Interrelatedness of 5S RNA sequences studied by correspondence analysis

ARTICLE *in* JOURNAL OF MOLECULAR EVOLUTION · FEBRUARY 1987

Impact Factor: 1.68 · DOI: 10.1007/BF02111236

Carmen A Mannella

Wadsworth Center, NYS Department of Health

J Mol Evol (1987) 24:228–235

Journal of
Molecular Evolution

Interrelatedness of 5S RNA Sequences Investigated by Correspondence Analysis

Carmen A. Mannella,[1] Joachim Frank,[1] and Nicholas Delihas[2]

[1] Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, New York 12201, USA
[2] Department of Microbiology, School of Medicine, State University of New York at Stony Brook, Stony Brook, New York 11794, USA

Summary. Correspondence analysis (a form of multivariate statistics) applied to 74 5S ribosomal RNA sequences indicates that the sequences are interrelated in a systematic, nonrandom fashion. Aligned sequences are represented as vectors in a 5N-dimensional space, where N is the number of base positions in the 5S RNA molecule. Mutually orthogonal directions (called factor axes) along which intersequence variance is greatest are defined in this hyperspace. Projection of the sequences onto planes defined by these factorial directions reveals clustering of species that is suggestive of phylogenetic relationships. For each factorial direction, correspondence analysis points to regions of "importance," i.e., those base positions at which the systematic changes occur that define that particular direction. In effect, the technique provides a rapid determination of group-specific signatures. In several instances, similarities between sequences are indicated that have only recently been inferred from visual base-to-base comparisons. These results suggest that correspondence analysis may provide a valuable starting point from which to uncover the patterns of change underlying the evolution of a macromolecule, such as 5S RNA.

Key words: 5S RNA — Correspondence analysis — Multivariate statistics — Evolution — Phylogeny

Offprint requests to: C.A. Mannella

Introduction

With the recent explosive increase in the numbers of sequenced nucleic acids and proteins has come increased interest in ways to analyze phylogenetic relationships. Schemes commonly used to this end involve pairwise comparisons of aligned sequences, and produce scores that reflect the relatedness of each sequence to every other one in the set. Phylogenetic "trees" can then be inferred using distances between species that vary according to these scores (e.g., Fitch and Margoliash 1967; Hori 1975; Dayhoff 1976; Hori and Osawa 1979; Fox et al. 1980).

This report illustrates an alternative approach to the problem of the classification of sequences, one that uses multivariate statistics, in particular, correspondence analysis. Multivariate statistical techniques have become a commonplace tool in numerous scientific and technical disciplines. The reader is referred to the excellent text by Lebart et al. (1984) for a formal introduction. Correspondence analysis has been used in molecular biology, for example, to look for patterns in the distribution of bases along particular genomes (Limaiem and Henaut 1984) and to explore the relationship between codon composition and the physico-chemical properties of amino acids (Sjostrom and Wold 1985). In our application, we apply correspondence analysis directly to a set of aligned sequences of a particular macromolecule, 5S ribosomal RNA, to determine whether such an analysis might provide a rapid assessment of the interrelatedness of the sequences and insight into the patterns of change that underlie the evolution of the molecule. The 5S RNA molecule was chosen for this pilot study because it is a small, readily sequenced RNA that has been the subject of numerous phylogenetic studies, the results of which can be compared with the results of correspondence analysis.

As will be explained in Methods, our approach starts by representing sequences of a particular nucleic acid N bases long as a set of vectors with 5N variables. Correspondence analysis is used to find the directions (factor axes) in 5N-space along which the sequence variance is greatest. Projection of the data along low-order factorial directions is found to result in phylogenetically meaningful clustering of the 5S RNA sequences. The factor axes are defined by eigenvectors that are linear combinations of the components of the original sequence vectors. Those base positions that contribute maximally to each eigenvector (termed "important" positions) can be readily identified. For phylogenetically interesting factor axes, these positions represent the regions in the 5S RNA molecule at which systematic changes have occurred in the course of evolution. Thus the technique provides rapid recognition of species-specific signatures identified previously only by tedious visual inspection.

Methods

The primary structure, S, of a nucleic acid can be represented by a 5 × N binary matrix, where N is the number of base positions, each of which can be occupied by one of four bases (adenine, A; guanine, G; cytosine, C; and uracil or thymine, U or T) or be blank (B). An example would be

$$\text{ACGBU}\ldots = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & \ldots N \\ \hline \text{A} & \text{X} & & & & & \ldots \\ \text{G} & & & \text{X} & & & \ldots \\ \text{C} & & \text{X} & & & & \ldots \\ \text{U} & & & & & \text{X} & \ldots \\ \text{B} & & & & \text{X} & & \ldots \end{array}$$

$$j \xrightarrow{\qquad\quad}$$

$$= i\downarrow \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \ldots \\ 0 & 0 & 1 & 0 & 0 \ldots \\ 0 & 1 & 0 & 0 & 0 \ldots \\ 0 & 0 & 0 & 0 & 1 \ldots \\ 0 & 0 & 0 & 1 & 0 \ldots \end{bmatrix}$$

$$= S \tag{1}$$

This 5 × N matrix can be represented as a binary string of 5N elements,

$$\underbrace{10000\ldots}_{\substack{i=1 \\ j=1 \text{ to } N}}\underbrace{00100\ldots}_{\substack{i=2 \\ j=1 \text{ to } N}}\underbrace{01000\ldots}_{\substack{i=3 \\ j=1 \text{ to } N}}\underbrace{00001\ldots}_{\substack{i=4 \\ j=1 \text{ to } N}}\underbrace{00010\ldots}_{\substack{i=5 \\ j=1 \text{ to } N}} \tag{2}$$

in which each element represents the coefficient F_k $[k = j + N(i - 1)]$ of each component in a 5N-dimensional vector

$$\mathbf{S} = \sum_{k=1}^{5N} F_k \mathbf{u}_k \tag{3}$$

where $\mathbf{u}_k$ is the unit vector in the k-th direction. Note that in this mathematical representation, each position along the sequence and each type of base (including "blanks") is weighted equally. No pre-knowledge regarding nearest-neighbor frequencies or conversion probabilities for particular bases is included.

According to this vector representation, a set of P aligned sequences corresponds to a "cloud" of P points in 5N-space. In correspondence analysis (Benzecri 1969), the chi-square distance between any two points is used as a measure of the overall relatedness of the respective sequences. The orthogonal directions in the data cloud along which intersequence variance is greatest are determined, and these directions (called "factor axes") are used to define a new set of unit vectors, $\mathbf{v}_m$(called "eigenvectors"). [See Lebart et al. (1982) or Bretaudiere et al. (1981) for details of the mathematical procedures.] In effect, the new coordinate system is adapted to the shape of the vector cloud itself. Projection of the cloud into two-dimensional spaces defined by low-order eigenvectors (i.e., those associated with the greatest intersequence variance) may reveal clustering of related sequences.

The eigenvectors found by correspondence analysis are linear combinations of the original unit vectors:

$$\mathbf{v}_m = \sum_{k=1}^{5N} C_{m,k}\mathbf{u}_k \tag{4}$$

Each coefficient $C_{m,k}$ is the contribution of the k-th original unit vector to defining the m-th eigenvector. Thus, the sum of the absolute values of the five coefficients associated with a given base position, j, is a measure of the total contribution of that position to defining the m-th eigenvector, i.e., the "importance," I, of the j-th base position in spanning the m-th factorial direction:

$$I_{m,j} = \sum_{i=1}^{5} |C_{m,j+(i-1)N}| \tag{5}$$

For the purposes of this article, correspondence analysis has been applied to a set of 74 sequences of 5S RNA. Included in the analysis (Table 1) are RNAs from 28 eubacteria, 7 chloroplasts, 3 archaebacteria, 35 eukaryotes, and 1 mitochondrion, each identified by a two-digit number in Table 1 and Fig. 3.

The FORTRAN computer programs used for the multivariate statistical analysis were run on a VAX 11/780 computer, and are available on request. These programs were originally implemented for image-processing applications, i.e., to establish the relatedness of different views of macromolecules in electron micrographs (van Heel and Frank 1981; Frank and van Heel 1982). Since the analysis is essentially an element-by-element comparison of binary strings like (2) above, correct alignment of the sequences is a critical prerequisite. Alignment of the sequences was done according to a consensus scheme for the generalized secondary structure of 5S RNA (Fig. 1a; see Delihas et al. 1984). Blanks were used in the alignment procedure to compensate for apparent insertions and deletions in the different sequences. Because of this, the 5S RNA sequence is represented as 140 base positions in length, which is greater than the actual number of bases in any individual sequence (typically 120). The five-base insert at the 3′ end of helix III in wheat mitochondrial 5S RNA was not included in the analysis. Since it is unique to one sequence, it has no more bearing on the similarity of that sequence to one sequence in the set than it does to any other. Also, the first three and last two base positions in each alignment were "silenced" (set to "B") because the occurrence of these terminal bases in sequences of mature 5S RNAs is highly variable (Erdmann et al. 1985).

Table 1. Species included in the correspondence analyses[a]

Eubacteria	
Gram negative	
01	*Escherichia coli*
02	*Beneckea harveyi*
03	*Photobacterium phosphoreum*
04	*Thermus thermophilus*
05	*Pseudomonas fluorescens*
06	*Thermus aquaticus*
07	*Azobacter vinelandii*
08	*Paracoccus denitrificans*
09	*Rhodospirillum rubrum*
10	*Rhodopseudomonas sphaeroides*
11	*Rhodocyclus gelatinosa*
12	*Pseudomonas cepacia*
Gram positive	
13	*Lactobacillus viridescens*
14	*Bacillus subtilis*
15	*Bacillus acidocaldarius*
16	*Bacillus stearothermophilus*
17	*Clostridum pasteurianum*
18	*Micrococcus lysodeikticus*
19	*Streptococcus cremoris*
20	*Streptococcus faecalis*
21	*Streptomyces griseus*
Mycoplasmas	
22	*Mycoplasma capricolum*
23	*Spiroplasma* sp.
24	*Mycoplasma pneumoniae*
Cyanobacteria	
25	*Anacystis nidulans*
26	*Prochloron*
27	*Synechococcus lividus* III
28	*Synechococcus lividus* II
Archaebacteria	
29	*Thermoplasma acidophilum*
30	*Halobacterium cutirubrum*
31	*Sulfolobus acidocaldarius*
Organelles	
Chloroplasts	
32	*Spinacia oleracea* (spinach)
33	*Dryopteris acuminata* (fern)
34	*Nicotiana tabacum* (tobacco)
35	*Lemna minor* (duckweed)
36	*Spirodela oligorhiza*
37	*Euglena gracilis*
38	*Marchantia polymorpha* (moss)
Mitochondrion	
39	*Triticum aestivum* (wheat)
Eukaryotes	
Fungi	
40	*Dictyostelium discoideum*
41	*Phycomyces blakesleeanus*
42	*Blastocladiella simplex*
43	*Schizosaccharomyces pombe*
44	*Saprolegnia ferax* oomycete
45	*Torulopsis utilis*
46	*Neurospora crassa*
47	*Phlyctochytrium irregulare*
Protozoa	
48	*Acanthamoeba castellanii*
49	*Physarum polycephalum*
50	*Crithidia fasciculata*
51	*Tetrahymena thermophila*
52	*Euglena gracilis*
53	*Chlamydomonas* sp.
Marine protists	
54	*Schizochytrium aggregatum*
55	*Thraustocytrium visurgense*
Algae	
56	*Chlorella pyrenoidosa* (green)
57	*Ulva pertusa* (green)
58	*Eisenia bicyclis* (brown)
59	*Gracilaria compressa* (red)
Plants	
60	*Spinacia oleracea* (spinach)
61	*Triticum aestivum* (wheat embryo)
62	*Linum usitatissimum* (flax)
Animals	
Insects	
63	*Bombyx mori* (silk glands)
64	*Philosamia cynthia-ricini* (silkworm)
Nematode	
65	*Caenorhabditis elegans*
Brachiopod	
66	*Lingula anatina*
Marine invertebrates	
67	*Asterias vulgaris* (starfish)
68	*Illex illecebrosus* (squid)
69	*Aurelia aurita* (jellyfish)
Vertebrates	
70	Human KB cells
71	*Xenopus laevis* (frog, somatic)
72	*Xenopus laevis* (frog, oocyte)
73	*Salmo gairdneri* (rainbow trout)
74	*Iguana iguana* (lizard)

[a] All 5S RNA sequences are taken from the compilation by Erdmann et al. (1985), except for no. 24 (Rogers et al. 1985) and no. 28 (Delihas et al. 1985)

Results and Discussion

Correspondence analysis was applied to the complete set of 74 5S RNA sequences (the "global" analysis) and to two subsets, one containing 36 eubacterial and organellar sequences, the other consisting of the 35 eukaryotic sequences. The variances associated with the factorial directions in each analysis are plotted in Fig. 2a–c. The first factor found by correspondence analysis for each of the data sets represents about 10% of the total variance, and the first five factors together account for 29–38% of the total variance. These results can be contrasted with those obtained by correspondence analysis of a set of 64 random 140-base sequences (Fig. 2d). In the

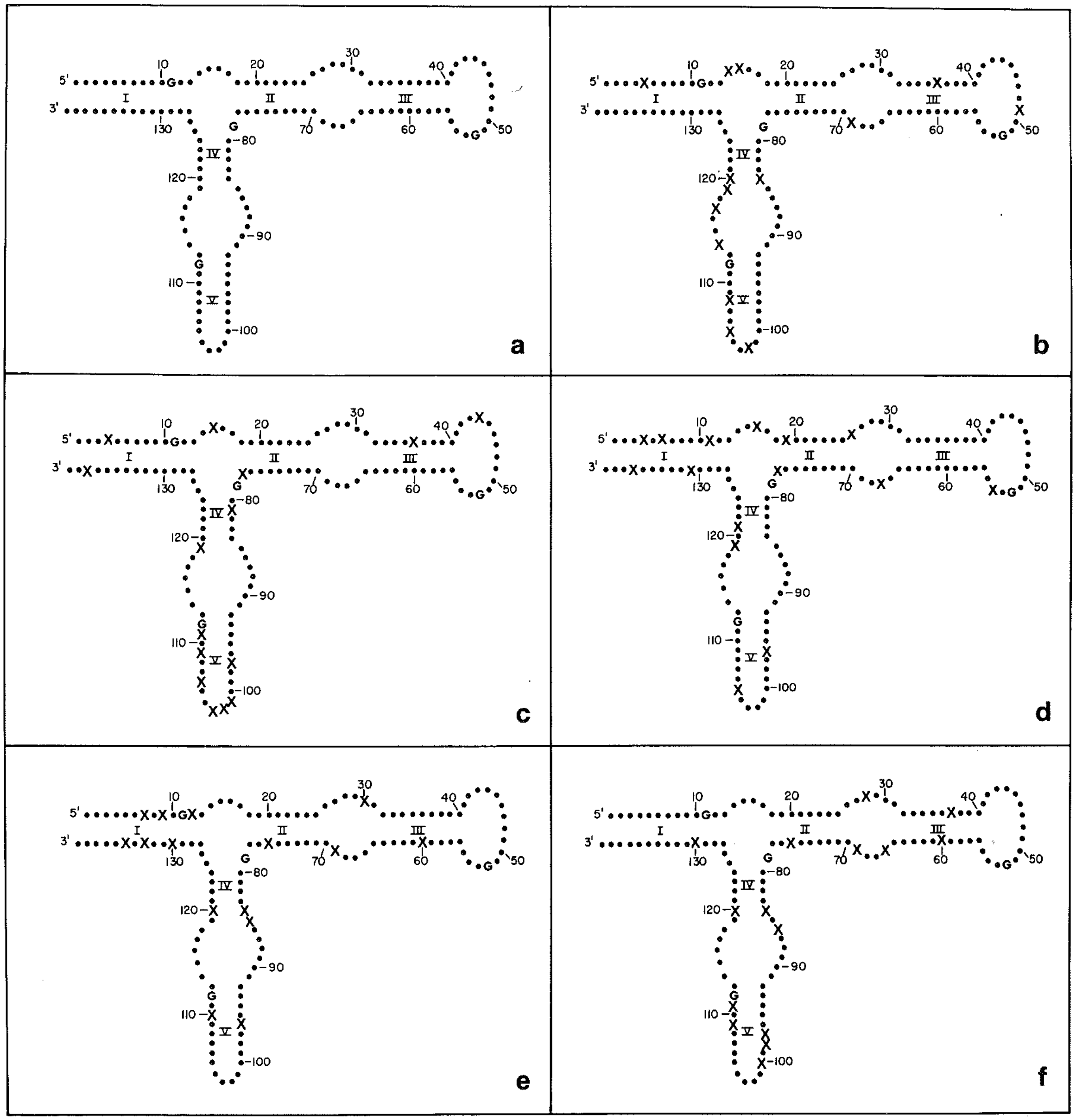

Fig. 1a-f. Important regions in 5S ribosomal RNA sequences. **a** Consensus model for 5S RNA secondary structure from Delihas and Andersen (1982), used in aligning the sequences. Each base position is represented by a dot, except for four universally conserved guanines (G), included as reference points. Base-paired regions (helices I–V) are indicated as parallel linear rows of bases. **b–f** 5S RNA models showing the base positions of high "importance" (labeled "X") associated with particular factorial directions in the three analyses undertaken: **b** global analysis, factor axis 1; **c** eubacteria plus organelles, factor axis 1; **d** eubacteria plus organelles, factor axis 3; **e** eukaryotes, factor axis 1; **f** eukaryotes, factor axis 2

latter case the variance associated with the first factor axis is small (less than 3%) and the first five factors together represent less than 13% of the total variance. Thus the actual data do not behave as if the differences between the sequences were random.

Projection Maps

For each of the 5S RNA sequence sets, maps were obtained representing the projections of the data clouds into two-dimensional subspaces defined by the five statistically most significant factor axes. Figure 3 shows a representative map for each data set.

a. Global Set (All 74 Sequences). Eubacterial and eukaryotic sequences cluster to either side of the origin on factor axis 1 (Fig. 3a), indicating major differences between the 5S RNAs of organisms in the two kingdoms (see, e.g., Delihas and Andersen 1982; Delihas et al. 1984). Both mitochondrial and chloroplast sequences fall on the eubacterial side of the origin on factor axis 1, consistent with the endosymbiotic hypothesis (Margulis 1970). Archae-

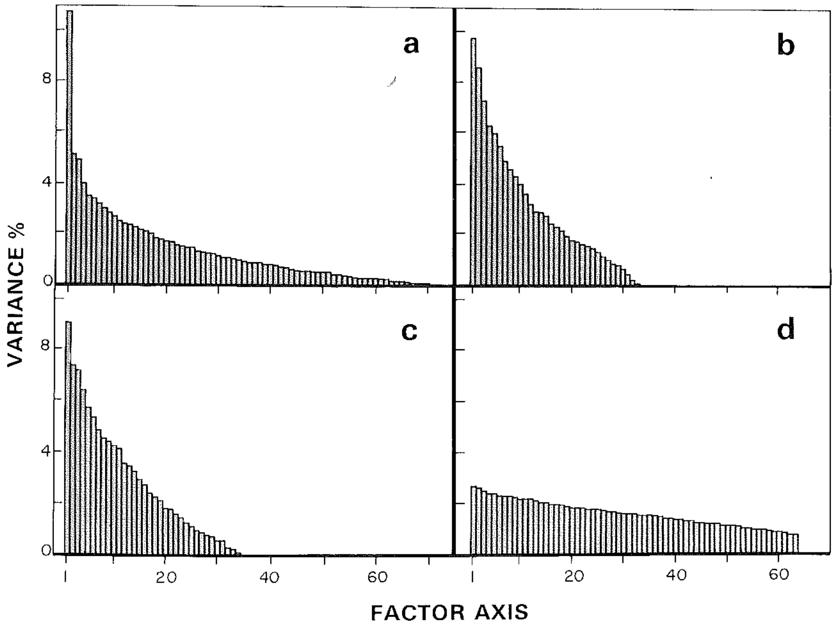

Fig. 2a–d. Variance histograms associated with correspondence analyses of the 5S ribosomal RNA sequence sets undertaken in this study: **a** global; **b** eubacteria plus organelles; **c** eukaryotes; **d** set of 64 random sequences

bacterial 5S RNAs fall near the center of factor axis 1, between the eubacterial and eukaryotic 5S RNAs (see Delihas and Andersen 1982; Fox et al. 1982). While more archaebacterial sequences need to be included in the analysis, this has been prevented by difficulties in aligning these sequences in the 3′ half of the molecule.

b. Eubacteria and organelles. Cyanobacterial and chloroplast sequences separate from the other eubacterial sequences along factor axis 1 (Fig. 3b). There is considerable evidence that chloroplasts evolved from symbiotic cyanobacteria (e.g., Margulis 1970; Schwartz and Dayhoff 1978; Delihas et al. 1984). While the green-plant chloroplast 5S RNAs cluster tightly in two-dimensional projection maps such as Fig. 3, the *Euglena* chloroplast sequence splits off along factor axes 2–5. That the latter sequence differs markedly from the green-plant chloroplast sequences (see Karabin et al. 1983) suggests that the chloroplasts may have polyphyletic origins. The *Clostridium pasteurianum* sequence segregates with the cyanobacterial sequences on the chloroplast side of the origin on factor axis 1 (see discussion below). Similarities between the 5S RNA sequences of *C. pasteurianum* and of spinach chloroplast have been reported by MacKay et al. (1982).

Factor axis 3 in this analysis separates the sequences of Gram-negative bacteria from those of Gram-positive bacteria, with the exception that the two *Thermus* species fall on the Gram-positive side of the origin on this axis. (Factor axis 2 in this analysis was phylogenetically less generally interesting, since it was dominated by differences between the *Euglena* and green-plant chloroplasts.) The mycoplasma sequences cluster with the Gram-positive sequences in the map of Fig. 3b, consistent with the conclusions of Rogers et al. (1985) about possible origins of the mycoplasmas. The mitochondrial sequence lies at the extreme of the Gram-negative side of the origin on factor axis 3, closest to the so-called purple photosynthetic bacteria groups α (*Rhodopseudomonas sphaeroides* and *Paracoccus denitrificans*) and β (*Rhodocyclus gelatinosa* and *Pseudomonas cepacia*) (see Villanueva et al. 1985).

c. Eukaryotes. In the projection of these sequences into the subspace defined by factor axes 1 and 2 (Fig. 3c), protozoan sequences occur in a cluster near the origin, and those of the fungi and the higher eukaryotes branch out from this cluster. There are four phyletically well-defined quadrants in this map: upper left, fungi; lower left, animals; lower right, plants and green algae; and upper right, protozoa, red and brown algae, and marine protists. Green-algal sequences segregate with those of plants, while the red- and brown-algal sequences segregate closer to those of protozoa. The sequence of *Saprolegnia ferax,* classified as a water mold, falls away from those of fungi in this projection, occurring between the sequences of red and brown algae and of the marine protists. Among the protozoan sequences, that of *Chlamydomonas* lies closest to those of the green plants (see Kumazaki et al. 1983), while the *Euglena gracilis* sequence falls nearest to the origin of the projection map. In fact, the *Euglena* 5S RNA sequence falls near the origin of each of the five lowest-order factor axes, suggesting that this sequence may be considered the most typically eukaryotic. The sequences of vertebrates and marine

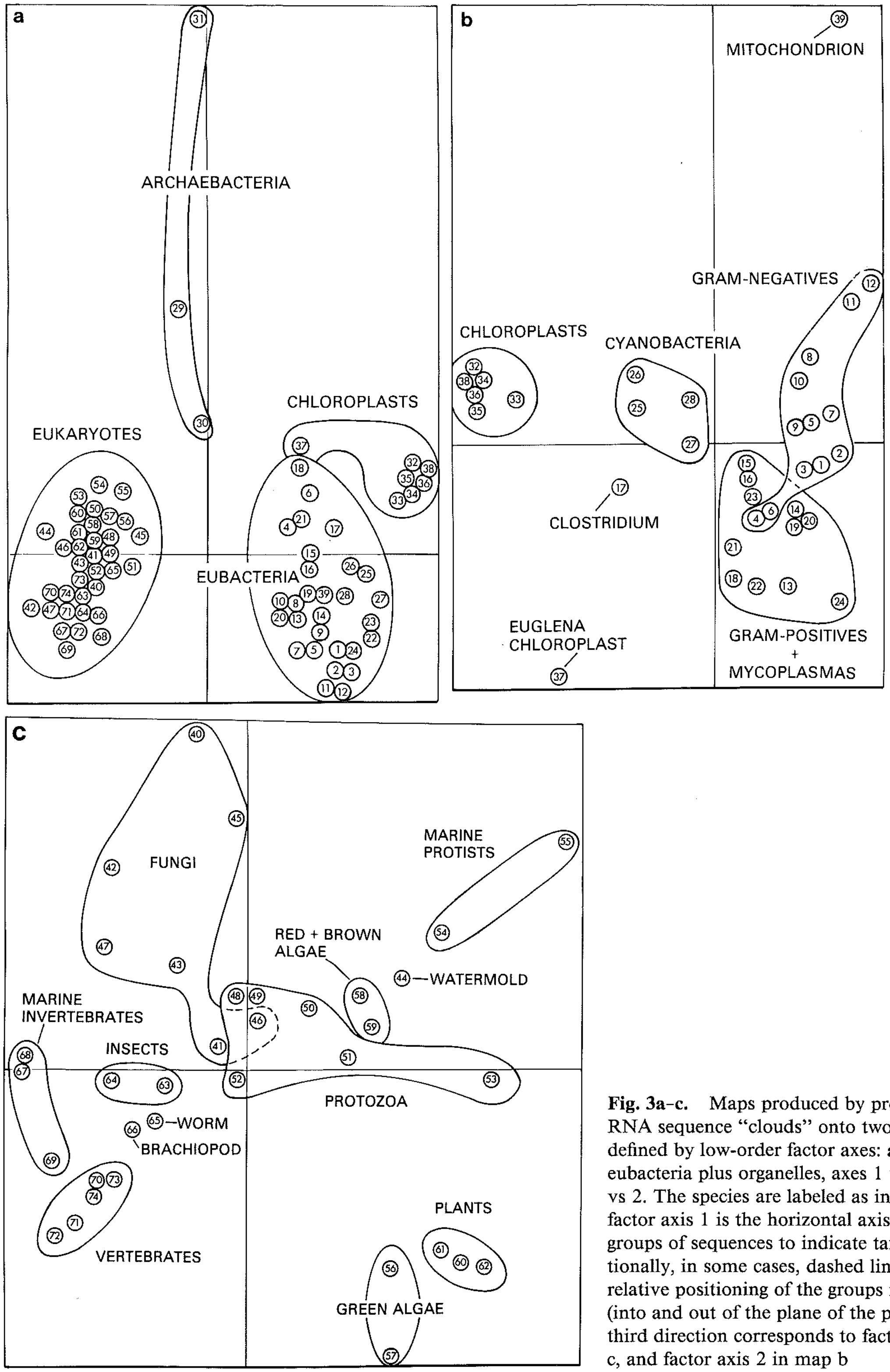

Fig. 3a–c. Maps produced by projecting the 5S ribosomal RNA sequence "clouds" onto two-dimensional subspaces defined by low-order factor axes: **a** global, axes 1 vs 2; **b** eubacteria plus organelles, axes 1 vs 3; **c** eukaryotes, axes 1 vs 2. The species are labeled as in Table 1. In each case, factor axis 1 is the horizontal axis. Lines are drawn around groups of sequences to indicate taxonomic groupings. Additionally, in some cases, dashed lines are used to indicate relative positioning of the groups in the third direction (into and out of the plane of the page) in each map. This third direction corresponds to factor axis 3 in maps a and c, and factor axis 2 in map b

invertebrates branch out from a group containing sequences of insects, a nematode, and the brachiopod *Lingula anatina.*

Important Positions

As explained in Methods, correspondence analysis determines the importance (I) of each base position in the 5S RNA sequences to defining the factor axes in maps like those of Fig. 3 [see Eqs. (4) and (5)]. The "most important" positions are those at which systematic changes occur that are most responsible for the different sequences falling to one or the other side of a particular axis. The 14 or 15 base positions (about 10% of the sequence) that are most important in defining the five phylogenetically interesting factor axes described in the previous section are indicated in Fig. 1b–f. In general, while the important base positions may occur anywhere along the 5S RNA sequence, there are definite indications that these positions cluster at particular regions in each case. For example: (1) For factor axis 1 of the global analysis (which separates eubacterial from eukaryotic sequences), 10 of the 14 most important positions occur in the arm containing helices IV and V and in the short loop between helices I and II. (2) For factor axis 1 of the eubacteria plus organelle analysis (along which chloroplast sequences separate from those of eubacteria), 9 of the 15 most important positions are localized in the stem loops of helices III and V. (3) For factor axis 3 of the eubacteria plus organelle analysis (which separates the sequences of Gram-negative bacteria from those of the Gram-positive bacteria and mycoplasmas), 7 of the 15 most important positions occur in helix I and immediately adjacent regions. (4) Similarly, for factor axis 1 of the eukaryote analysis (which segregates sequences of plants from those of animals, and fungal sequences from protozoan), 6 of the 15 most important base positions are in helix I. (5) For factor axis 2 of the eukaryote analysis (which approximately separates sequences of higher eukaryotes from those of lower eukaryotes), 8 of the 15 most important base positions are in helix V and the regions immediately adjacent to it. Five of the remaining 7 positions are in helix III and in the nonhelical region between it and helix II.

Detailed analysis of the systematic variations at the base positions designated as important by correspondence analysis in each case is beyond the scope of this paper. However, the ability of the analysis to pick out significant trends in the way these 5S RNA sequences vary can be demonstrated by close examination of one case, factorial direction 1 of the eubacteria plus organelle analysis. Ten of the top 15 important positions are strictly conserved among the green-plant chloroplast 5S RNAs [positions 4, U; 36, A; 44, B (deletion); 78, C; 81, U; 101, G; 102, G; 109, U; 119, A; 138, A] and 5 of these positions are also closely conserved among the cyanobacterial 5S RNAs (positions 4, 78, 101, 102, 119). There is little similarity at the same positions with other eubacterial 5S RNA sequences. Only three of the five important base positions conserved in both chloroplast and cyanobacterial sequences had been determined previously by visual inspection (Delihas et al. 1985); the others had been missed.

As noted above, the 5S RNA of *C. pasteurianum* also segregates to the chloroplast side of the origin on factor axis 1. An inspection of the positions shared by the *Clostridium* 5S RNA with a representative plant chloroplast 5S RNA (that from moss) reveals 25 shared bases at positions otherwise highly variable among eubacterial 5S RNAs, including most of the positions in helices II and V. More significantly, the *Clostridium* and chloroplast sequences are homologous at 9 of the top 15 important positions for factor axis 1 in this analysis (positions 4, 81, 97, 101, 102, 103, 106, 109, 111). One is inclined to dismiss the clustering of *C. pasteurianum* with chloroplast and cyanobacterial 5S RNAs as an anomaly, perhaps resulting from sequence convergence. In fact, this result points out the need for more information about the evolution of the 5S RNAs of different species of *Clostridium,* especially in view of the extreme evolutionary stability of the plant chloroplast 5S RNAs and the ancient origin of the clostridia.

Conclusions and Future Directions

The results of applying correspondence analysis to sets of 5S RNA sequences suggest that this technique provides a valuable starting point from which to uncover the patterns of change underlying macromolecular evolution. It should be possible to apply correspondence analysis to other kinds of nucleic acids, as well as to proteins, the limiting factor being the availability of reliably aligned sequences.

The clustering of species in projection maps like those of Fig. 3 is generally consistent with current phylogenetic thought. In fact, such two-dimensional projections are inadequate for displaying the multidimensional correlations that may exist between the sequences. This fundamental limitation may be overcome by techniques that map the sequences nonlinearly, according to their distributions along several of the lowest-order (statistically most significant) factor axes (Radermacher and Frank 1985). Also, mathematical techniques exist that find stable clusters in multidimensional data, and that further determine a hierarchical ranking of these clusters (Lebart et al. 1984). Application of such hierarchical clustering techniques to factorial coordinates produced by correspondence analysis of sequences may be used to generate evolutionary "trees" directly.

As demonstrated in this report, in addition to determining the interrelatedness of sequences, correspondence analysis simultaneously points to those positions in the sequences at which the systematic changes occur that underlie species divergence. In effect, the technique rapidly finds group-specific "signatures" (patterns specific to related organisms), which are difficult to uncover by visual inspection of sequences.

In summary, correspondence analysis, as we have applied it to the study of 5S RNA sequences, may prove to be a powerful tool in the investigation of phylogenetic relationships among macromolecules.

Acknowledgments. The authors are indebted to Drs. D.S. Berns and C.E. Lawrence of the Wadsworth Laboratories for helpful discussions in the course of this research. Dr. S. Kim of the Computer Systems Management and Data Processing Group of the New York State Department of Health is gratefully acknowledged for his assistance in transferring sequence data into the VAX 11/780 computer used in the analyses. Finally, I. Auger is thanked for providing the set of 64 randomly generated base sequences used as controls in the analysis. This research was supported by NIH grant GM29169 and NSF grants PCM 83-02127 and DMB 85-02213.

References

Benzecri JP (1969) Statistical analysis as a tool to make patterns emerge from data. In: Watanabe S (ed) Methodologies of pattern recognition. Academic Press, New York, p 35

Bretaudiere J-P, Dumont G, Rej R, Bailly M (1981) Suitability of control materials. General principles and methods of investigation. Clin Chem 27:798–805

Dayhoff MO (1976) The origin and evolution of protein superfamilies. Fed Proc 35:2132–2138

Delihas N, Andersen J (1982) Generalized structures of the 5S ribosomal RNAs. Nucleic Acids Res 10:7323–7344

Delihas N, Andersen J, Singhal RP (1984) Structure, function and evolution of 5S ribosomal RNAs. Prog Nucleic Acid Res Mol Biol 31:161–190

Delihas N, Andersen J, Berns D (1985) The structure of the 5S ribosomal RNA from the thermophilic cyanobacterium *Synechococcus lividus* II. J Mol Evol 21:334–337

Erdmann, UA, Wolters J, Huysmans E, DeWachter R (1985) Collection of published 5S, 5.8S and 4.5S ribosomal RNA sequences. Nucleic Acids Res 13:r105–163

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:269–284

Fox GE, Stackebrandt E, Hespell RB, Gison J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, Zalben LB, Blakemore R, Gupta R, Bonnen L, Lewis BJ, Stahl DA, Luehrsen KR, Chen KN, Woese CR (1980) The phylogeny of prokaryotes. Science 209:457–463

Fox GE, Luehrsen KR, Woese CR (1982) Archaebacterial 5S ribosomal RNA. Zentralbl Bakteriol Hyg I [C] 3:330–345

Frank J, van Heel M (1982) Correspondence analysis of aligned images of biological particles. J Mol Biol 161:124–137

Hori H (1975) Evolution of 5S RNA. J Mol Evol 7:75–86

Hori H, Osawa S (1979) Evolutionary change in 5S RNA secondary structure and a phylogenetic tree of 54 5S RNA species. Proc Natl Acad Sci USA 76:381–385

Karabin GD, Narita JO, Dodd JR, Hallick RB (1983) *Euglena gracilis* chloroplast ribosomal RNA transcription units. J Biol Chem 258:14790–14796

Kumazaki T, Hori H, Osawa S (1983) Phylogeny of protozoa deduced from 5S rRNA sequences. J Mol Evol 19:411–419

Lebart L, Morineau A, Warwick KA (1984) Multivariate descriptive statistical analysis. John Wiley & Sons, New York

Limaiem J, Henaut A (1984) Etude de la fluctuation de la frequence des quatre bases le long du genome mitochondrial des Mammiferes au moyen de l'analyse factorielle des correspondances. C R Seances Acad Sci 298:279–286

MacKay RM, Salgado D, Bonen L, Stackebrandt E, Doolittle WF (1982) The 5S ribosomal RNAs of *Paracoccus denitrificans* and *Prochloron.* Nucleic Acids Res 10:2963–2970

Margulis L (1970) Origin of eukaryotic cells. Yale University Press, New Haven, Connecticut

Radermacher M, Frank J (1985) Use of nonlinear mapping in multivariate image analysis of molecule projections. Ultramicroscopy 17:117–126

Rogers MJ, Simmons J, Walker RT, Weisburg WG, Woese CR, Tanner RS, Robinson IM, Stahl DA, Olsen G, Leach RH, Maniloff J (1985) Construction of the mycoplasma evolutionary tree from 5SrRNA sequence data. Proc Natl Acad Sci USA 82:1160–1164

Schwartz RM, Dayhoff MO (1978) Origins of prokaryotes, eukaryotes, mitochondria and chloroplasts. Science 199:395–403

Sjostrom M, Wold S (1985) A multivariate study of the relationship between the genetic code and the physical-chemical properties of amino acids. J Mol Evol 22:272–277

van Heel M, Frank J (1981) Use of multivariate statistics in analysing the images of biological macromolecules. Ultramicroscopy 6:187–194

Villanueva E, Luehrsen KR, Gibson J, Delihas N, Fox GE (1985) Localization and the phylogenetic origins of the plant mitochondrion based on a comparative analysis of 5S ribosomal RNA sequences. J Mol Evol 22:46–52

Received April 2, 1986/Revised and accepted August 15, 1986