J. molec. Evolution 2, 277—292 (1973)

Doublet Frequency Analysis of Bacterial DNAs

G. J. Russell*, D. J. McGeoch, R. A. Elton* and J. H. Subak-Sharpe

Institute of Virology, University of Glasgow, Glasgow

Received June 5, 1973

Summary. Nearest neighbour base frequency analyses of the DNAs of fifteen bacteria and two blue-green algae are reported. When expressed in terms of deviations from random expectation, the frequencies can be placed in four distinct groups sharing similarities not dependent on the G+C contents of the DNAs. The majority of the groupings found are in agreement with those of conventional taxonomy but several interesting discrepancies are shown to exist, some of which confirm other recent molecular evidence. The frequencies for the algal DNAs closely resemble those of the largest group of bacteria. The results are considered in relation to possible evolutionary pressures on polypeptide-specifying DNA and inferences are made about the relative usage of alternative codons in different species.

Key words: Doublet Analysis — Bacterial DNA.

1. Introduction

The degree of relatedness of bacteria has been studied by several criteria. Primarily, bacteria have been classified by their morphological characteristics which represent "the combined expression of many genetic loci" (Marmur, Falkow, and Mandel, 1963; Mandel, 1969). This is taken to a high order by the Adansonian method ("numerical taxonomy") in which coefficients of similarity are established using a large number of available phenotypic expressions. Other criteria which have been used include transfer of genetic material, comparison of biochemical pathways and of the properties of informational macromolecules. The deduced relationships have generally been found to be in broad agreement with conventional taxonomy.

In the study of informational macromolecules, Lee, Wahl, and Barbu suggested, as early as 1956, that DNA base compositions might be an important taxonomic aid. Since then, it has become clear that organisms exhibiting homology by genetic criteria generally have base composition similarity. The converse is clearly not true. Moreover as the DNAs of all analysed bacteria fall within the base composition limits of 25–75% G+C content, this criterion provides only limited information. Variation in base composition has also been related to change in phenotype within a genus

* Member of the scientific staff of the Medical Research Council Virology Unit.

e.g. change in flagellation as a function of base composition in the Rhizobia (De Ley and Rassel, 1965).

The feasibility of utilising the DNA to investigate the evolutionary relationships of bacteria was greatly increased by the introduction of the technique of nearest neighbour analysis (Josse, Kaiser, and Kornberg, 1961; Swartz, Trautner, and Kornberg, 1962; Hurwitz *et al.*, 1961) with the demonstration that the frequencies of the sixteen possible dinucleotides or "doublets" in a natural DNA are characteristic and non-random. The six bacteria studied above showed considerable differences in doublet frequencies which appeared to be generally related to base composition. The possibility that doublet frequency relationships have genetic and evolutionary significance was suggested by Kaiser (1962) and Freese (1962), but only one other such study on bacteria has been reported (Skalka, Fowler, and Hurwitz, 1966). Strong support for this view has come from doublet frequency studies on viral and viral host DNAs in animals, plants and bacteria (Subak-Sharpe *et al.*, 1966; Subak-Sharpe, 1969; McGeoch, Crawford, and Follett, 1970; Russell, Follett, Subak-Sharpe, and Harrison, 1971).

Subak-Sharpe *et al.* (1966) described a method of normalization which allowed comparison of the doublet frequencies of DNAs irrespective of their base composition. Their patterns of deviation from random expectation (general designs) were compared in order to determine if there was any underlying similarity not directly dependent on the precise base composition. The rationale was that related DNAs would have similar general designs. When this procedure was applied to the data of Josse *et al.* (1961), the six bacteria studied fell into two distinct groups (Subak-Sharpe *et al.*, 1966) and the data of Skalka *et al.* (1966) gave rise to a third group.

The present work was undertaken to extend these investigations over a much wider range of bacteria so as to obtain some idea of the extent and number of characteristic general design groups and also of their relationship to the existing bacterial classification. A special effort was made to include bacterial species having DNAs with extreme base composition because of the possible doublet frequency limitations imposed on the genetic information in such DNAs. This aspect is considered more closely in the following paper (Elton, 1973). Also included here are the analyses on two blue-green algae which are not distinguished from bacteria by definition of the characteristics of bacteria (Stanier and Van Niel, 1962).

2. Materials and Methods

a) Enzymes and Chemicals

DNA Polymerase. DNA dependent DNA polymerase was isolated and purified from *Escherichia coli* (strain B) by the method described by Richardson *et al.* (1964). Steps 5 and 6 were omitted and the purification terminated after fractionation on

DEAE-cellulose (step 7). The purified enzyme was dialysed against 0.05 M tris + 0.01 M β-mercaptoethanol, pH 7.5, and stored on ice.

DNA polymerase was also obtained from the Boehringer Corporation (London) Ltd. (DNA polymerase; Grade I from *E. coli*).

Pancreatic DNase I was obtained from the Worthington Biochemical Corp., Freehold, N. J., and the Sigma Chemical Co., London.

Micrococcal nuclease was obtained from Worthington and was dissolved at 15000 units/ml H_2O and heated at 100° for 1 min. This solution was stored at —20°.

Spleen phosphodiesterase was obtained from Sigma as a lyophilised powder. This was dissolved at 20 units/ml H_2O to give a specific activity of approximately 5 units/mg of protein. This solution denatured with freezing and thawing and was therefore kept at 4° for a maximum of one week.

Spleen phosphodiesterase was also obtained from Boehringer with an estimated specific activity of 100 units/mg of protein relative to the Sigma product. This was supplied in an ammonium sulphate suspension and was dialysed against 0.05 M tris, pH 7.5, before use.

Pronase (Grade B) was obtained from Calbiochem Ltd., London, and was preincubated for 2 hours at 37° at a concentration of 10 mg/ml H_2O before use.

Caesium chloride was obtained from two sources.

1. Hopkins and Williams, Chadwell Heath, Essex. A saturated solution (20°) of this product contained a suspension of insoluble material. Filtered saturated solutions were found to have an absorbance of 0.06–0.12 odu/ml at 260 nm. This compound was used for large scale preparative buoyant density gradients.

2. Harshaw Chemical Co., Solon, Ohio. A saturated solution of this product (optical grade) had an absorbance of 0.005–0.030 odu/ml at 260 nm. This was used for all analytical and final small scale preparative buoyant density gradients.

Deoxynucleoside-5'-triphosphates. Non-radioactive deoxynucleoside-5'-triphosphates were obtained from Sigma. Deoxynucleoside-5'-triphosphates, labelled with ^{32}P in the α-phosphate, were obtained from the International Chemical and Nuclear (ICN) Corp., Irvine, California. The quality of these products was very variable and batches showing more than 10% contamination were purified by paper chromatography using the isobutyrate solvent described later.

b) Organisms and DNAs

We gratefully acknowledge the receipt of various preparations of bacteria and other organisms from the following sources.

Rodent mycoplasma (*M. pulmonis:* strain Kon) from Miss Kathryn Forshaw, Department of Infectious Diseases, Glasgow University.

Staphylococcus aureus, Photobacterium phosphoreum, Streptococcus pneumoniae, Rhodopseudomonas capsulata, Rhodopseudomonas sphaeroides, and *Pseudomonas aeruginosa* from Professor A. Wardlaw, Department of Microbiology, Glasgow University.

Anabaena variabilis and *Chlorogloea fritschii* from Dr. N. Carr, Department of Biochemistry, University of Liverpool.

Bacterium NCIB 8250 ("Vibrio 01") from Dr. C. Fewson, Department of Biochemistry, Glasgow University.

We would also like to express our thanks to Dr. D. Söll, Department of Molecular Biophysics, the Josiah Willard Gibbs Research Laboratories, Yale University, for a preparation of *Kid mycoplasma* DNA and to Dr. M. Leng, Centre de Biophysique Moleculaire, Orleans, for a preparation of *Sarcina lutea* DNA.

Professor H. M. Keir kindly prepared DNA from *Bacillus megaterium, Proteus vulgaris, Rhodospirillium rubrum* and *Serratia marcescens* which had been specially grown by the late Professor I. Lominski of Glasgow University, Department of Microbiology.

DNA from *Clostridium perfringens* was obtained commerically from Sigma.

c) Preparation of DNA

Unless described separately below, DNA was isolated in the following manner:

Wet-packed cells were suspended in 5 volumes of 0.15 M NaCl + 0.1 M EDTA, pH 8.0 (if necessary, suspension was aided by a few passes in a Dounce-type homogeniser). The suspension was made approximately 1 % with respect to sodium dodecyl sulphate (SDS) and then heated at 60° for 10 min. After cooling, the solution was diluted with an equal volume of H_2O, pronase added to a concentration of 100 μg/ml and after 2 hours' incubation at 37° the solution was dialysed overnight at 4° against 0.1 × SSC (0.15 M NaCl, 0.015 M Na Citrate).

The resulting solutions containing up to 100 μg/ml of DNA were adjusted to density 1.7 g/cm^3 by addition of solid CsCl. Solution densities were determined by refractive index measurements (Schildkraut, Marmur and Doty, 1962). 20–25 ml aliquots were then centrifuged at 150000 × g (M.S.E. 8 × 35 ml titanium angle rotor) for 20 hours at 20° and 0.5 ml fractions collected. The fractions containing the DNA band were located by measuring absorbance at 260 nm. It must be stressed that at these conditions of centrifugation and DNA concentration equilibrium is not attained. When necessary, a second cycle of CsCl buoyant density centrifugation, now to equilibrium, was used with DNA concentrations of 20–30 μg/ml and centrifugation at 100000 × g for 60–70 hours at 20°. All DNA samples were finally dialysed extensively against 0.05 M tris + 0.005 M EDTA, pH 7.5.

Rodent mycoplasma DNA. Difficulty was encountered in releasing DNA from mycoplasma. Initial attempts to isolated mycoplasma DNA by the method described above resulted in very poor yields ($<$ 100 μg DNA/g dried cells). A more satisfactory method of lysis was that of Rottem *et al.* (1968).

Dried mycoplasma cells (50 mg) were suspended in 1 ml 2 M glycerol and incubated for 10 min at 37°. This suspension was then injected into 10 ml distilled H_2O at 0° through a No. 1 needle, made 0.15 M with respect to NaCl, and after treatment with SDS at 60° the purification was continued as already described. This gave a higher yield of DNA ($>$ 100 μg DNA/100 mg dried cells), but it appeared to be very prone to degradation possibly due to a nuclease closely associated with the DNA (D. Söll, personal communication).

Staphylococcus aureus: Lysis of this organism was achieved by sonication followed by treatment with lyzozyme at 500 μg/ml for 1 hour at 37°. Thereafter DNA isolation was as already described.

d) Ultracentrifugal Analysis

Buoyant densities in CsCl were determined for all DNA samples using the method of Schildkraut, Marmur, and Doty (1962). A Beckman Model E analytical ultracentrifuge was used with 0.5–2.0 μg of test DNA and 0.5–1.0 μg of marker DNA. Herpes virus type 1 DNA was used as a dense marker taking the density as 1.7254 g/cm^3 (Halliburton, Hill, and Russell, 1971). *Clostridium perfringens* DNA, with a density of 1.6915 g/cm^3 (Szybalski, 1968), served as light marker.

Sedimentation velocity experiments were by the band sedimentation method (Vinograd and Brunner, 1966) using 0.5–2.0 μg DNA in 0.01–0.05 ml of low salt concentration layered on to 1.0 M NaCl.

e) Doublet Frequency Analysis

The technique of nearest neighbour or doublet frequency analysis has been described in detail (Josse *et al.*, 1961; Swartz *et al.*, 1962; Josse and Swartz, 1963).

Our experimental procedure was essentially identical to Josse *et al.*'s technique with some minor modifications which we have already reported (Subak-Sharpe *et al.*, 1966; McGeoch *et al.*, 1970; Russell *et al.*, 1971).

Before analysis, each DNA was "activated" by the limited action of pancreatic DNAse and heating to 80° as described by Aposhian and Kornberg (1962). Initially separation of the deoxynucleoside-3'-monophosphates was by high voltage electrophoresis (McGeoch, Crawford, and Follett, 1970). In most cases, however, separation was achieved by descending paper chromatography with isobutyrate solvent on Whatman No. 1 paper for 30–36 hours. For optimum resolution of the 3'-monophosphates, isobutyrate solvent at pH 4.25 was used. This was made by mixing 66 volumes of isobutyric acid with 30 of H_2O, then adjusting the pH to 4.25 with concentrated NH_4OH (sp. gr. 0.880) and making up with H_2O to a final volume of 100. The order of increasing mobility is dGMP, dTMP, dCMP, and dAMP.

f) Normalization of Doublet Frequencies

When comparing DNAs of differing overall G+C content, by definition there will be differences in the doublet frequencies. For example, those doublets containing only G and C will tend to occur with a high frequency in a DNA of high G+C content and with a low frequency in a DNA of low G+C content, the reverse being true for those doublets containing only A and T. This expectation makes comparison of the data difficult to interpret. If the effect of G+C differences *per se* could be removed from the data, then one could compare directly the extent to which the doublet frequencies of the DNAs differ from their random expectations. The rationale is that the 16 doublet frequencies from DNAs which are related will show similar patterns of deviation from random expectation.

This can be achieved by determining the ratio of the observed frequencies to the random expected frequencies. The random expected frequency for a particular doublet is the product of the frequencies of the two bases comprising that doublet i.e. the product of the base composition factors gained from nearest neighbour frequency analysis. In a completely random DNA the ratio for every doublet would be one. The ratios can therefore be plotted in histogram form as deviations from unity. This gives a diagram which illustrates deviation from random expectation irrespective of the G+C content of the DNA. Our presentation is similar to that described by Subak-Sharpe *et al.* (1966), where the normalization was carried a step further by multiplying the ratios by the expected random frequency at 50% G+C (i.e. 1/16 or 0.0625). This gives a set of values with 50% G+C content, equivalent to the original frequencies in deviation from random expectation, which are plotted in histogram form as deviations from the random expectation (62.5 parts/1000 at 50% G+C content).

Our new and the previous format therefore differ only in the values shown in the ordinate so that patterns in the two formats can be compared directly. In our new format a ratio for a given doublet, say 0.5, means that this doublet occurs with a frequency 0.5 times that expected in a random DNA of that particular G+C content. In the previous format this value would be −31.25 doublets/1000 showing the relative shortage of this doublet compared to the random expectation of 62.5/1000.

We consider that this simpler presentation is more easily appreciated and as the end product is identical the previously established term "general design" (Subak-Sharpe, 1967) will be retained, redefined as "the set of deviations from unity of the ratios of the observed doublet frequencies to their expected random frequencies".

g) Cluster Analysis

Cluster analysis of the general designs was carried out using the unweighted average linkage method (Sokal and Sneath, 1963), with Euclidean squared distance (that is, the sum of squares of differences of ratios of observed to random frequencies between two DNAs) as the measure of distance. Essentially the same classification of the DNAs was achieved when Euclidean absolute distances or weighted clustering methods were used instead.

3. Results

a) Buoyant Density Determinations

All of the DNAs investigated gave a single symmetrical peak and the buoyant density values for each are shown in Table 1, together with the G+C contents derived from these values by the method of Schildkraut, Marmur, and Doty (1962). Also shown where available are the G+C contents previously reported in the literature (Hill, 1966; Shapiro, 1968).

With two exceptions, all bands were fairly narrow suggesting that the native DNA molecules had relatively high molecular weights. Both mycoplasma DNAs gave broadly dispersed bands. The symmetry of these bands suggested that this was due to low molecular weight rather than density heterogeneity. This was confirmed by sedimentation analysis showing a very broad dispersed band with an average *s* value of less than 5, equivalent to a mean molecular weight of less than 0.1×10^6 daltons (Studier, 1965).

Buoyant Density in Alkali. In order to ascertain if strand separation, due to differing G+T contents was possible (Vinograd, Morris, Davidson, and Dove, 1963), DNAs with low G+C contents from *Kid mycoplasma,*

Table 1. Buoyant densities and base compositions of bacterial DNAs

	ϱ g/cm^3	% G+C from ϱ	% G+C[a]	From doublet frequency analysis						
				% G+C	A	T	G	C	A/T	G/C
Rodent Myco.	1.687	28	[c]	25	35.6	39.5	12.0	12.9	0.90	0.93
Kid myco.	1.687	28	25	26	35.5	38.8	12.9	12.8	0.92	1.01
C. perfringens	1.691	32	27	28	34.7	37.5	12.8	14.0	0.93	0.99
S. aureus	1.695	35	34	33	32.8	34.5	17.2	15.6	0.95	1.10
B. megaterium	1.693	34	37	34	32.9	33.2	17.2	16.8	0.99	1.02
P. vulgaris	1.696	37	37	37	30.8	32.6	18.8	17.8	0.94	1.06
"Vibrio"	1.699	39	[c]	37	30.0	32.8	19.0	18.2	0.91	1.04
P. phosphoreum	1.701	42	[c]	39	30.2	31.3	19.5	19.0	0.96	1.03
S. pneumoniae	1.701	42	39	40	29.1	30.9	21.1	19.0	0.94	1.11
A. variabilis[b]	1.703	44	[c]	42	27.4	30.2	22.1	20.2	0.91	1.10
C. fritschii[b]	1.704	45	[c]	43	26.9	30.5	22.2	20.4	0.88	1.09
S. marcescens	1.714	55	57	57	20.8	22.4	28.3	28.5	0.93	0.99
R. rubrum	1.722	63	63	62	18.4	20.1	30.3	31.3	0.92	0.97
P. aeruginosa	1.726	67	64	64	16.9	18.7	33.1	31.3	0.90	1.06
R. capsulata	1.725	66	[c]	67	15.8	17.6	32.7	33.9	0.90	0.96
R. sphaeroides	1.728	70	[c]	68	15.1	17.1	23.2	34.7	0.88	0.96
S. lutea	1.730	71	71	68	15.4	16.6	34.2	33.8	0.93	1.01

[a] From Shapiro (1968) and Hill (1966).

[b] Blue-green algae.

[c] Values not available.

Clostridium perfringens and *Staphylococcus aureus* were subjected to buoyant density analyses in alkaline CsCl, pH 12–13. *C. perfringens* and *S. aureus* DNAs both showed a single symmetrical band with no indication of strand separation. The *Kid mycoplasma* DNA showed a very broad dispersed pattern due to the low molecular weight. While there was no indication of a clear bi-modal distribution, it was not possible to rule out at least partial strand separation.

These results contrast with the clear separation of *Tetrahymena pyriformis* DNA into two strands achieved by Woese and Bleyman (1972); the implications for the composition of coding DNA in these low G+C species are discussed in the following paper (Elton, 1973).

b) Doublet Frequency Analyses

Base Composition. The base composition data gained from doublet frequency analysis are shown in Table 1. From these it can be seen that the G+C contents from the analyses are generally in good agreement with those calculated from buoyant density. However, there is a general tendency to a slightly lower G+C content. This is most marked in *C. perfringens* and *P. aeruginosa.* Also there is invariably an excess of T over A. These tendencies have already been noted previously (Swartz *et al.*, 1962; Subak-Sharpe *et al.*, 1966) but no satisfactory explanation of these discrepancies is, as yet, available.

Doublet Frequencies. The sixteen possible doublets present in a double stranded DNA can be divided into two categories.

1. The 12 *dependent doublets* form six pairs because in the combined anti-parallel strands the complementary doublet frequencies ApA–TpT, GpT–ApC, TpG–CpA, GpA–TpC, ApG–CpT, and GpG–CpC are theoretically identical. It follows that doublet analyses give only limited information as to the frequency of dependent doublets on any one strand of the DNA.

2. The *independent doublets* ApT, TpA, GpC, and CpG which are their own anti-parallel complements and are therefore each present with their measured frequency in both strands of the DNA.

The results of the doublet frequency analyses are shown in Table 2 with the dependent doublets shown in complementary pairs. The results are further arranged to show those doublets containing only A and T and those containing only G and C grouped together at the top and bottom respectively.

From Table 2 it can be seen that the values for the dependent doublets are generally in good agreement. The only consistent major discrepancies were found in the ApA–TpT values, particularly of the low G+C DNAs up to *C. fritschii,* in which ApA<TpT. (This discrepancy is removed by normalization, as will be discussed later.)

Table 2a. Doublet frequencies of bacterial DNAs

	Rodent mycoplasma (25)	*Kid mycoplasma* (25)	*C. perfringens* (28)	*S. aureus* (33)	*B. megaterium* (34)	*P. vulgaris* (37)
AT	112	123	121	114	108	102
TA	95	118	124	100	95	88
AA TT	164 192	143 167	129 149	115 127	120 122	105 120
GT AC	35 35	43 39	41 38	55 49	51 50	54 48
TG CA	51 51	53 50	48 45	69 59	62 58	70 64
GA TC	47 56	44 50	49 53	54 49	55 53	51 48
AG CT	45 56	50 55	59 65	48 50	52 51	53 50
GG CC	20 19	20 18	26 26	28 24	31 31	36 35
GC	19	22	23	34	35	47
CG	4	6	5	25	28	29

Table 2b

	Vibrio 01 (37)	*P. phosphoreum* (39)	*S. pneumoniae* (40)	*A. variabilis* (42)	*C. fritschii* (43)	*S. marcescens* (57)
AT	97	98	83	77	76	59
TA	74	86	66	71	66	35
AA TT	106 125	99 107	99 109	84 105	87 109	50 59
GT AC	53 48	57 53	53 48	59 49	57 46	54 50
TG CA	74 68	70 67	69 62	73 61	72 60	69 67
GA TC	52 55	51 51	66 64	58 56	56 58	56 61
AG CT	49 53	51 52	64 63	60 62	60 63	49 53
GG CC	38 33	36 35	44 38	54 45	51 42	69 69
GC	46	51	42	52	58	105
CG	28	37	29	36	39	96

Table 2c

	R. rubrum (62)	*P. aeruginosa* (64)	*R. capsulata* (67)	*R. sphaeroides* (68)	*S. lutea* (68)
AT	56	37	45	40	26
TA	22	24	10	11	12
AA TT	37 47	33 36	32 39	23 29	19 20
GT AC	47 47	54 46	42 39	43 39	64 58
TG CA	59 59	65 53	58 54	57 50	69 63
GA TC	66 73	59 62	63 69	67 75	61 65
AG CT	45 51	53 60	43 51	49 59	51 55
GG CC	95 98	96 83	91 100	96 106	108 106
GC	95	123	133	127	108
CG	105	117	135	131	114

Numbers in parenthesis are the % G+C contents.

Comparison of the analyses show that, as might be expected, the frequencies of the doublets containing only A and/or T and those containing only G and/or C reflect the differing G+C contents. However, groups do exist within the DNAs which share common characteristics not necessarily dependent on the overall G+C content e.g. the distinct shortage of the doublet CpG compared to GpG, CpC, and GpC in *Kid mycoplasma, Rodent mycoplasma,* and *C. perfringens* and the excess of the doublet ApT relative to TpA, ApA, and TpT in the photosynthetic bacteria.

Because of the obvious difficulty of analysing these data in detail and simultaneously allowing for the differences in the overall G+C contents, the general design for each DNA was determined as described in the Methods section. These values are shown in Figs. 1–4 together with the corresponding patterns for *Hemophilus influenzae, Bacillus subtilis, Escherichia coli, Aerobacter aerogenes, Mycobacterium phlei,* and *Micrococcus luteus* calculated from the data of Josse *et al.* (1961) and for *Clostridium pasteurianum*

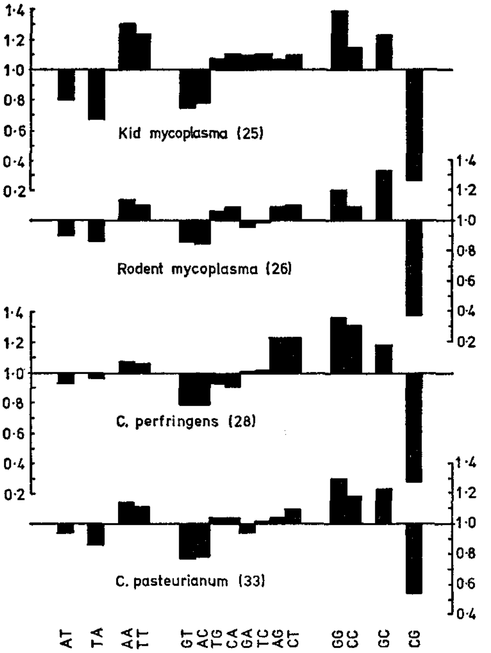

Fig. 1. General design patterns of bacterial DNAs of the low G+C group. Values on the ordinate are the ratios of the observed doublet frequencies to the expected random frequencies. The values in parenthesis are the G+C contents determined from doublet frequency analysis

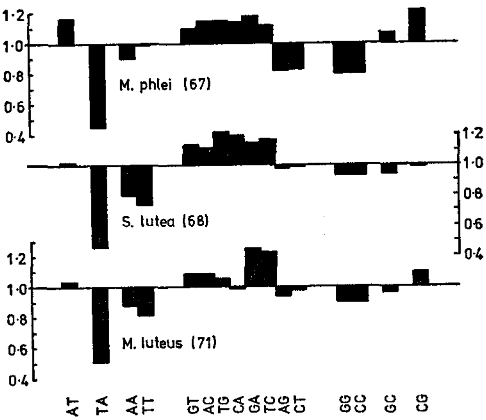

Fig. 2. General design patterns of bacterial DNAs of the extreme high G+C group. Values on the ordinate are the ratios of the observed doublet frequencies to the expected random frequencies. The values in parenthesis are the G+C contents determined from doublet frequency analysis

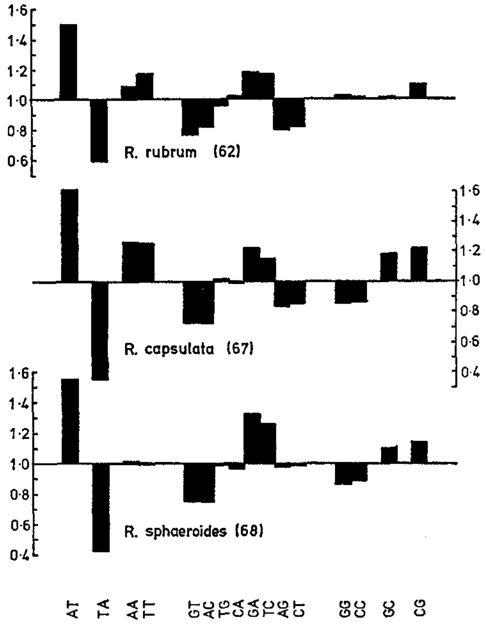

Fig. 3. General design patterns of the photosynthetic bacteria. The values on the ordinate are the ratios of the observed doublet frequencies to the expected random frequencies. The values in parenthesis are the G+C contents determined from doublet frequency analysis

calculated from the data (obtained with RNA polymerase) of Skalka, Fowler, and Hurwitz (1966).

It can now be seen that there are indeed distinct groups with similar patterns of deviations from random expectation. The major groups are:

1. The extreme low G+C group including the *Mycoplasma* and *Clostridia* (Fig. 1).

2. The extreme high G+C group including *M. phlei*, *S. lutea*, and *M. luteus* (Fig. 2).

3. The photosynthetic bacteria group (which also has high G+C) including *R. rubrum*, *R. capsulata*, and *R. sphaeroides* (Fig. 3).

4. The large group covering a range of G+C content of 33–64% which is arbitrarily termed the *E. coli* group. This group also includes the two blue-green algae (Fig. 4). A case could be made for further subdivision which could for example separate *Streptococcus pneumoniae* from the other members of the *E. coli* group.

The above grouping was supported by cluster analysis which resulted in the dendrogram shown in Fig. 5. In this figure, the clustering of the species is measured by the Euclidean squared distance scale; the point where

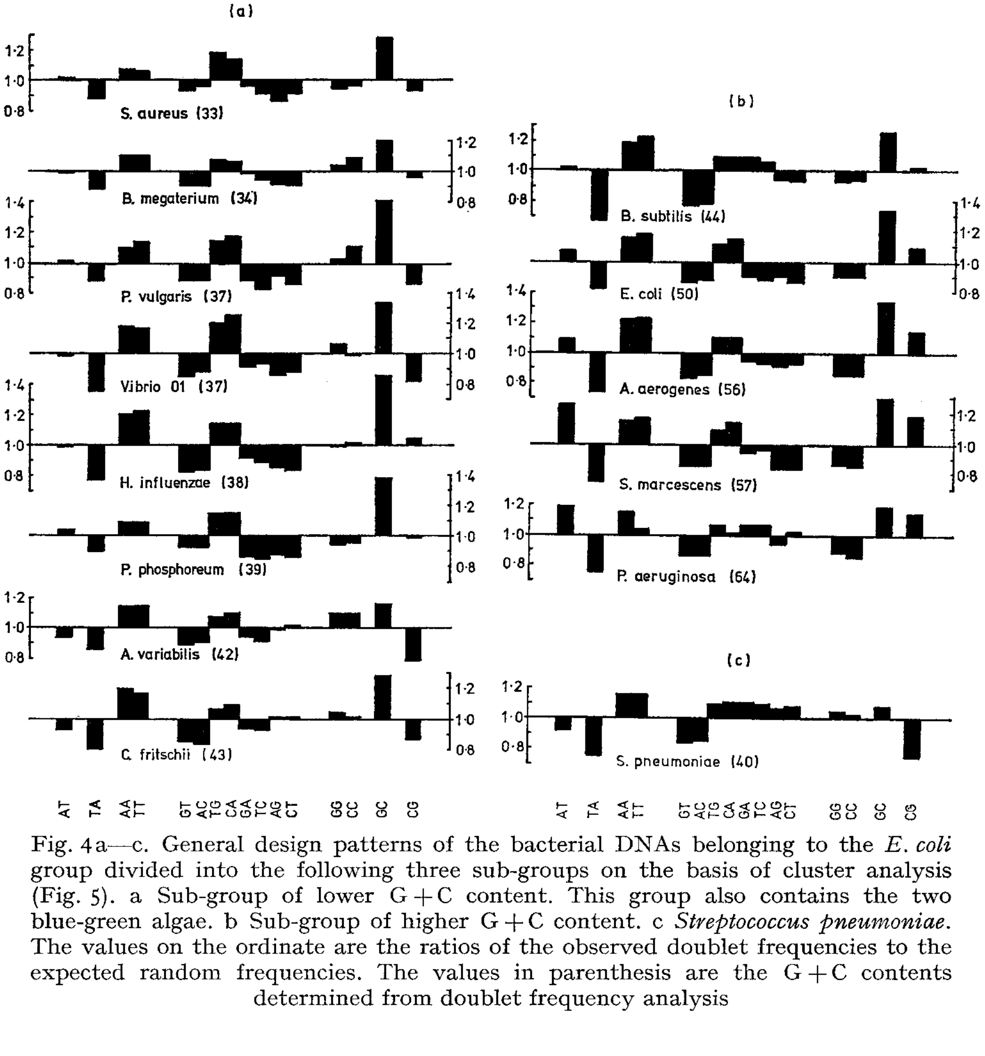

Fig. 4a—c. General design patterns of the bacterial DNAs belonging to the *E. coli* group divided into the following three sub-groups on the basis of cluster analysis (Fig. 5). a Sub-group of lower G + C content. This group also contains the two blue-green algae. b Sub-group of higher G + C content. c *Streptococcus pneumoniae*. The values on the ordinate are the ratios of the observed doublet frequencies to the expected random frequencies. The values in parenthesis are the G + C contents determined from doublet frequency analysis

two branches of the dendrogram are joined gives the mean Euclidean squared distance between the species occurring on those branches. A similar technique has been used before on nearest neighbour frequency data by Bellett (1967), whose conclusions for the six bacterial DNAs then available are in line with our own. The dendrogram also suggests subdivision of the *E. coli* group not only with reference to *S. pneumoniae* but also between species of higher and lower G + C content (Fig. 4a–c).

4. Discussion

The success of the normalization technique in classifying the data into a small number of discrete groups is striking, especially in the case of the *E. coli* group of DNAs. Here, DNAs with a G + C difference of over 30%

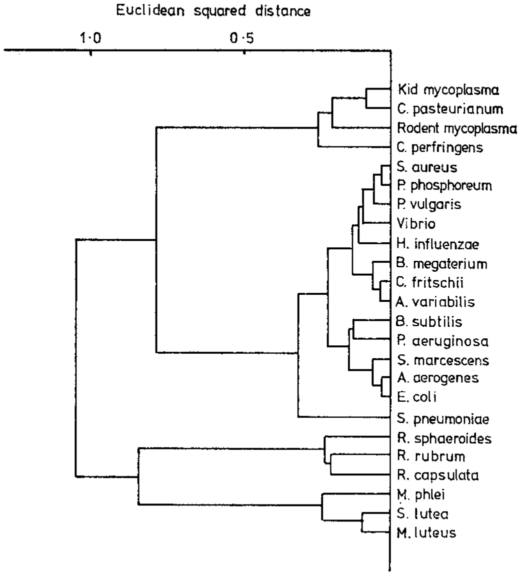

Fig. 5. Cluster analysis of general designs of bacterial DNAs

(*S. aureus* and *P. aeruginosa*) share the major characteristics of this group (excess of ApA–TpT, TpG–CpA, and GpC and shortage of GpT–ApC) although inevitably showing large differences in actual doublet frequencies (Table 2). While the other groups at present span much more limited ranges of G+C content, they also illustrate the usefulness of normalisation in grouping the data. The general design, which reflects the DNA's past response to all those evolutionary pressures on the genetic material which are broadly independent of systematic increase or decrease in G+C content, clearly forms a most useful method of classification. The results demonstrate that organisms with DNA of widely different G+C contents can still be seen to have responded similarly to this aspect of natural selection during their evolution. The most reasonable interpretation is that the general design similarities closely reflect evolutionary relationship.

When the data are considered in relation to conventional taxonomic groupings, many of the findings confirm relationships established by other criteria. Examples of related species with similar general designs are the Enterobacteriaceae (*P. vulgaris, E. coli, A. aerogenes,* and *S. marcescens*) the *Mycoplasma* species, the *Clostridium* species, the *Bacillus* species, *M. luteus* and *S. lutea* and the photosynthetic genera *Rhodospirillium* and *Rhodopseudomonas*.

In several cases, however, discrepancies are revealed between general design and accepted conventional relationships. Perhaps the most striking is the assignment of the genera Bacillus and Clostridium (both members of the family Bacillaceae) to different general design groups; even the two species with very similar G+C contents, *B. megaterium* and *C. pasteurianum*, show many important differences in general design. By contrast, the Pseudomonadales *P. aeruginosa*, *P. phosphoreum*, and *Vibrio* belong to the general design group characteristic of most of the Eubacteriales. Both these observations are consistent with the studies of Sogin, Sogin and Woese (1972) on 5 *s* RNA sequences, which also suggested a closer relationship of the Enterobacteriaceae to *Pseudomonas* than to *Clostridium*. The placing of the two blue-green algae in the *E. coli* general design group can also be correlated with other molecular evidence of evolutionary relationship, in this case the hybridisation observed by Kung (1973) between the DNAs of blue-green algae and *B. subtilis*. The classification of the Micrococcaceae *M. luteus* and *S. aureus* into separate groups reinforces the suspicion which derives from consideration of their G+C contents alone, that they are really only very distantly related. The differences in doublet frequencies between *S. aureus* and *Clostridium*, despite their similar G+C contents, may contribute to the different X-ray diffraction fibre patterns observed in these DNAs by Bram (1972).

The most obvious examples of apparently unrelated organisms with similar doublet patterns occur at the extremes of G+C content. At the high G+C end of the range, the two genera *Micrococcus* and *Mycobacterium* both exhibit extreme shortage of TpA, whereas the low G+C *Clostridium* and *Mycoplasma* species both show a severe shortage of CpG. As Woese (1967) first recognised, and as Elton (1973) shows in more detail in the following paper, these properties are a result of the constraints imposed by amino acid composition and by the form of the genetic code on polypeptide-specifying sequences in the DNA of code limit organisms (i.e. those organisms in which the G+C content of the DNA has been maximised or minimised within the limits possible for coding DNA). It should not therefore necessarily be assumed that similarity of doublet pattern at the extremes of G+C content implies molecular or evolutionary relatedness; it may merely be that different groups of organisms have independently evolved in response to a selective pressure towards extremity of G+C content. Indeed, the evidence from the two low G+C species *C. perfringens* and *Tetrahymena pyriformis* is that they have achieved these G+C contents while accumulating different proportions of A and T in the single strands of their DNAs (Elton, 1973), a finding which reinforces the likelihood of independent evolution to their present similar doublet patterns.

The accurate prediction of the extreme G+C doublet patterns (Elton, 1973) suggests strongly that almost all the DNA in code limit species carries information for the specification of polypeptides, and that the main selection pressure on their doublet patterns is ultimately chanelled through the translation apparatus, although it need not originate there. While we have

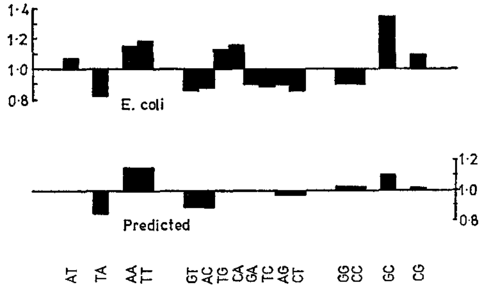

Fig. 6. Predicted general design pattern for protein-specifying DNA calculated for protein of the *E. coli* amino-acid composition assuming a uniform codon weighting. The experimentally determined general design pattern for *E. coli* is shown for comparison

no evidence to suggest that a similar situation does not apply equally to species with intermediate G+C contents, there seems to be no simple explanation of the *E. coli* type pattern in terms of coding DNA sequences. In fact, as Goel *et al.* (1972) have shown, a number of deviations from uniformity of codon usage in protein-specifying DNA need to be assumed in order to explain the doublet frequencies in *E. coli.* Assuming a uniform codon weighting, we have calculated the doublet frequencies to be expected in DNA coding for protein of the *E. coli* amino acid composition [for more details of the method, see the following paper (Elton, 1973)], and the result is shown in Fig. 6. The major discrepancies are clearly excess of GpC and TpG–CpA and shortage of GpA–TpC, ApG–CpT, and GpG–CpC. The deviations from uniform codon weighting predicted by Goel *et al.* (1972) correspond, as might be expected, to the discrepancies between observation and prediction in Fig. 6: e.g. restriction of arginine coding to the codons CGU and CGC and of serine coding to AGU and AGC are among their predictions, and maximal use of CGC and AGC would help to explain the excess in the GpC frequency. The same selective forces which produced these deviations may also be operating in the other members of the *E. coli* general design group, all of which show excess of GpC over random expectation. If this is so, then the common occurrence of rather specific restrictions in base sequences in these "intermediate G+C content" species would support the hypothesis of evolutionary relationship, an argument which we cannot apply to the extreme G+C organisms.

The photosynthetic Athiorhodaceae *R. rubrum, R. sphaeroides* and *R. capsulata* belong to a distinctive general design group, characterised by the large proportional difference between the frequencies of TpA and ApT. The frequency of ApT is so high in these species of comparatively high G+C content that an explanation of the doublet pattern in terms of coding DNA requires the assumption of almost maximum use of this doublet within the constraints of amino acid composition, implying a marked preference for

XAT codons over XAC ones. This contrasts with the situation for *P. aeruginosa* which is of comparable G + C content, and supports the assignment of the photosynthetic species to a distinct taxonomic group from the Pseudomonadaceae.

Perhaps it should be stressed that the spectrum of general designs so far found for the bacteria represents a very limited set in comparison to those already established (Josse *et al.*, 1961; Swartz *et al.*, 1962) and shown by our own unpublished results. In particular, quite different characteristic general designs are found in vertebrates, invertebrates, insects, higher plants and unicellular eukaryotes.

In comparison to other molecular methods, such as polypeptide sequence analysis or nucleic acid hybridization, doublet analysis is clearly an insensitive technique for probing homology differences. However doublet analysis should become the method of choice when *ancient relationships* need to be discerned because, in contrast to molecular hybridization, it provides a measure of the past response of the DNA to selection pressures *averaged over the whole genome*. Information from molecular hybridisation and polypeptide sequence analysis can be complemented by the broader picture of evolutionary relationships obtainable from doublet analysis. With this in mind, our present rather uneven coverage of the bacteria is being extended as DNAs become available.

It is perhaps ironic that the best pointer to the most ancient relationship of living organisms continues to reside in the sum total of their genetic information, their genome, which also represents the organisms potential for future evolution.

Acknowledgements. We are pleased to acknowledge the generous gifts of organisms and DNA preparations as listed under "Materials and Methods". Our thanks are also due to Mr. A. Dolan for his expert technical assistance.

References

Aposhian, H. V., Kornberg, A.: J. biol. Chem. **237**, 519 (1962)
Bellett, A. D. J.: J. molec. Biol. **27**, 107 (1967)
Bram, S.: Biochem. biophys. Res. Commun. **48**, 1088 (1972)
DeLey, J., Rassel, A.: J. gen. Microbiol. **41**, 85 (1965)
Elton, R. A.: J. molec. Evolution **2**, 263 (1973)
Freese, E.: J. theoret. Biol. **3**, 82 (1962)
Goel, N. S., Subba Row, G., Yčas, M., Bremermann, H. J., King, L.: J. theoret. Biol. **35**, 399 (1972)
Halliburton, I. W., Hill, E. A., Russell, G. J.: Biochem. J. **124**, 62P (1971)
Hill, L. R.: J. gen. Microbiol. **44**, 419 (1966)
Hurwitz, J., Furth, J. J., Anders, M., Ortiz, P. J., August, J. T.: Cold Spr. Harb. Symp. quant. Biol. **26**, 91 (1961)
Josse, J., Kaiser, A. D., Kornberg, A.: J. biol. Chem. **236**, 864 (1961)
Josse, J., Swartz, M.: Meth. Enzymol. **6**, 739 (1963)
Kaiser, A. D.: J. molec. Biol. **4**, 275 (1962)
Kung, S. D.: F.E.B.S. Lett. **29**, 259 (1973)

Lee, K. Y., Wahl, R., Barbu, E.: Ann. Inst. Pasteur **91**, 212 (1956)
McGeoch, D. J., Crawford, L. V., Follet, E. A. C.: J. gen. Virol. **6**, 33 (1970)
Mandel, M.: Ann. Rev. Microbiol. **23**, 239 (1969)
Marmur, J., Falkow, S., Mandel, M.: Ann. Rev. Microbiol. **17**, 329 (1963)
Richardson, C. C., Schildkraut, C. L., Aposhian, H. V., Kornberg, A.: J. biol. Chem. **239**, 222 (1964)
Rottem, S., Stein, O., Razin, S.: Arch. Biochem. **125**, 46 (1968)
Russell, G. J., Follett, E. A. C., Subak-Sharpe, J. H., Harrison, B. D.: J. gen. Virol. **11**, 129 (1971)
Schildkraut, C. L., Marmur, J., Doty, P.: J. molec. Biol. **4**, 430 (1962)
Shapiro, H. S.: In: Handbook of biochemistry, H. A. Sober, Ed., p. H-30. Cleveland: Chemical Rubber 1968
Skalka, A., Fowler, A. V., Hurwitz, J.: J. biol. Chem. **241**, 588 (1966)
Sogin, S. J., Sogin, M. L., Woese, C. R.: J. molec. Evolution **1**, 173 (1972)
Sokal, R. R., Sneath, P. H. A.: Principles of numerical taxonomy. San Francisco: Freeman 1963
Stanier, R. Y., Van Niel, C. B.: Arch. Mikrobiol. **42**, 17 (1962)
Studier, F. W.: J. molec. Biol. **11**, 373 (1965)
Subak-Sharpe, J. H.: Brit. med. Bull. **23**, 161 (1967)
Subak-Sharpe, J. H.: In: Proceedings of the 8th Canadian Cancer Research Conference, p. 242. London: Pergamon 1969
Subak-Sharpe, J. H., Bürk, R. R., Crawford, L. V., Morrison, J. M., Hay, J., Keir, H. M.: Cold Spr. Harb. Symp. quant.' Biol. **31**, 737 (1966)
Swartz, M. N., Trautner, T. A., Kornberg, A.: J. biol. Chem. **237**, 1961 (1962)
Szybalski, W.: Fractions **1**, 1 (1968)
Vinograd, J., Brunner, R.: Fractions **1**, 2 (1966)
Vinograd, J., Morris, J., Davidson, N., Dove, W. F.: Proc. nat. Acad. Sci. (Wash.) **49**, 12 (1963)
Woese, C. R.: The genetic code. New York: Harper and Row 1967
Woese, C. R., Bleyman, M. A.: J. molec. Evolution **1**, 233 (1972)

Dr. G. J. Russell
Dr. R. A. Elton
Prof. J. H. Subak-Sharpe
M.R.C. Virology Unit
University of Glasgow
Church Street
Glasgow, G11 5JR, United Kingdom

Dr. D. J. McGeoch
Institute of Biochemistry
University of Glasgow
Glasgow, United Kingdom