Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 2437–2441, April 1985
Evolution

Pattern analysis of 5S rRNA

(sequence analysis/phylogeny/dendrogram/triplet pattern/early evolution)

MANFRED EIGEN*, BJÖRN LINDEMANN*, RUTHILD WINKLER-OSWATITSCH*, AND COLIN H. CLARKE†

*Max-Planck-Institut für biophysikalische Chemie, D3400 Göttingen, Federal Republic of Germany; and †School of Biological Sciences, University of East Anglia, Norwich, NR4 7TJ, England

Contributed by Manfred Eigen, November 26, 1984

ABSTRACT **Some 200 different 5S rRNA sequences from eubacteria, chloroplasts, mitochondria, archaebacteria, and eukaryotes were analyzed for evolutionary kinship relationships and associated sequential features. Group-specific occupation schemes for the 149 positions of an overall alignment were established. Eubacterial, archaebacterial, and intermediate occupation schemes all yield a strongly biased base triplet pattern in one of the three possible reading frames strongest for eubacterial, chloroplastic, and archaebacterial, but still detectable for mitochondrial and eukaryotic cytoplasmic sequences. The frequency of triplets decays in the order RNY > RNR > YNY > YNR; R being a purine (guanine or adenine), Y is a pyrimidine (cytosine or uracil), and N is any base. A strong preference for guanine or cytosine was found in all triplet positions. The effects show no exceptions and are clearly above the level of statistical fluctuations.**

In this paper, we report a comparative study of the ≈200 5S rRNA sequences known today. Preliminary analysis of some mainly eubacterial 5S rRNA sequences (1) revealed a clear bias for the presence of a triplet pattern 5′ RNY 3′ where R is a purine (guanine or adenine), Y is a pyrimidine (cytosine or uracil), and N is any nucleotide. A similar phenomenon was found previously for tRNA sequences (2, 3). While the reading frame for tRNAs is defined through the position of the anticodon and the common assignment of the 5′ terminus, 5S rRNA sequences vary in length and therefore had to be tested for the three possible reading frames. For each individual sequence, the RNY bias shows up in only one of the frames, varying with respect to the 5′-terminal position; in the two corresponding alternative frames, a weaker YNR bias always appears. We conjectured and proved with this study that the variable reading frames can be synchronized through proper alignment.

Our analysis is essentially based on data from two sources. Most of the sequences are compiled in an alignment produced by Erdmann *et al.* (4), of which we used only nondegenerate sequences in order to avoid statistical bias. These comprise 115 eukaryotic, 37 eubacterial, 9 chloroplastic, and 1 mitochondrial sequence. In addition, 17 archaebacterial sequences were kindly provided by G. E. Fox, C. R. Woese, and K. R. Luehrsen (personal communication).

All data were filed and processed on a Philips P2000 M computer so as to yield alignment, common reading frames, tree topology, base composition, and periodic patterns.

Alignment and Common Reading Frames

Any comparative analysis of base composition and pattern structures is critically dependent on proper alignment of the sequences. There are sufficient homologies and invariances distributed along the entire sequence that assignment of positions does not pose any serious problem for an overwhelming majority of sequences. Minor uncertainties remain for only a few positions in the mitochondrial and some archaebacterial sequences.

Alignment was greatly aided by the determination of master sequences, which show the nucleotide appearing most frequently for each position in a group of sequences (2). If the genealogy of the group shows bundle-like topology, representing simultaneous or temporally parallel divergence, the master sequence may closely resemble the common ancestor. However, for a tree-like topology of consecutive divergence, as is typical for phylogeny, the master sequence is merely a consensus sequence that need not be identical with the root of the distribution. It is, nevertheless, representative for the degree of homology. The average deviations of individuals from master sequences for 117 corresponding positions, where the master sequences are uniformly occupied, are as follows: eubacteria, 29.5 (10–47), 74.8% homology; archaebacteria, 31.6 (17–44), 73.0% homology; eukaryotes, 25.1 (8–53), 78.6% homology, with the numbers in parentheses indicating the extremes of individual deviations. (The nine chloroplastic sequences analyzed represent an extremely homologous group showing on average <10 deviations.) These data refer to absolute differences of nucleotide composition. They would have to be corrected for positions in double-stranded regions as well as for parallel and reverse mutations to represent evolutionary distances.

The true problem of pattern analysis is not alignment as such, but rather the assignment of those positions that are not uniformly occupied. The master sequences shown in Fig. 1 are stretched into a "procrustean bed" of 149 positions. This is the consequence of a consensus drawn from an alignment of all sequences. The small numbers below certain positions indicate how many individual sequences in the alignment deviate from the consensus. As shown in Fig. 1, deviations from the occupation scheme exhibited by the majority of sequences are minor and restricted to small related groups that diverged from the main branch (e.g., cyanobacteria, prochlorophytes, and chloroplasts associated with the group of eubacteria). Unequivocal decisions can therefore be made for each major group. Any small surplus of occupation then means "insertion," and any small deficiency of occupation means "deletion."

Each master sequence defines a unique "occupation scheme" of the 149 positions. Their mutual insertion/deletion divergences are represented as a dendrogram in Fig. 2, which includes the nine chloroplast and one mitochondrial nondegenerate sequences available. While chloroplast and eubacterial sequences agree wherever eubacteria differ from other groups, chloroplasts show two characteristic features that they share only with cyanobacterial and prochlorophytic sequences. These unique changes indicate that chloroplast evolution started in a branch of prokaryotic precursors. The mitochondrial sequence is quite unique, being responsible for the large internal gaps showing up in the alignment of

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.

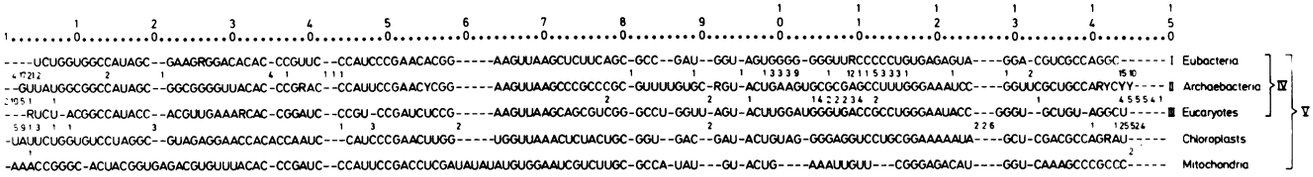

FIG. 1. Alignment of master sequences. Numbers below positions specify minorities of sequences that show occupation at otherwise unoccupied positions (insertion) or no occupation at otherwise occupied positions (deletion). Since, apart from terminal regions, these numbers are small compared to the numbers of sequences analyzed—i.e., 115 eukaryotes, 37 eubacteria, 17 archaebacteria, and 9 chloroplasts (but only 1 mitochondrion)—an unequivocal assignment of insertions or deletions can be made relative to the bulk of sequences in the group. The master sequences thus obtained are read as uninterrupted sequences that define the occupation schemes I, II, and III. The two other occupation schemes are the consensus of master sequences I, II, and III (scheme IV) and of all five sequences (scheme V).

the three main groups (i.e., insertions at positions 20, 21, 59, 60, 61, 62, and 63, which are balanced by deletions at positions 11, 92, 100, 101, 102, 112, 113, and 114). Otherwise, they are closest to the eubacterial occupation scheme. Only at position 85 does the mitochondrial occupation scheme agree with archaebacteria and eukaryotes rather than with eubacteria.

If we compare the occupation schemes of the three main groups, we identify only 5 interior differences between eubacteria and archaebacteria. At two of these positions (86 and 132), eubacteria agree with eukaryotes, while at three positions (85, 90, and 103), archaebacteria agree with eukaryotes.

Eukaryotes, on the other hand, have departed farthest from any common node (ignoring mitochondria). Their uniform appearance as a group must thus be due to some bias that appeared after their branching off from both eubacteria and archaebacteria, requiring a change of structure mandatory for all eukaryotes.

Occupation schemes I–V, as defined in Figs. 1 and 2, were used to search for sequential patterns. Schemes IV and V are consensus schemes of the three main group master sequences I–III and of all master sequences, respectively.

In principle, any of the five schemes or any of their intermediates could represent an ancestral sequence. However, it is very unlikely that this is true for the eukaryotic scheme III. The smallest divergence from nodes was found for eubacteria. It is likely for the root to be close to this scheme. Note that its relative position is characteristic of this type of

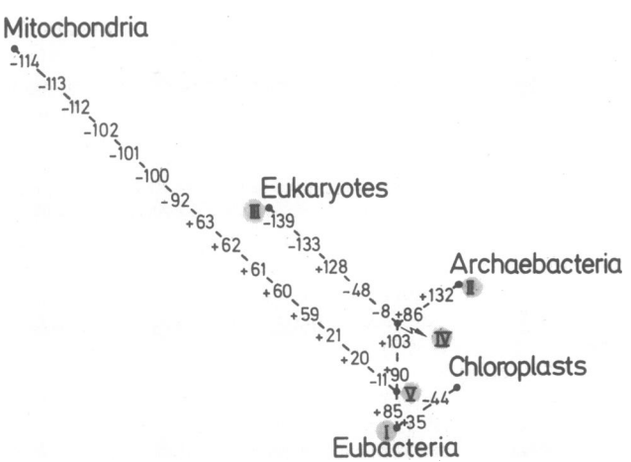

FIG. 2. Dendrogram of occupation schemes. Aligned master sequences (cf. Fig. 1) are compared with respect to occupation or nonoccupation of positions. Each number refers to a position where plus means occupation at an otherwise nonoccupied position (insertion) and minus means nonoccupation at an otherwise occupied position (deletion). Assignment of insertion or deletion here is relative to a preceding node. As reference (earliest node), the eubacterial master sequence was chosen. Roman numerals correspond to occupation schemes defined in Fig. 1.

dendrogram and not necessarily identical with that in the sequence dendrograms considered below.

In concluding this discussion on alignment and occupation we stress the following: (*i*) the main groups have quite unique occupation schemes; (*ii*) according to these schemes archaebacteria are more closely related to eubacteria, but are otherwise intermediate between eubacteria and eukaryotes; (*iii*) uncertainties about the most likely common precursor are restricted to three nucleotides between positions 85 and 103 (scheme IV); (*iv*) chloroplasts and mitochondria support the eubacterial scheme, although they have features characteristic of themselves; and (*v*) the interior of the eukaryotic scheme differs most from all others, probably because of structure-sensitive insertions and deletions that must have occurred after eukaryotes diverged from the other groups.

Patterns and Base Composition

A search for periodic patterns may start from the master sequences in which biases are more pronounced. The eubacterial and archaebacterial master sequences in Fig. 1 reveal even on visual inspection an RNY pattern, while nontriplet patterns cannot be found to any significant extent.

The evaluation proceeds as follows: All sequences are expressed as continuous sequences in RY form, the gaps being closed according to one of the five occupation schemes. All consecutive triplets (i.e., RNY, RNR, YNY, or YNR) are then identified and their frequencies of appearance are recorded. This procedure is carried out for the three possible triplet reading frames. For the bulk of sequences, uniform occupation is found only between positions 5 and 143 inclusive; hence, we start throughout at positions 5, 6, and 7, respectively, and count triplets for each occupation scheme up to position 143. This procedure is not critically dependent on the choice of both terminal positions as long as it is carried through consistently.

In an unbiased distribution of 39 triplets among four classes, one would expect each class to be represented on average 9.75 times. The data in Table 1 referring to the eubacterial occupation scheme as applied to all groups shows triplet frequencies that are obviously biased—namely, in reading frame 1 strongly for RNY over YNR and weakly for RNR over YNY. For eubacteria, the mean ratios RNY/YNR and RNR/YNY are 14.3/6.1 and 10.2/7.3, and for the master sequence 17/3 and 11/8. Only the half sums (RNY + YNR)/2 and (RNR + YNY)/2 appear balanced, the values being 10.2 and 8.75 for the averages and 10.0 and 9.5 for the master values. Hence, in frame 1 there is a clear order of frequencies RNY > RNR > YNY > YNR, which is strongest for eubacteria, median for archaebacteria and chloroplasts (and mitochondria), and weakest for eukaryotes. On the other hand, the histograms in Fig. 3 representing the complete frequency distributions for all groups clearly demonstrate that in all cases (including the eukaryotes), we are dealing with a

Table 1. Pattern distribution for 37 eubacteria, 17 archaebacteria, 115 eukaryotes, and 9 chloroplasts in the three alternative reading frames of occupation scheme I (cf. Figs. 1 and 2)

	Frame 1				Frame 2				Frame 3			
Sequence	RNY	RNR	YNY	YNR	RNY	RNR	YNY	YNR	RNY	RNR	YNY	YNR
						Eubacteria						
Mean	14.3	10.2	7.3	6.1	5.6	14.1	7.0	10.5	8.3	8.1	9.2	12.1
Master	17	11	8	3	2	17	8	10	7	7	10	13
						Archaebacteria						
Mean	13.4	9.5	8.1	7.0	5.5	12.6	9.4	10.4	9.6	7.1	10.6	10.5
Master	15.5*	9	7.5*	7	6	12	9	11	10	6	11	12
						Eukaryotes						
Mean	12.6	9.0	6.1	9.2	6.4	11.5	7.7	10.1	9.0	7.3	8.7	10.5
Master	12	9	6	10	8	10	7	11	10	7	7	11
						Chloroplasts						
Mean	13.9	10.3	5.7	8.3	7.0	11.3	7.6	11.4	7.4	10.2	10.6	8.0
Master	14	11	6	8	7	11	8	11	7	10	11	8

*One of these triplets in the master sequence refers to equally frequent RNY and YNY appearance.

well established experimental fact rather than some arbitrary fluctuation.

Prevalence of RNY in one reading frame implies prevalence of YNR in a different one. The following scheme demonstrates the order to be expected for the different reading frames (Fr) if the main frame is biased by one of the four triplets:

Fr	RNY bias	Fr	RNR bias
1	RNY>RNR≈YNY>YNR	1	RNR>RNY≈YNR>YNY
2	RNR+YNR>RNY+YNY	2	RNR+YNR>RNY+YNY
3	YNR+YNY>RNR+RNY	3	RNR+RNY>YNR+YNY

Fr	YNR bias	Fr	YNY bias
1	YNR>RNR≈YNY>RNY	1	YNY>RNY≈YNR>RNR
2	RNY+YNY>RNR+YNR	2	RNY+YNY>RNR+YNR
3	RNY+RNR>YNR+YNY	3	YNR+YNY>RNR+RNY

The data of Table 1 yield the best fit for RNY and some additional RNR bias in the main reading frame. The relatively high RNR frequencies in frame 2 are quite in agreement with this conclusion. The alternative interpretation, that frame 2 is an RNR biased main frame, creates inconsistencies with the order of appearance of other triplets suggested by the above scheme.

A bias, of course, should be rated against some control, for which the unbiased middle position may qualify. Table 2 shows the results for eubacteria, archaebacteria, and eukaryotes. The table lists, for each position in the triplets, the frequency of occurrence for the sums (A + G) = R, (U + C) = Y, (A + U), and (G + C). Without exception the table shows the following: (*i*) The sequences are rich in (G + C). At all three positions, (G + C) is in excess over (A + U), the average (G + C)/(A + U) ratio being 1.61. (*ii*) The first position is dominated by R in all sequences as R/Y = 1.73 and in eubacteria as R/Y = 2.55. (*iii*) The third position is dominated by Y, in all sequences as Y/R = 1.34 and in eubacteria as Y/R = 1.60. (*iv*) The middle position is unbiased as to R/Y, the average for all sequences being 1.01. These results support the conclusions reached previously.

Next, we repeat the procedure of pattern analysis for all five occupation schemes defined in Figs. 1 and 2. Since master sequences were shown to be representative for the appearance of codon patterns, we demonstrate in Table 3 the results obtained for master sequences only. The data refer to the reading frame that shows the strongest bias. We realize (*i*) that an RNY bias is found throughout with only minor exceptions, (*ii*) that it is strongest for eubacteria as a group and generally for the occupation scheme that refers to the eubacterial master sequence, (*iii*) that it is weakest (or absent) for eukaryotes as a group and for the occupation scheme that refers to the eukaryotic master sequence, and (*iv*) that a slight preference is found for application of an occupation scheme that refers to the very group from which it was derived. The qualitative correspondence of the results does not come as a surprise, because most schemes are closely related to one another. Schemes I, II, IV, and V are identical up to position 85 and schemes I and V are out of phase for triplets only between positions 85 and 104. Only scheme III, showing a weak or no bias, is out of phase with the other schemes for the larger part of the sequences. If the bias were the leftover of an ancient pattern, it should refer to one occupation scheme—namely, the one that most closely resembles the ancient sequence. However, insertions or deletions occurring in the course of evolution may have blurred that scheme and led to readaptations in some base-paired regions of the molecule, favoring some more individual occupation scheme. It is interesting to note that for eubacteria and archaebacteria, the strongest RNY over YNR bias appears prior to position 98—i.e., in a region where both occupation schemes largely agree.

Sequence Topologies

So far, we have not introduced any temporal argument. The only experimental clue that may be interpreted as a temporal argument is the fact that deviations from equipartition are

Table 2. Occupation at the three triplet positions in master sequences analyzed according to the first reading frame of occupation scheme I (cf. Table 1)

	Eubacteria			Archaebacteria			Eukaryotes*		
	1	2	3	1	2	3	1	2	3
R	28	20	14	24.5†	19	16	21	19	19
Y	11	19	25	14.5	20	23	17	18	19
A+U	15	15	15	12.5†	15	15.5†	17.5†	15	13
G+C	24	24	24	26.5	24	23.5	20.5	22	25

*The numbers R + Y or A + U + G + C for eukaryotes do not add up to 39 because, according to occupation scheme I, some positions remain open.

†One triplet in master sequence refers to equally frequent appearance of two nucleotides.

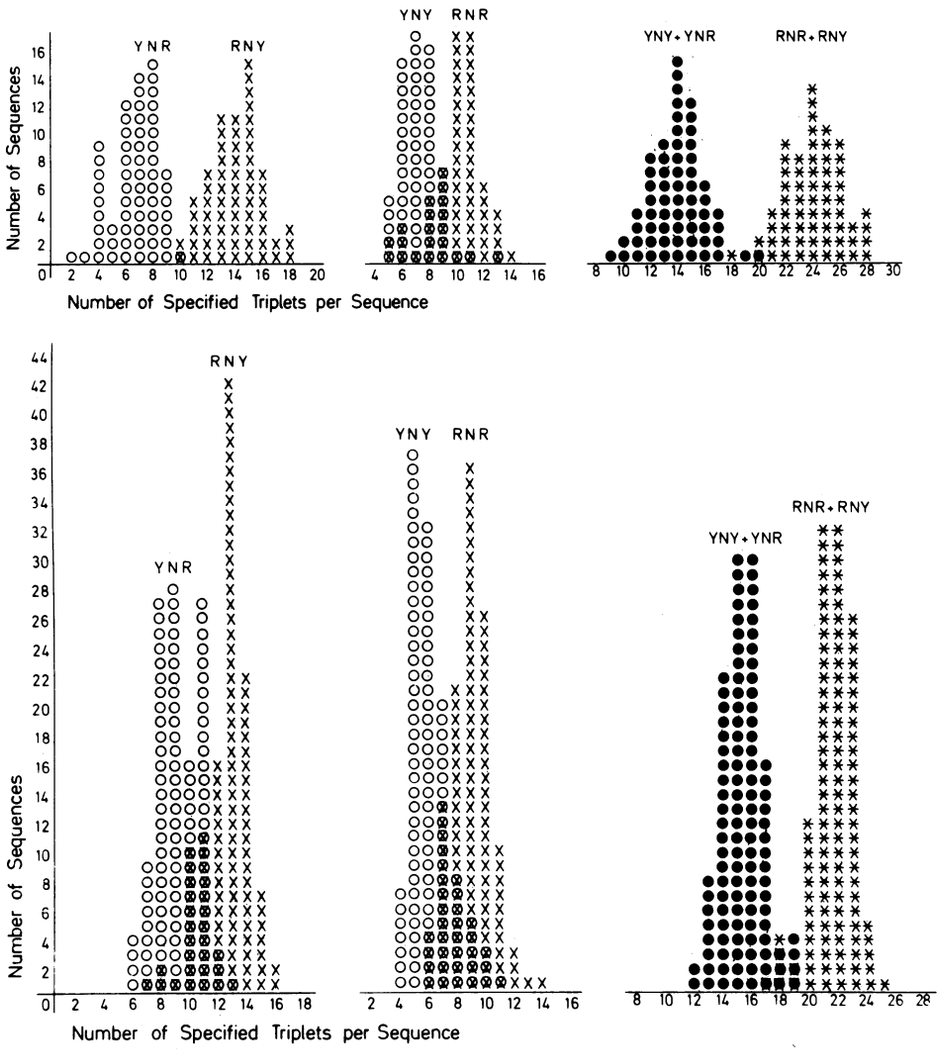

FIG. 3. Histograms for frequency distributions of triplets in reading frame 1. (*Upper*) 37 eubacteria, 17 archaebacteria, and 9 chloroplasts. (*Lower*) 115 eukaryotes.

more pronounced in master sequences than in averages, a phenomenon typical for biased behavior (3). Bias itself may point either to the past or to the future. It may mean reverberation of some ancient structure or evolutionary pressure—i.e., a common goal to be reached through adaptive changes. The question could be decided by the reconstruction of early nodal sequences. Phylogenetic trees for various ensembles of 5S rRNA have been reported in the literature (6–9). Dendrograms based on mutation distances, however, yield only relative branching orders, in which the earliest node remains uncertain. A method based on sequence rather than mere distance space that is especially suitable for testing the topology of branching and reconstructing nodal sequences has been developed (together with A. Dress); its application to 5S rRNA will be presented elsewhere. Here we shall use a related more qualitative procedure that allows a comparison of present and early (i.e., near-nodal) sequences.

We look for ensembles of four sequences that are so little related that their nodes of divergence reflect early periods of evolution. Consider four sequences of two mutually unrelated couples of eubacterial and archaebacterial species (8), which certainly diverged long ago (2). In our example, we use the Gram-positive eubacterium *Bacillus pasteurii*, the cyanobacterium, or bluegreen alga *Synechococcus a.n.*, previously called *Anacystis nidulans*, and the two archaebacteria *Methanococcus vannielli* and *Halobacterium salinarium*. If the branching nodes of all four sequences can be localized in sequence space and shown to be close to each other, the common root of eubacteria and archaebacteria should be not too far away. One then should be able to determine whether the pattern bias is stronger near the common

Table 3. RNY/YNR ratios according to different occupation schemes specified in Figs. 1 and 2

Group	Scheme				
	I	II	III	IV	V
Eubacteria	17/3	13 + 2/5 + 1	8 + 2/6 + 3	13/4	13/7
Archaebacteria	15.5/7	14/6	10/8 + 1	13/8	14/5
Eukaryotes	12 + 2/10 + 1*	12 + 4/9 + 1	12/9	9 + 2/12	9 + 4/10
Chloroplasts	14/8	12 + 1/6 + 2	8 + 3/7 + 1	11/11 + 2	11/6 + 1
Mitochondria	12 + 3/7 + 1	15 + 3/6 + 3	7 + 2/6 + 4	11 + 1/8 + 2	12/5 + 4

*Certain positions remain open if an occupation scheme is applied to a group to which it is not inherent (such as eubacterial scheme I applied to eukaryotes). The small additional numbers then refer to the most likely adjustment of these open positions based on homology with other master sequences.

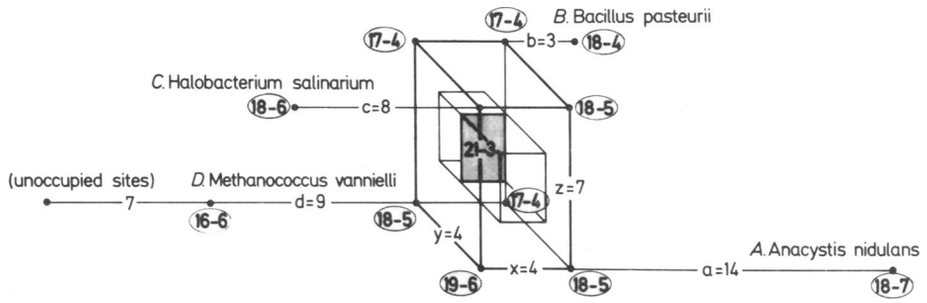

FIG. 4. Dendrogram of four individual sequences in sequence space. Four binary (i.e., RY) sequences A, B, C, and D specify seven types of positions, which after summing up define the seven distances *a, b, c, d, x, y,* and *z* geometrically represented in this diagram. The four peripheral distances *a, b, c,* and *d*, each of which refers to one of the four sequences, specify the number of those positions in which the particular sequence deviates from the three others, which themselves are homologous. The three distances *x, y,* and *z* (defining the dimensions of the box) specify the number of those positions at which two sequences (localized at one of the three planes) mutually agree but differ from the two other sequences (localized at the opposite plane, separated by the third box dimension). A detailed mathematical description of sequence space analysis will appear in another paper (unpublished data). Numbers refer to RNY–YNR differences at the particular position. Dark plane in the midst of the box refers to the highest RNY/YNR ratio of possible precursors. Box represents a total of 2^{x+y+z} sequences.

branching point, representing an ancient sequence, or at the periphery of the diverged bundle—i.e., in the present sequences.

In Fig. 4, an example of this analysis is shown. Four sequences in RY notation are aligned and analyzed according to the following criteria: (*i*) Which positions are identical in all four sequences? (*ii*) At which positions does one sequence deviate from the three others, which themselves are uniformly occupied? Counting such positions specifies the four peripheral distances *a, b, c,* and *d* in Fig. 4. (*iii*) At which positions do two of the four sequences mutually agree? There are three such situations—namely, for sequences designated A, B, C, and D: A = B ≠ C = D, A = C ≠ B = D, and A = D ≠ B = C. Counting such positions defines the three dimensions of the box *x, y,* and *z*.

The seven distances in the geometrical representation of Fig. 4 provide a complete characterization of the relative positions of the four sequences and their intermediates in sequence space, allowing an exact assignment to nodal positions. The pattern distributions at the nodal points show an increase of the difference of frequencies RNY–YNR toward the center of the box. This difference is a direct measure of the distance to a sequence with a completely randomized RNY pattern. Inside the box, there is a plane representing 12 sequences (differing only in the assignment of the middle positions of triplets) that comprise 21 RNY versus 3 YNR triplets. We have intentionally chosen 4 sequences that show large differences RNY–YNR as compared to the averages presented in Table 1. If the large bias refers to a present requirement or convergent evolution, then it should, on the whole, decrease toward the center of the box.

Discussion

For a discussion of the biological significance of the results it is important to distinguish facts from conjectures. The facts are as follows:

(*i*) All sequences studied clearly reveal a significant bias of RNY patterns. These patterns show up—with differing strengths—in one reading frame of all occupation schemes and are balanced by weaker YNR biases in the two alternative reading frames.

(*ii*) The bias is strongest for those sequences that appear in the phylogenetic analysis to be more conservative (2).

(*iii*) In all cases, the frequency of appearance decays in the order RNY > RNR > YNY > YNR. In a very few cases, the YNY/YNR order can no longer be distinguished because the bias has decayed to the limits of randomization.

(*iv*) The infrequent appearance of YNR triplets in the main reading frame is equivalent to the absence of any of the three stop codons in this frame for nearly all sequences inspected. (The three stop codons in the genetic code are UAA, UAG, and UGA.)

(*v*) In all triplets of the main reading frame, the middle position is balanced with respect to R and Y—i.e., 50 ± 5%/50 ± 5%.

(*vi*) (G + C) appears more frequently than (A + U), the percentage ratio being on average 62/38. This bias again is slightly larger in eubacterial and archaebacterial than in eukaryotic, chloroplastic, and mitochondrial sequences (the latter being 54/46).

(*vii*) The features reported are stronger near the roots than at the tips of the branches of the phylogenetic tree.

Our conjecture is that we are dealing with an ancient phenomenon that still reverberates in present day structures.

The order of triplet frequencies RNY > RNR > YNY > YNR according to Shepherd (5, 10) is also a general attribute of coding sequences. It may reflect the evolution of the genetic code from an RNY structure, providing a comma-free readout via wobble-intermediates to the present form. The reflection of this order in the data, more than the mere existence of a triplet pattern as such, led us to believe that tRNA and 5S rRNA have descended from coding sequences used at the time the genetic code originated.

We want to thank Drs. G. E. Fox, C. R. Woese, and V. A. Erdmann for their kind support through making available to us unpublished data and sequence compilations. We also acknowledge very stimulating discussions about sequence topology with Dr. A. Dress. Dr. W. C. Gardiner kindly read and critically reviewed the manuscript. We thank him for most useful comments and suggestions.

1. Clarke, C. H. (1984) Evolution of Prokoryotes, *Abstr. FEBS Symp.*, Munich.
2. Eigen, M. & Winkler-Oswatitsch, R. (1981) *Naturwissenschaften* **68,** 217–228.
3. Eigen, M. & Winkler-Oswatitsch, R. (1981) *Naturwissenschaften* **68,** 282–292.
4. Erdmann, V. A., Wolters, J., Huysmans, E., Vandenberghe, A. & deWachter, R. (1984) *Nucleic Acids Res.* **12,** r133–r161.
5. Shepherd, J. C. W. (1981) *J. Mol. Evol.* **17,** 94–102.
6. Fox, G. E., Luehrsen, K. R. & Woese, C. R. (1982) *Zentralbl. Bakteriol. Parasitenk. Infektionskr. Hyg. Abt. 1 Orig.* **C3,** 330–345.
7. Kuentzel, H., Heidrich, M. & Piechulla, B. (1981) *Nucleic Acids Res.* **9,** 1451–1461.
8. Hori, H., Itoh, T. & Osawa, S. (1982) *Zentralbl. Bakteriol. Parasitenk. Infektionskr. Hyg. Abt. 1 Orig.* **C3,** 18–30.
9. Klotz, L. C., Komar, N., Blanken, R. L. & Mitchell, R. M. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 4516–4520.
10. Shepherd, J. C. W. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 1596–1600.