

- in vitro*: a preliminary communication. The Immunologist, Abstr. 10th International Conference of Immunology, New Delhi, 1–6 November 1998, p. 624.
5. Park, Raekil, Young-Choi, D. U., Kim, Myung-Sun, Soh, Hong-sub, Jung, Byung-Hak, Jun, ChangDuk and Chung, Hun-Taeg, Bacterial endotoxin, lipopolysaccharide, induced the activation of stress – activated protein kinase in RAW 264.7 cells. The Immunologist, Abstr. 10th International Conference of Immunology, New Delhi, 1–6 November 1998, p. 286.
 6. Boyum, A., Separation of leukocytes from blood and bone marrow. *Scand. J. Clin. Lab. Invest. (Suppl.)*, 1968, **21**, 77.

ACKNOWLEDGEMENTS. I thank Dr M. Jagesh Kamath and Mrs Latha Jagesh for help; Prof. R. N. Sreenivasa Gowda, IAH & VB, Hebbal, Bangalore and Dr R. Manjunath, Indian Institute of Science, Bangalore for help and guidance; Mr Ramakrishnan, Sigma-Aldrich, Bangalore for his generous gift of Histopaque Reagent.

Received 6 January 2004; revised accepted 12 May 2004

CAGCAG – the most consistent repeating pattern in evolution of small subunit of rRNA gene sequences

D. S. Iyer, D. V. Raje, H. J. Purohit*,
A. Gupta and R. N. Singh

National Environmental Engineering Research Institute, Nehru Marg,
Nagpur 440 020, India

Conserved patterns in nucleotide sequences are often suspected for their possible structural or functional implications. In this exercise, repeating patterns of nucleotides of size six or more that are conserved in rRNA sequences of three evolutionary domains, have been targetted. The pattern CAGCAG was found to be the most consistent repeating pattern in 16S rRNA of *Proteobacteria* and 18S rRNA of *Eucarya*, but the repetitiveness was not observed in *Archaea*. This pattern or the residues within have not been reported for their biological relevance; but still the information contained between the repeats of the pattern was found to be of much relevance in classification using similarity and multiple discriminant analysis.

THE universal phylogenetic tree of life has been proposed based on small subunit of rRNA gene sequences using alignment techniques. Three lines of evolutionary descent, viz. *Eucarya* (eukaryotes), *Proteobacteria* and *Archaea* have been explored thoroughly^{1,2}. Further, it has been

shown that the residues of small subunit rRNA molecule play a crucial role in protein synthesis³. Amongst different small subunit rRNAs, 16S rRNA has been extensively explored for its association with bacterial diversity⁴ and also for its functional role in protein biosynthesis⁵. The 16S rRNA is found in the 30S sub-unit of the ribosome, which has similar secondary structure with its counterpart 18S rRNA in the 40S subunit of the ribosome in eukaryotes⁶.

Analysing the genetic information in terms of patterns or identifying regions that are preserved during evolution and relating the findings with structure and function of a gene, are issues of immense interest since the last decade. There are some deterministic pattern-discovery algorithms available, which can find sparse amino or nucleic acid patterns matching in protein or DNA sequences^{7–9}. The origin, evolution and distribution of repetitive elements in genome sequences have been studied, both experimentally and computationally. There are programs available for identifying such repeated patterns in large genome sequences and identify repeating patterns of size at least 20 bases as mini or micro satellite information to characterize DNA. Amongst these, the recently developed REPuter (<http://bibiserv.techfak.uni-bielefeld.de/reputer>) has been found to be efficient and provides exhaustive repeats in sequences⁹. Although the programs provide the list of repeats in a sequence, they do not have an option to automatically provide repeating patterns, which are conserved across the input set of homologous sequences by considering the separating distance criterion. We have developed a program, Repeat Tuple Search (RTSearch; www.ebi.ac.uk/~liijnzaad/RepeatTupleSearch), which has this additional feature to determine the consistent repeating patterns (CRPs) across the set of sequences. By consistent, we mean the repeating pattern (length at least six bases) occurring across majority of the input sequences, such that the separating distance between the two repeats in these sequences remains constant. The repeats are exact and do not allow even single base ambiguity. The program works efficiently for small sequences of up to size 2 kb. It has two basic components – the first determines repeating patterns of length more than six (default setting) along with the separating distance between the repeats in each input sequence. The search for repeating patterns is exhaustive, without asking for any input conditions from the user. The second component processes the collective data to get the number of sequences in which different repeating patterns make their appearances, considering the constant separating distance criterion. Patterns with high frequency of occurrence are considered as CRPs.

Earlier, we had reported that four repeating patterns occur with more than 80% consistency across the sampled set of fifty different 16S rRNA sequences of *Pseudomonas*, with CAGCAG being the most consistent repeating pattern. The sub-sequences between the repeats were analysed using information theory to obtain the signature for genus

*For correspondence. (e-mail: hemantdrd@hotmail.com)

Table 1. Most consistent repeating patterns in 16S/18S sequences of the three selected groups

Repeating pattern	Group						
	Archaea (16S)	Proteobacteria (16S)					Eucarya (18S)
		Alpha	Beta	Delta	Gamma	Epsilon	
CAGCAG	–	14 (142)	17 (166, 167)	15 (167, 168)	16 (167)	17 (142, 143)	16 (143, 144)
CGCAAC	–	12 (9)	18 (9)	12 (9)	13 (9)	18 (9)	–
GCAACG	–	18 (133, 135, 136, 137)	16 (133, 135)	4 (718)	12 (132, 133)	13 (684, 685)	–
TGGGGAG	–	20 (111, 112)	15 (111, 113)	11 (112, 113)	14 (112)	18 (112)	–
TACGGG	56 (17)	–	–	–	–	–	–
GAGAGG	16 (379)	–	–	–	–	–	–
GGGTAG	16 (115)	–	–	–	–	–	–
AATTGG	16 (41)	–	–	–	–	–	–
ACGGGG	15 (45)	–	–	–	–	–	–
ATTGAC	–	–	–	–	–	–	18 (87)
GTGGAGC	–	–	–	–	–	–	15 (125)
TTAATT	–	–	–	–	–	–	15 (129)
CGAAAG	–	–	–	–	–	–	14 (38)
CTTTAA	–	–	–	–	–	–	13 (9)
GGTGGTG	–	–	–	–	–	–	13 (3)
AGAGGT	–	–	–	–	–	–	12 (85)

Numbers indicate frequency in each group/sub-group out of 20 sample sequences. Numbers in parenthesis indicate separating distance between repeats.

*Pseudomonas*¹⁰. We carried out CRP search in nearly forty different bacterial genera using 16S rRNA sequences and found that CAGCAG is the most consistent repeating pattern observed in majority of the sequences. Even the separating distance and the region of occurrence of the patterns was found to be more or less same in the sequences. This was precisely the motivation of the study, where we have used representative sequences from all the three lines of evolutionary descent and observed the most consistent repeating patterns in 16S rRNA of *Archaea* and *Proteobacteria* and 18S rRNA of *Eucarya* sequences.

The sample sequences for 16 rRNA and 18S rRNA belonging to three different groups, were retrieved from EMBL, GenBank (www.ebi.ac.uk/embl/contact/collaboration) or NCBI (www.ncbi.nlm.nih.gov/Nucleotide). Following the taxonomic classification, the 16S rRNA group was further divided into *Archaea* and *Proteobacteria* (alpha, beta, gamma, delta and epsilon). The 18S rRNA sequences representing *Eucarya* were also retrieved from GenBank. Twenty sequences from each group with size ~1.5 kb for 16S rRNA and ~1.8 kb for 18S rRNA were randomly picked and used in the study.

The groupwise sequences were used as input to the RTSearch program. Accordingly, the most consistent repeating patterns in each group were obtained (Table 1) with repeating patterns occurring in more than 60% of sequences of the respective groups. The pattern CAGCAG, occurs in all the sub-groups of *Proteobacteria* and dominantly in the 18S rRNA of *Eucarya*, while it is not a dominant feature in *Archaea*. Except CAGCAG, other patterns do not appear simultaneously in two or more groups,

although they are dominant in the respective groups/sub-groups. Since the repeat of CAGCAG is a common feature to both *Proteobacteria* and *Eucarya*, we targetted only this pattern and the information contained between the repeats. It is evident that the separating distance between the repeats of this pattern is nearly the same, with some variations due to insertions or deletions of bases. The observed separating distance for alpha, epsilon and 18S of *Eucarya* is 142 bases, while beta, delta and gamma sub-groups show a separating distance of 167 bases. Hence these groups were classified into two, considering the separating distance as a criterion, to know whether the information contained within the repeats has any relevance in distinguishing the groups from each other. This idea was in support of the earlier work wherein considerable variation in the sub-sequences between the repeats of CAGCAG was observed, even at species level of a bacterial genus¹⁰.

Accordingly, the sub-sequences between the repeats were extracted from the sequences belonging to *Proteobacteria* and *Eucarya* and were aligned using CLUSTAL W method, resulting in a dendrogram. The MegAlign module of Laser Gene software¹¹ was used to generate dendrograms for the two cases (Figures 1 and 2). The relationship amongst the sequences of groups alpha, epsilon and 18S of *Eucarya* have been shown in Figure 1. The three distinct groups are well separated from each other. To support this observation, we alternatively carried out multiple discriminant analysis for the three groups. We have earlier shown that the closely related bacterial groups, viz. *Acinetobacter*, *Alcaligenes*, *Burkholderia*, *Moraxella* and *Pseudomonas*, can be well discriminated based on the dinucleotide pro-

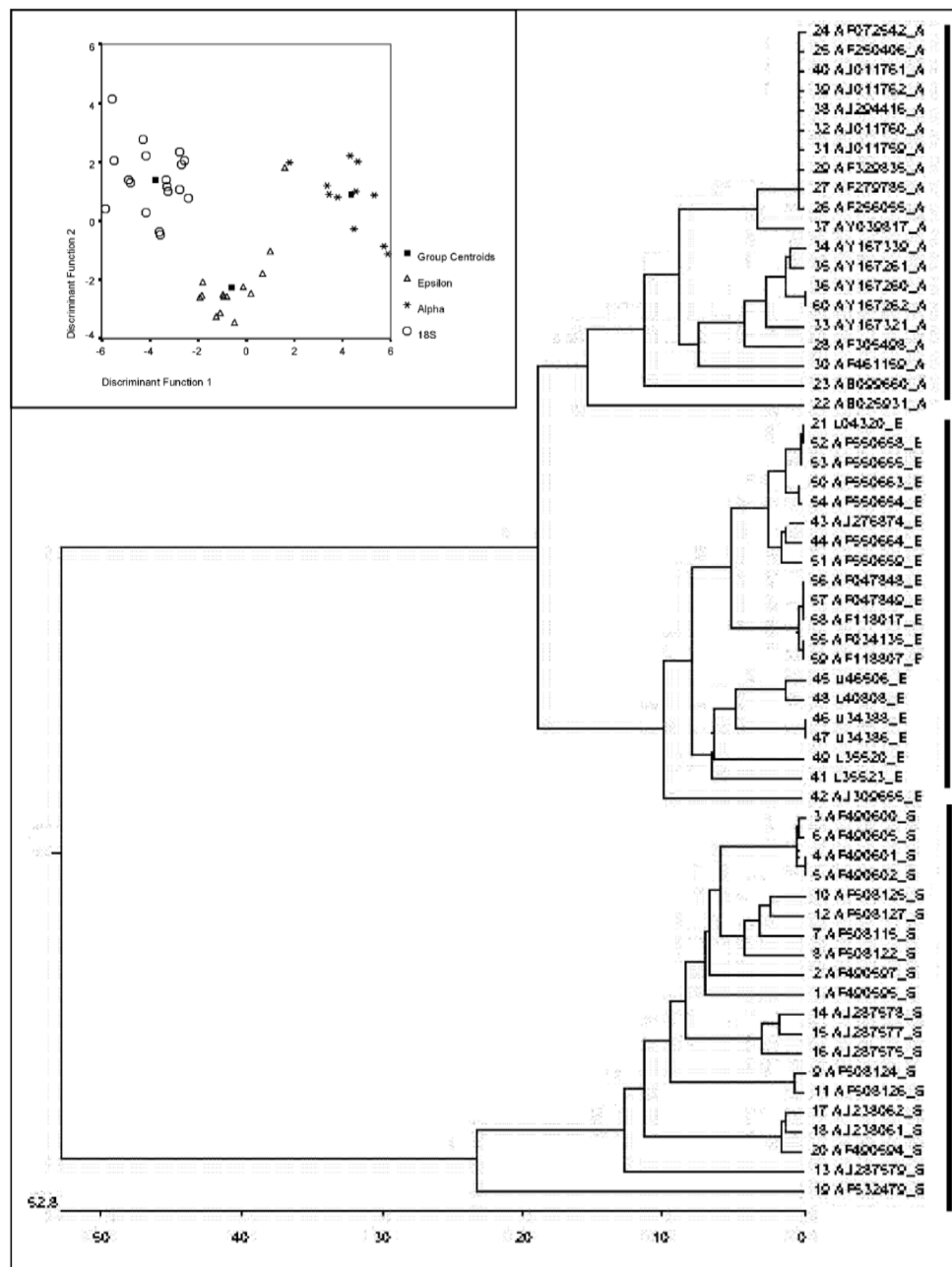


Figure 1. Dendrogram showing the association of sequences from alpha (A), epsilon (E) and *Eucarya* (S) groups using sub-sequence data between repeating pattern CAGCAG. Accession numbers of sequences are according to GenBank. (Inset) Scatter plot for sequences of these groups based on the first two discriminant functions.

abilities as feature space and using multiple discriminant analysis¹². In this study also, data on 16-dimensional feature space was generated for each sub-sequence in each group. Assuming that all the dinucleotides may not have equal contribution in classification, stepwise feature selection was carried out with Wilk's lambda as a selection criterion¹³. It was observed that the features AA, AG, CA,

CT, GC, GT and TC could classify sequences of alpha, epsilon and *Eucarya* with 98% accuracy. The scatter plot giving the classification is shown in the inset of Figure 1. The classification is based on the first two discriminant functions, which provide reasonably good prediction accuracy. The separation of the group centroids based on discriminant function 1 is much wider thereby supporting

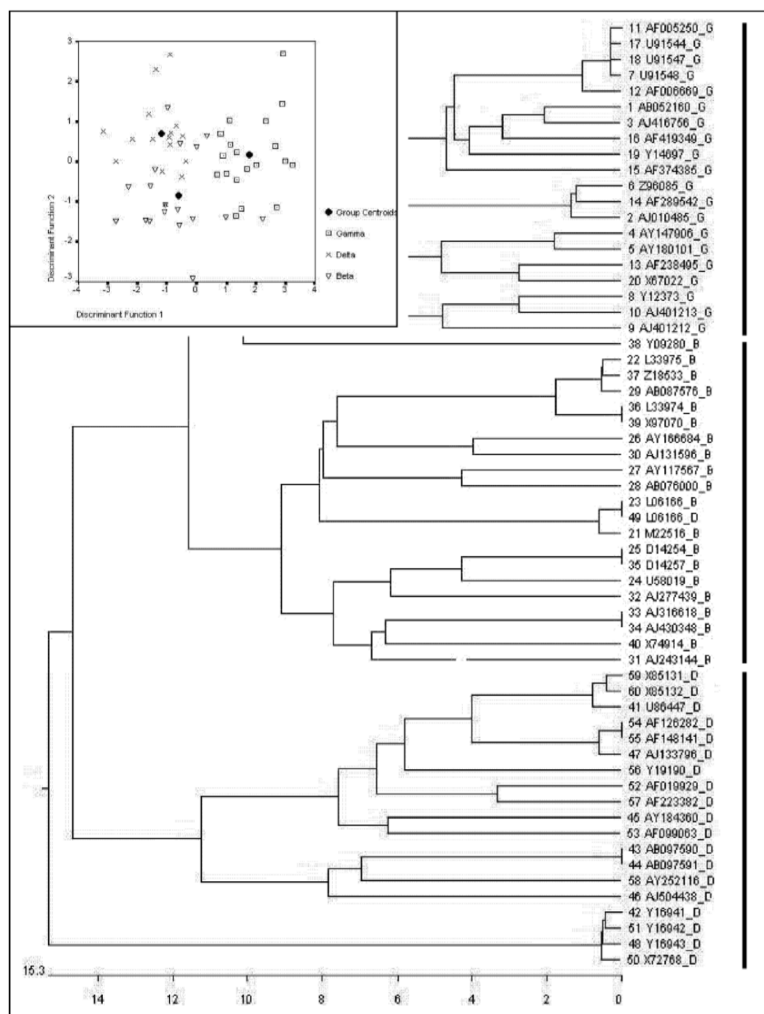


Figure 2. Dendrogram showing the association of sequences from beta (B), delta (D) and gamma (G) groups using sub-sequence data between repeating pattern CAGCAG. Accession numbers of sequences are according to GenBank. (Inset) Scatter plot for sequences of these groups based on the first two discriminant functions.

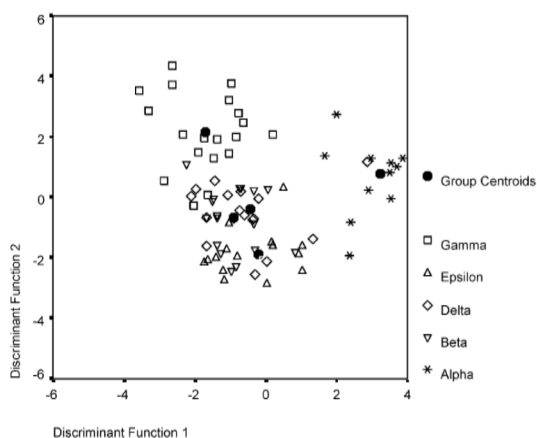


Figure 3. Scatter plot showing the distribution of sequences from five groups of *Proteobacteria* based on stepwise multiple discriminant analysis.

the classification shown through the dendrogram. This reveals that although sequence lengths are the same (~142 bases), information contained between the repeats is sufficient to provide good distinction among these groups. A similar exercise was carried out for the sequences of the other three groups, viz. beta, delta and gamma, having the same separating distance of ~167 bases. Figure 2 displays their relatedness through a dendrogram. The separation of sequences is quite evident, with the exception of one sequence from delta group closely associated with the sequence from beta group. The clustering of groups has been shown in the inset of Figure 2. Multiple discriminant analysis provides 78% classification accuracy using features AA, AG, CA and GG. This is lower in comparison with the earlier case, but still sufficient to distinguish the sequences from each other. The separation of group centroids is also lesser compared to the earlier case, which agrees with the

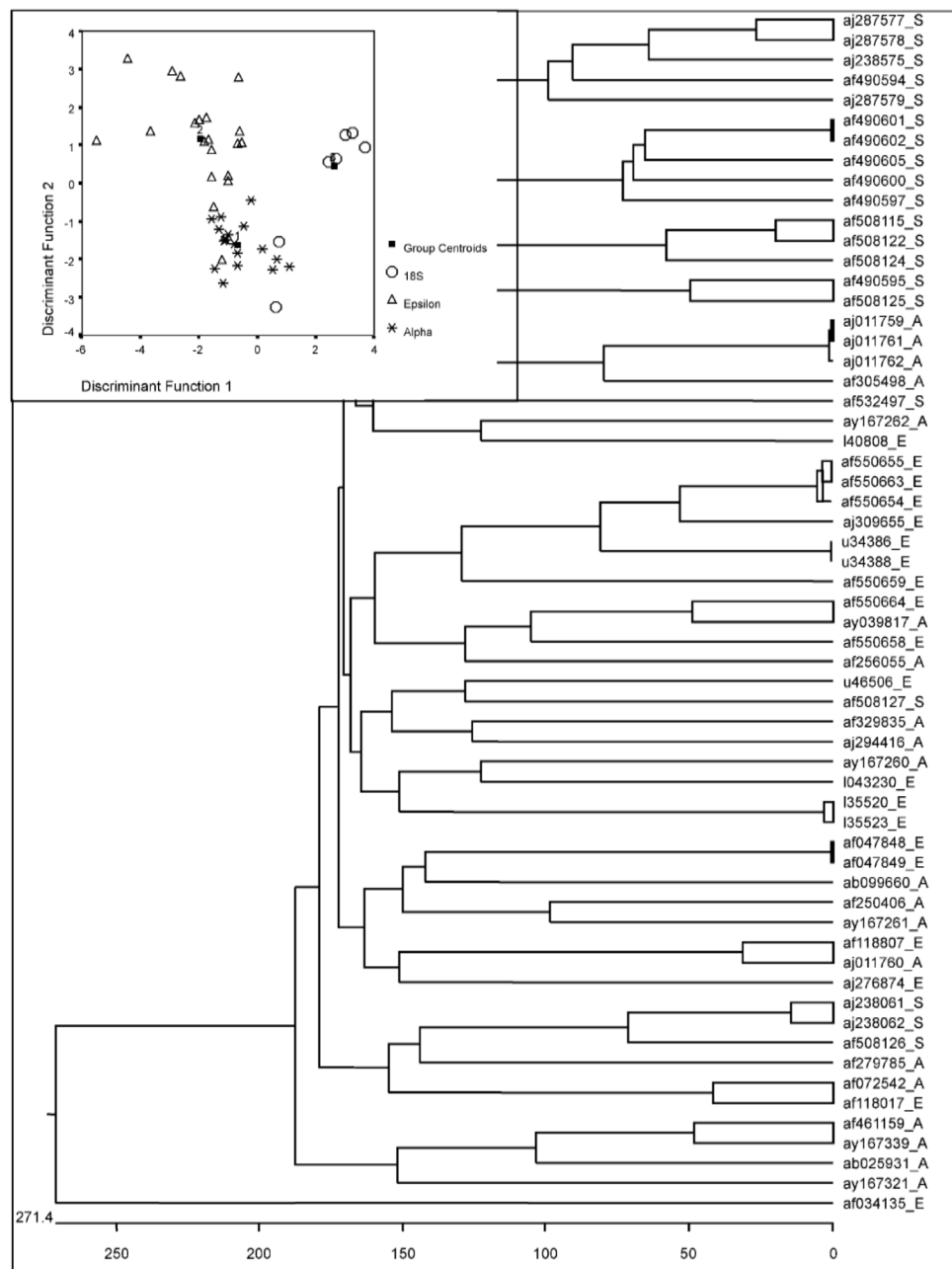


Figure 4. Dendrogram showing the association of sequences from alpha (A), epsilon (E) and *Eucarya* (S) groups using complete sequences. Accession numbers of sequences are according to GenBank. (Inset) Scatter plot for sequences of these groups based on the first two discriminant functions.

group separation using a dendrogram. The scatter plot in Figure 3 further supports this, where five *Proteobacteria* have been discriminated from each other based on their dinucleotide compositions in sub-sequences.

In order to establish the relevance of the selected sub-sequence in classification, the above exercise was carried out using complete sequences from all the groups. In the

first case, sequences of alpha, epsilon and *Eucarya* were used to obtain the dendrogram, while in the second case sequences from beta, delta and gamma were used to generate the dendrogram. The separation in both cases using complete sequences was not as clear as that obtained using the sub-sequence data between repeating tuple CAGCAG. Classification using complete sequence for alpha, epsilon

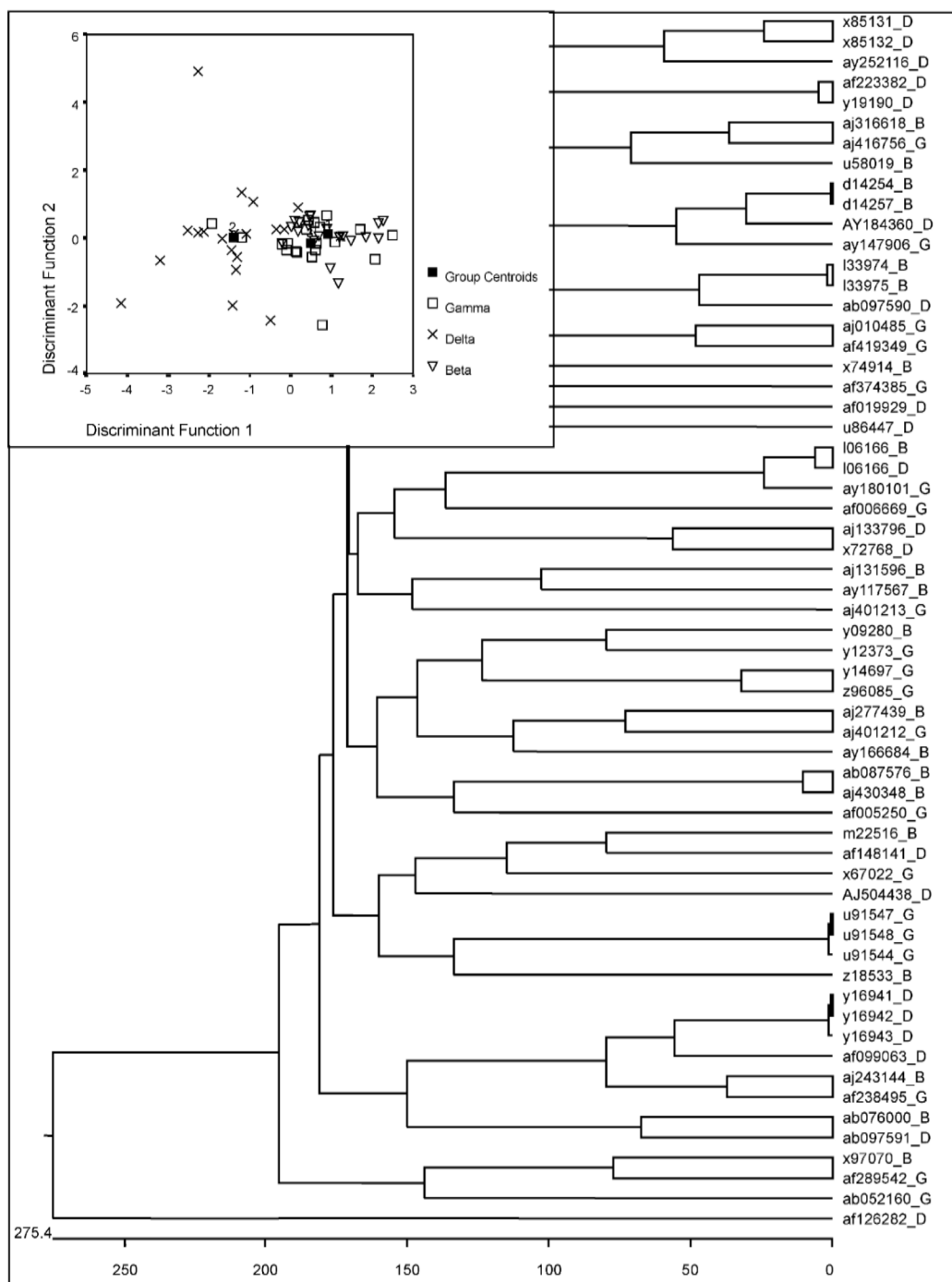


Figure 5. Dendrogram showing the association of sequences from beta (B), delta (D) and gamma (G) groups using complete sequences. Accession numbers of sequences are according to GenBank. (Inset) Scatter plot for sequences of these groups based on the first two discriminant functions.

and *Eucarya* is shown in Figure 4, while that for beta, delta and gamma is shown in Figure 5. Classification accuracy using multiple discriminant analysis in the first case

was 93%, while for the other it was 61%. On the contrary, sub-sequence-based classification yielded fairly good separation indicating that the nucleotide composition within

the region holds sufficient diversity to distinguish the sequences of the groups in the above two cases.

Thus, the application of computational tools in sequence analysis can be useful in extracting patterns of different types. This study demonstrates one such utility by considering the most consistent repeating pattern. The pattern search, in this case, has even yielded supplementary information that is vital to distinguish the groups from each other. Likewise, using different logic for pattern search and with 16S rRNA as a model, the generated knowledge would provide better understanding of the molecular mechanisms of ribosome function as well as the relationship of organisms on an evolutionary scale.

1. Pace, N. R., A molecular view of microbial diversity and the biosphere. *Science*, 1997, **276**, 734–740.
2. Lesk, A. M., *Introduction to Bioinformatics* (ed. Lesk, A. M.), Oxford University Press, New York, 2002, pp. 19–21.
3. Triman, K., Peister, A. and Goel, R., Expanded versions of the 16S and 23S ribosomal RNA mutation databases (16SMDBexp and 23SMDBexp). *Nucleic Acids Res.*, 1998, **26**, 280–284.
4. Amann, R. L., Ludwig, W. and Schleifer, K. H., Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 1995, **59**, 143–169.
5. Ogle, J. M., Brodersen, D. E., Clemons, W. M., Tarry, M. J., Carter, A. P. and Ramakrishnan, V., Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science*, 2001, **292**, 902–987.
6. Huysmans, E. and DeWachter, R., Compilation of small ribosomal subunit RNA sequences. *Nucleic Acids Res.*, 1986, **14**, 73–81.
7. Califano, A., SPLASH: structural pattern localization analysis by sequential histograms. *Bioinformatics*, 2000, **16**, 341–357.
8. Gorodkin, J., Heyer, L. and Stormo, G., Finding the most significant common sequences and structure motifs in a set of RNA sequences. *Nucleic Acids Res.*, 1997, **25**, 3724–3732.
9. Kurtz, S. and Schleiermacher, C., REPuter: fast computation of maximal repeats in complete genomes. *BMC Bioinformatics*, 1999, **15**, 426–427.
10. Purohit, H. J., Raje, D. V. and Kapley, A., Identification of signatures and primers specific to genus *Pseudomonas* using mismatched patterns of 16S rDNA sequences. *BMC Bioinformatics*, 2003, **4**, 19.
11. Lasergene software: DNA Star Inc., 1998.
12. Raje, D. V., Purohit, H. J. and Singh, R. N., Distinguishing features of 16S rDNA gene for five dominating bacterial genus observed in bioremediation. *J. Comput. Biol.*, 2002, **9**, 819–829.
13. Dillon, W. R. and Goldstein, M., Multiple discriminant analysis. In *Multivariate Analysis: Methods and Applications*, John Wiley, New York, 1984, pp. 394–415.

ACKNOWLEDGEMENTS. This work was supported by a grant from the Department of Biotechnology, New Delhi and CSIR Network Program (SM0002). We thank Mr G. Balamurgan, Mr M. Ravichandran, Mr A. Padmanabhan, Mr M. Parthiban and Ms S. Sugunadevi, Bharathiar University, Coimbatore, who developed the group-wise database on 16S rRNA as part of their project study.

Received 28 August 2003; revised accepted 10 March 2004

Decline in reactive oxygen species at low light intensity can overcome necrosis barrier in hybrid wheat

Geetanjali Sharma¹, K. V. Prabhu² and Renu Khanna-Chopra^{1,*}

¹Water Technology Centre and ²National Phytotron Facility, Indian Agricultural Research Institute, New Delhi 110 012, India

In an attempt to overcome the barrier to genetic transfer in wheat (*Triticum aestivum*), necrotic hybrids Kalyansona × C306, WL711 × C306, J24 × C306) showing different degrees of necrosis were grown in phytotron along with their parents Kalyansona, WL711, J24 and C306 under lower light intensity (25% of that in the field), long days and higher mean temperatures compared to those existing in the field. Leaf hydrogen peroxide levels, pollen viability and characters from ears of hybrids and their parents were determined. Leaves from plants grown in phytotron revealed lower levels of hydrogen peroxide in comparison to those in the field. Hybrids and their parents grown in phytotron exhibited early flag leaf and ear emergence than those seen in the field. Less severely necrotic hybrids formed viable seeds in the developing ears on the plant itself, whereas the most severely necrotic hybrid of the three formed seeds only after culturing in a medium. Main shoot ear from hybrids was smaller than that of the parents and revealed lower spikelet number, grain number, total grain weight and weight per grain when compared to parents. Thus, lowering of reactive oxygen species content at low light intensity coupled with long days and higher mean temperatures enabled the hybrids to complete their life cycle that resulted in overcoming the necrosis barrier in F₁ progeny in wheat.

HYBRID necrosis is a phenomenon wherein premature and gradual death of leaves and leaf sheaths occurs in certain wheat hybrids¹. It results due to two complementary genes, *Ne1* and *Ne2*. Multiple allelism of these two genes leads to various degrees of necrosis in different crosses. Transfer of desirable traits into well-adapted genetic backgrounds is therefore dependent on the genetic constitution of the donor and recipient parents at the *Ne* loci. A natural mutant selection of wheat, C306 is recognized as the most widely adapted source for drought tolerance in wheat, which does not form viable hybrids with most high yielding wheat lines due to its genetic constitution², *Ne1Ne1ne2ne2*. Drought-resistant *Ne1* carrier wheat cultivar C306 (C) was crossed with high yielding *Ne2* carrier cultivars, Kalyansona (K), WL711 (W) and J24 (J). These exhibited different degrees of hybrid necrosis. Under field conditions, hybrids of Kalyansona × C306 (K × C)

*For correspondence. (e-mail: renu_wtc@rediffmail.com)