 591, RNR = 422 and YNR = 314. Similarly in fd, RNY = 750, YNY = 521, RNR = 368 and YNR = 314. Thus the number of the original form RNY is greater than of the once Y $\rightleftharpoons$ R mutated forms RNR or YNR, and the smallest count is given for the twice mutated form YNR. Mutation away from an original RNY strand gives the best explanation yet found to explain such a distribution and give the basic Y, R pattern producing the present rhythms.

The fact that the RNY message is best preserved in the degenerate codons may seem contradictory to observations of high rates of mutation in such codons particularly in the third position (cf. Nichols and Yanofsky 1979; Jukes and King 1979). It is well to remember, however, that the present effects are an indication of how a much larger number of Y $\rightleftharpoons$ R mutations have occurred over the whole era of time since such a primeval message existed, and not of phenomena in more limited periods. The original RNY strand could only code for a maximum of eight amino acids, assuming that the present genetic code applied, and great improvements in protein function must have been achieved in the earlier part of this time. Many amino acids changes would be made and additional amino acids would also come into use when the translation system became more sophisticated and the comma-less code was no longer necessary. It seems likely that in this early period and in subsequent rapid periods of protein change (e.g. in the globins of early vertebrates; see Goodman et al. 1975) changes of amino acids will often be advantageous and be selected for. In times of relative stability (e.g. more recently in globins; ibid.) when the protein's function has been largely optimized for the purpose then required, changes of amino acid will usually reduce functional efficiency and be selected against. During these stabler periods, changes in degenerate codons, which may be neutral or give only a marginal advantage to the organism, have more chance of surviving (e.g. in divergences between the globin genes of later vertebrates; cf. Jukes and King 1979). Thus, from these considerations and from the present findings of good preservation of the RNY message in degenerate codons, mutations in these codons seem to represent a minority of the total point mutations away from the primeval strand, and in the majority of cases a change of amino acids has produced the prototype which has survived (as suggested earlier in this discussion).

With regard to insertions and deletions, the above simulations indicate that, at a considerable distance from each other, the number which has occurred corresponds approximately to the number of changes of reading frame between the individual $\phi$X174 or fd genes. The other interesting indication is that there are short 'out of phase' stretches within these genes. The existence of such stretches can be checked by randomly mixing the codons within the genes amongst each other. The amplitude of the rhythms corresponding to correlations at small distances then decreases, showing that the arrangement of the codons is important and that the effect is not merely due to the proportional numbers of different codon types used. This reduction in amplitude of the correlations is due to the mixing of the 'out of phase' codons from these short stretches amongst the other 'in phase' codons. The present effects of past mutation by insertion and deletion have thus been detected.

**Conclusion**

More recently such Y, R rhythms, generally weaker than in the DNA viruses $\phi$X174, G4 and fd, but having the same characteristics in phase and amplitude, have been found in a variety of other genomes, for example the RNA virus MS2 (Fiers et al. 1976 and references therein), a ribosomal protein gene cluster of *E. coli* (Post et al. 1979), the transposon Tn3 from *E. coli* containing three protein genes (Heffron et al. 1979), the mRNA for chicken ovalbumin (McReynolds et al. 1978) and the sea urchin histone genes (Schaffner et al. 1978). (In contrast, however, no rhythms at all could be detected in 16S and 23S ribosomal RNA genes from *E. coli* (Brosius et al. 1978, 1980), which clearly have no direct connection with protein coding.) Hence the rhythmic phenomena seem to be of a general nature occurring in plasmids, viruses (both DNA and RNA), procaryotes and eucaryotes. Moreover, additional evidence for the comma-less code hypothesis has been obtained by examining all the present genomes (subdivided into sections) in each of the three reading frames and seeing which frame gives the best approximation to an original RNY message. In this way it can be shown that almost all of the proteins in these genomes are being read in their original frames, although some short 'out of frame' stretches are seen within the genes. A good indication of the present extent and reading frame for each protein can thus be obtained purely from purine-pyrimidine information (Shepherd 1980). It is also found that only the combination RRY with RYY of all the possible Y, R comma-less codes will predict the correct reading frame for these many proteins.

Finally, although the periodic correlations appear as a significant effect in a wide variety of genomes and their characteristics, particularly the phases, indicate a common origin as now proposed, many aspects of the evolutionary processes involved must remain obscure until more evidence is available and perhaps the formation and function of such a primitive strand becomes experimentally demonstrable.

*Acknowledgments.* Many thanks are due to Dr. A. Labhardt for valuable help with some of the early computer programming, to Dr. C. Paul for advice with regard to computer technicalities, to Drs. T. Bickle and G. Büldt for helpful discussions, and to Professors W. Arber and G. Schwarz for their generous support. Acknowledgement is also made to the Swiss National Science Foundation.

## References

Beck E, Sommer R, Auerswald EA, Kurz Ch, Zink B, Osterburg G, Schaller H, Sugimoto K, Sugisaki H, Okamoto T, Takanami M (1978) Nucleotide sequence of bacteriophage fd DNA. Nucl Acids Res 5:4495–4503

Brosius J, Palmer ML, Poindexter JK, Noller HF (1978) Complete nucleotide sequence of a 16S ribosomal RNA gene from *Escherichia coli*. Proc Natl Acad Sci USA 75:4801–4805

Brosius J, Dull TJ, Noller HF (1980) Complete nucleotide sequence of a 23S ribosomal RNA gene from *Escherichia coli* Proc Natl Acad Sci USA 77:201-204

Crick FHC, Griffith JS, Orgel LE (1957) Codes without commas. Proc Natl Acad Sci USA 43:416-421

Crick FHC, Brenner S, Klug A, Pieczenik G (1976) A speculation on the origin of protein synthesis. Orig Life 7: 389-397

Eigen M (1978) The hypercycle. A principle of natural selforganisation. Part C: The realistic hypercycle. Naturwiss 65:341-369

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) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260: 500-507

Fiers W, Contreras R, Haegman G, Rogiers R, Van de Voorde A, Van Heuverswyn H, Van Heereweghe J, Volkaert G, Ysebaert M (1978) Complete nucleotide sequence of SV40 DNA. Nature 273:113-120

Fuller W, Hodgson A (1967) Conformation of the anticodon-loop in tRNA. Nature 215:817-821

Godson GN, Barrell BG, Staden R, Fiddes JC (1978) Nucleotide sequence of bacteriophage G4 DNA. Nature 276:236-247

Goodman M, Moore GW, Matsuda G (1975) Darwinian evolution in the genealogy of haemoglobin. Nature 253:603-608

Heffron F, McCarthy BJ, Ohtsubo H, Ohtsubo E (1979) DNA sequence analysis of the transposon Tn3: Three genes and three sites involved in transposition of Tn3. Cell 18:1153-1663

Jukes TH, King JL (1979) Evolutionary nucleotide replacements in DNA. Nature 281:605-606

Linney E, Hayashi M (1974) Intragenic regulation of the synthesis of $\phi$X174 gene A proteins. Nature 249:345-348

McReynolds L, O'Malley BW, Nisbet AD, Fothergill JE, Givol D, Fields S, Robertson M, Brownlee GG (1978) Sequence of chicken ovalbumin mRNA. Nature 273:723-728

Nichols BP, Yanofsky C (1979) Nucleotide sequence of *trpA* of *Salmonella typhimurium* and *Escherichia coli*: An evolutionary comparison. Proc Natl Acad Sci USA 76:5244-5248

Post LE, Strycharz GD, Nomura M, Lewis H, Dennis PP (1979) Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit $\beta$ in *Escherichia coli*. Proc Natl Acad Sci USA 76:1697-1701

Reddy VB, Thimmappaya B, Dhar R, Subramanian KN, Zain BS, Pan J, Ghosh PK, Celma ML, Weissman SM (1978) The genome of Simian Virus 40. Science 200:494-502

Sanger F, Coulson AR, Friedmann T, Air GM, Barrel BG, Brown NL, Fiddes JC, Hutchison CA, Slocombe PM, Smith M (1978) The nucleotide sequence of bacteriophage $\phi$X174. J Mol Biol 125:225-246

Schaffner W, Kunz G, Daetwyler H, Telford J, Smith HO, Birnstiel ML (1978) Genes and spacers of cloned sea urchin histone DNA analyzed by sequencing. Cell 14:655-671

Shepherd JCW (1980) in preparation

Sutcliffe JG (1979) Complete nucleotide sequence of the *Escherichia coli* plasmid pBR322. Cold Spring Harbor Symp Quant Biol 43:77-90

Woese C (1970) Molecular mechanics of translation: a reciprocating ratchet mechanism. Nature 226:817-820

Received December 15, 1979/ Revised September 16, 1980