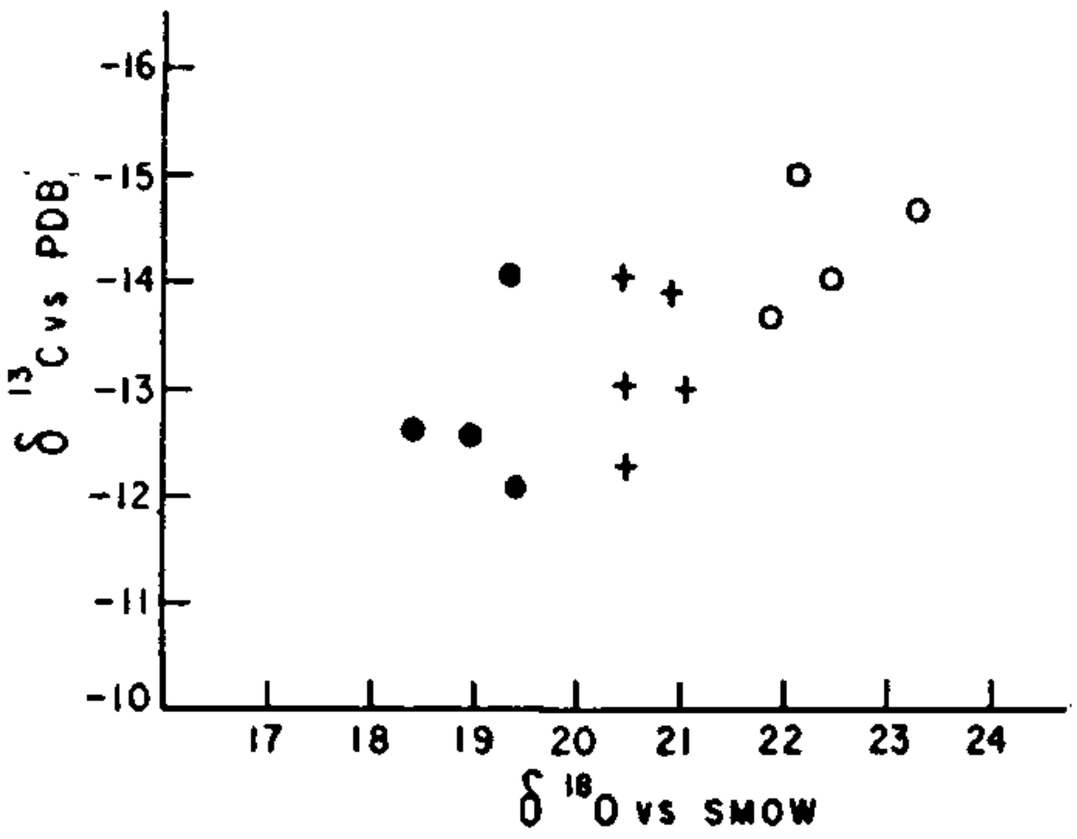

Figure 2. $\delta^{13}C$ [PDB] versus $\delta^{18}O$ [SMOW] diagram, representing '+' and '•' for south Andaman samples while 'o' represents samples from Great-Nicobar

The electron microprobe data (Table 1) and petrographic studies indicate the following features: (a) Relicts of parent rock minerals like chromite, magnetite, quartz and chlorite are present in samples of south Andaman and Great-Nicobar. (b) The carbonate matrix has the texture of clastic mud and constitutes about 70–80% of the rock volume. (c) Sparry calcite crysfals are present in vein structures and dolomite is relatively more abundant in the samples of Great-Nicobar. (d) No evidence of any microfossil.

Stable isotope data suggest significant depletion in $\delta^{13}C$ values (–12.0 to –15.02‰ vs PDB) and relatively low $\delta^{18}O$ values (18.42 to 23.32‰ vs SMOW) (Figure 2). The isotopic data and petrographic studies indicate that these rocks were probably derived by pedogenic transformation of ultramafic or serpentinous rocks under the influence of meteoric waters. Hydrothermal origin can be ruled out because the data show highly depleted $\delta^{13}C$ and somewhat enriched $\delta^{18}O$ values than expected for the hydrothermally derived carbonates.

Brecciation of the parent ultramafic rock could have been followed by cementation process. Breccias can result in this region of shearing by tectonic movement caused by accreting plates. Subsequently the rocks were exposed to subaerial environment and meteoric water may have caused the alteration of ultramafic rock fragments. As suggested earlier[17], cementation process could involve two types of fluids emanating from ultramafic rocks, the predominantly $Ca^{2+}OH^-$ and $Mg^{2+}HCO_3^-$. However, their isotopic signatures were reported to be similar[17]. Our isotopic data are in excellent agreement with both the calcite formed involving such fluids and to vein calcites formed by meteoric waters[13, 17].

In conclusion, the isotopic and mineralogical studies of these rocks suggest that they were formed by brecciation and cementation of ultramafic rocks under the influence of meteoric water at or near surface temperatures.

1 Curray, J. R., Enmel, F. J., Moore, D. G. and Raitt, R W , in *The Ocean Basins and Margins; The Indian Ocean* (eds Nairn, E. M and Stehli, F. G), 1982, pp 399–450
2. Hamilton, W., *U.S. Geol Surv. Prof. Pap.*, 1979, pp. 1070.
3. Karunakaran, C , Powde, M B , Raina, V K , Ray, K. K. and Saha, S. S., in *Proceedings of the 22nd Session of International Geol Congr* , 1964, Part XI, pp 79–100.
4 Jafri, S. H , Balaram, V. and Govil, P. K., *Mar Geol* , 1993, **112**, 291.
5 Karunakaran, C., Ray, K. K. and Saha, S S , *J. Geol. Soc. India*, 1968, **9**, 32
6. Srinivasan, M S and Chatterjee, B K , *J Geol. Soc India*, 1981, **22**, 536
7. Keith, M L. and Weber, J. N , *Geochim Cosmochim Acta*, 1964, **28**, 1280
8 Bathurst, R. G. C., in *Early Diagenesis of Carbonates Sediments* (eds Parker, A. and Sellwood, B. W.), Elsevier, Amsterdam, 1981, pp. 349–378
9 Tan, F C., in *Handbook of Environmental Geochemistry* (eds. Fritz, P. and Fontes, J C), 1989, pp 172–190.
10. Sacket, W M , in *Handbook of Environmental Geochemistry* (eds Fritz, P and Fontes, J. C.), Elsevier, Amsterdam, 1989, pp. 139–167.
11 Craig, H., *Science*, 1961, **133**, 1702.
12 Benson, L. V. and Mathews, R K., *J Sediment. Petrol.*, 1971, **41**, 1018.
13 Larson, S A. and Tullborg, E., *Lithos*, 1984, **17**, 117.
14. Deines, P. and Gold, D. P , *Geochim Cosmochim Acta*, 1973, **37**, 1709
15. Folk, R. L. and McBride, E. F , *Geology*, 1976, **4**, 327.
16 Bonatti, E , *Mar Geol* , 1974, **16**, 83.
17. Barnes, I and O'Neil, J. R., *Geol. Soc Am. Bull* , 1969, **80**, 1947.
18. McCrea, J M., *J Chem Phys.*, 1950, **18**, 849

ACKNOWLEDGEMENTS. We are grateful to Dr K Gopalan for encouragement and support to carry out this study. Thanks are due to Mr R Natarajan and Dr S N. Charan for help in microprobe analyses Mr D. J. Patil's assistance in isotopic measurements is thankfully acknowledged

Received 21 August 1993, revised accepted 15 December 1993

A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes

A. Nandy

Computer Division, Indian Institute of Chemical Biology, Calcutta 700 032, India

A novel graphical approach is proposed for representing DNA sequences in a two-dimensional cartesian

co-ordinate system. The map of a DNA sequence in this representation serves to highlight relative local abundances of the nucleotides and differentiate between regions with large variations between the nucleotide abundances. Applying this method to conserved gene families such as the globin genes, we find that the maps of different genes appear to have distinctive patterns leading to the possibility of using the patterns for global sequence homology; this also raises the possibility of applying this method to homology search in megabase sequences through pattern recognition techniques. These patterns are also seen to be useful for determining evolutionary changes. Comparison is made with the chaos generator diagram technique of Jeffrey to identify the distinctive features in the current method.

MANY different approaches have been taken to characterize the information content in gene sequences to provide an insight into the role of sequence composition and base distribution. The difficulties in identifying signals and sequence patterns from long sequences written in the normal letter code led Hamori and Ruskin[1] and Hamori[2] to devise a three-dimensional graphical plot of gene sequences. However, the technique requires very strong visual identification signals to pick out significant features; such a plot is difficult to set up and conceptualize, and there has actually been very little practical application. Among other attempts, Lathe and Findlay[3] proposed a line extension format for sequence representation to graphically aid in identifying sequence patterns. Hayashi and Munakata[4] formulated an acoustic method to more easily recognize and memorize significant stretches of DNA sequences. Recently, considerable interest has been generated on global characteristics of DNA sequences by Jeffrey's application of chaos generator techniques[5] to demonstrate chaos patterns for vertebrate sequences. Burma *et al.*[6] have extended the methodology to provide new insights into genome structure. Peng *et al.*[7] have shown using a variation of the random-walk model that several DNA sequences have long-range correlations that are not apparent from a perusal of the letter codes, while Voss[8] analysing over 25000 sequences found from spectral density measurements of individual base positions the existence of long-range fractal correlations and prominent short-range periodicities.

To date, however, the linear alphabetic representation remains the most widely used system for DNA representation and identification of DNA characteristics through homology searches and frequency determinations of poly-nucleotides[9, 10], partly due to simplicity in conceptualization. In this paper we present the methodology of a relatively easy graphical approach to DNA representation that has the advantage of enabling visualization of the nucleotide distribution pattern of the gene sequences and also making it possible to determine areas of approximate visual similarities between various sequences as a rough and ready reckoner of possible sequence homologies; it would also appear to be possible to extend the technique for a rapid search of megabase sequences for sequence homologies. This technique generates two-dimensional graphs that can be related to the *H*-curves of Hamori and Ruskin[1] by a projection of their three-dimensional curve onto two-dimensional planes and performing suitable axes transformations. The present technique, however, differs from their approach in that it essentially plots the differences in pairs of bases along the axes which makes it less sensitive to minor changes in base composition while making patterns arising out of significant changes more readily identifiable. In this paper we present the methodology of this graphical technique, in which we take the globin genes as examples to demonstrate the benefits of studying DNA sequences in this system.

In our approach to representation of DNA sequences, we construct a symmetric purine-pyrimidine graph on the cartesian co-ordinate system, with purines on the x-axis (A along the negative x-axis, G along the positive x-axis) and pyrimidines on the y-axis (T in the negative y-direction, C in the positive) and plot the sequence structure using the following rule: Starting from the origin, we plot a point for each succeeding nucleotide with a displacement depending on the type of nucleotide: for adenine the displacement is one unit in the negative x-direction; for cytosine, it is one unit up from the current position in the positive y-direction; for guanine the displacement is one unit to the right, and for thymine it is one unit down. Then the total displacement up to say the nth nucleotide on the sequence will be represented by a point on the purine-pyrimidine plot where the displacement along the x-axis represents the difference of total number of G's and the total number of A's up to this nth nucleotide, and the displacement along the y-axis represents the cumulative difference between the C's and T's up to that nucleotide. As n goes from 1 to N, the total sequence length, there will be a progression of points on the map generating a graphical representation of the pattern of composition of a DNA sequence which can be expected to differ for different sequences.

For any given sequence, there are, *a priori*, two fixed points: (i) the origin will be at the origin of the purine–pyrimidine plot by choice, (ii) the end point will be located at $(G - A, C - T)$, where A, C, G, T represent the total number of each of the nucleotides in the sequence under consideration. For a given base distribution on the sequence we can then expect that

(i) An ordered sequence of continuous A's, then C's followed by G's and T's will produce a quadrangular pattern starting at (0, 0) and ending at $(G - A, C - T)$;
(ii) For a sequence with homogeneous distribution of A, C, G, T, i.e. that any randomly chosen segment of the

sequence comprises four bases in the same ratio as the overall total, the plot generated will be a uniform line leading from the origin to the end point.

(iii) In a randomly distributed system with the same base composition where the presence of any nucleotide is independent of any succeeding or preceding nucleotide, the points will trace a path on the map as in a classical two-dimensional random-walk problem, terminating finally at the same end point.

(iv) In the case of an actual DNA sequence, the graphical pattern will be characteristic of the distribution of bases on the sequence and will wind in various directions depending on the variations in the relative abundances of the bases. A study of the macroscopic picture therefore can be of use in identifying large segments of the sequence with predominantly different base characteristics.

(v) Homologous sequences will exhibit similarities in shape and structure in this representation where minor differences in base distributions may not necessarily distort the overall pattern.

(vi) At a more detailed level, the fine structure of such a map can also serve to highlight the relative local abundances of nucleotides within the sequence from a study of the twists and turns along the sequence plot. Thus e.g. (Figure 1):

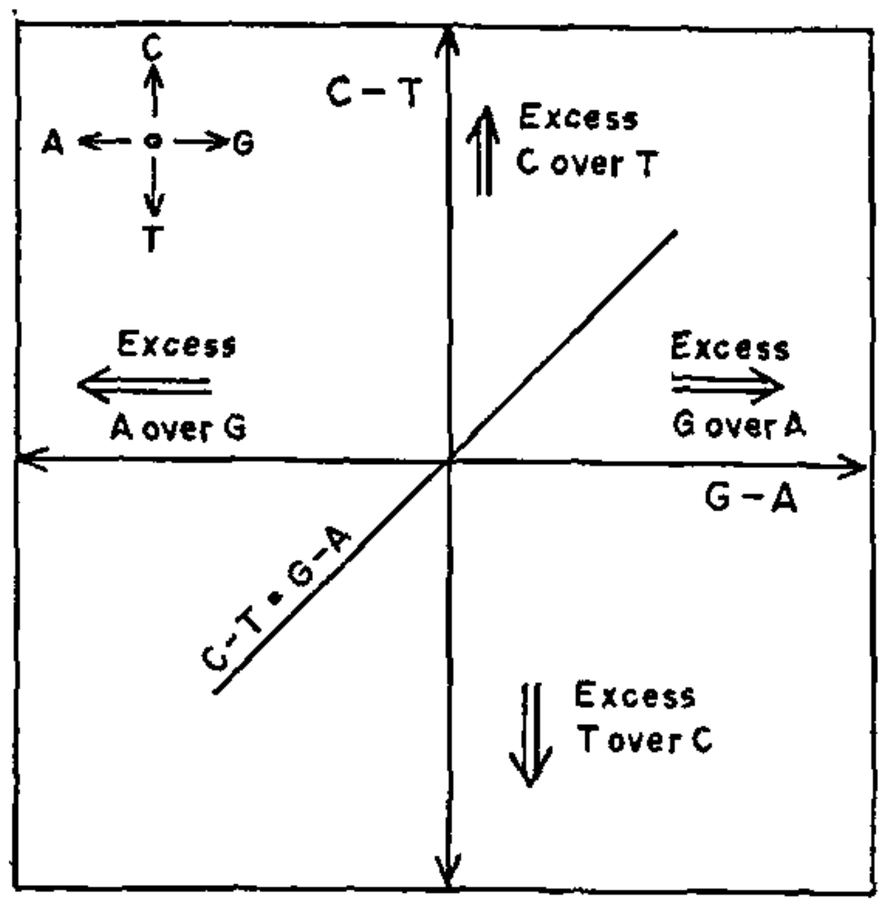

Figure 1. The co-ordinate system for the new graphical representation The bold arrows show direction of progression of the plot for different local base counts All maps are taken to start from the origin and progression is 1 step up if the next base is *C*, 1 step down for *T*, 1 step to the right for a *G* and 1 step left if the next base is *A* Thus progression of the map in different directions shows whether the local base composition is of an excess of *G* over *A* (movement in positive *x*-direction) or excess *T* over *C* (movement in negative *y*-direction), etc

(a) A segment along *x*-axis implies $C = T$ over that stretch, a segment along *y*-axis implies $A = G$.

(b) If, for instance, the map progresses generally horizontally in the positive *x*-direction, this implies a segment rich in *G* compared to *A*, and progression in the negative *y* direction will represent a *T*-rich segmen as compared to *C*, while progression along a diagonal in e.g. the third quadrant (*x*, *y* negative) represents a relatively *AT*-rich segment with the equality $C - T = G - A$ or $C + A = G + T$.

(c) A repeated sequence segment will appear as a repeated structure in the plot, except for repeats like *AG*, *CT*, etc. where the plot will oscillate between two neighbouring points only by construction (but see item (vii) below. This feature can be used to mask certain dinucleotides or chosen combinations of nucleotides and highlight others by a choice of step sizes and assignment of bases to axes.)

(d) Complementary strands will occur as mirror images of each other reflected along the diagonal from the upper right corner to the lower left.

(vii) It is also possible to interchange the *G* and *C* axes to highlight different base combinations such as $A - G$-rich and $C - T$-rich segments. In this case the second quadrant represents purine-rich segments, the fourth quadrant represents pyrimidine-rich segments and other parameters change accordingly. The following discussions however are based mainly on the co-ordinate system with *A*, *G* along the *x*-axis and *C*, *T* along the *y*-axis since this places purines along the horizontal and pyrimidines along the vertical axes and makes comparisons easier. This is also important in evolutionary studies where the transition types of substitutions in (*A*, *G*) or in (*C*, *T*) are expected to be much more prevalent than the purine to pyrimidine type of transversion substitutions (see ref. 11 and references therein): the changes in shape of the distribution will be minimal for transition type of changes in this representation since the *A* to *G* and *G* to *A* substitutions will tend to cancel each other in the global perspective, and so also for the $C - T$ substitutions.

Mapping a DNA sequence on this system will generate a graph according to the distribution of the bases along the DNA, and can therefore be expected to bear resemblances in relevant segments to other sequences that share significant homologies. In fact, application of this technique to a class of highly conserved gene family reveals characteristic patterns that are also conserved between species. Figures 2*a–e* show the different globin genes from the human β globin cluster on chromosome 11. Each of the β globin genes shows a characteristic pattern of a small *G*-rich segment followed by a large *A*, *T*-rich segment. The ε genes have a predominantly *A*, *T*-rich segment interspersed with *G*, *T*-rich parts that bends sharply from the top *G*- and *T*-rich part. The γ-globins have excess *T* over *C* while the *A* and *G*-rich segments are finely balanced to produce a pattern that zigzags down the *T*-axis from the *G*-rich head. The δ and β globins being evolutionarily closely related (see e.g Lewin[12]) have closely similar patterns; again there is an

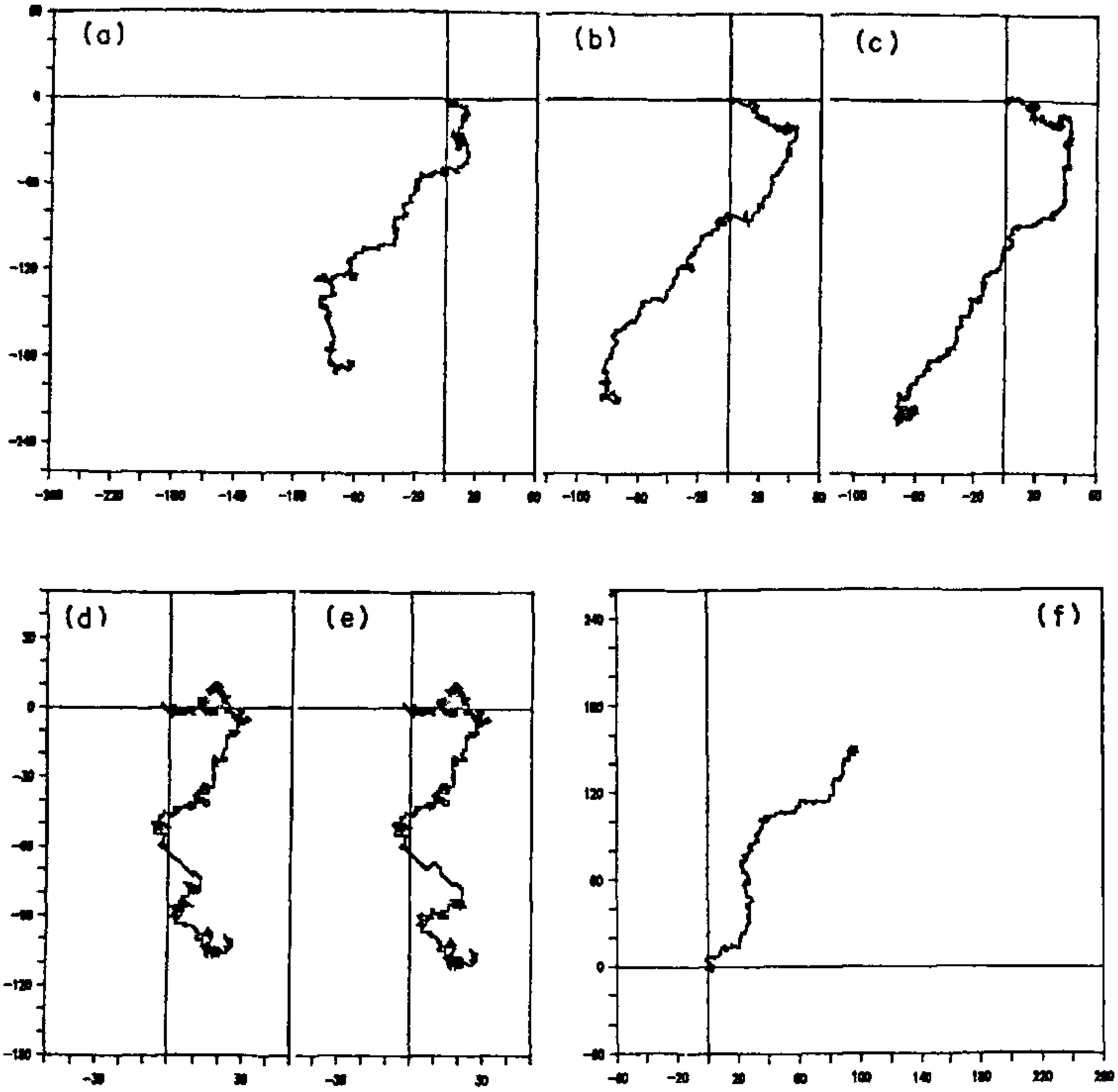

Figure 2. Maps of the globin gene family The maps are for the sequences from exon 1 to exon 3 inclusive of introns 1 and 2 Maps are of (*a*) ε-globin, (*b*) δ-globin, (*c*) β-globin, (*d*) *A*-γ-globin and (*e*) *G*-γ-globin genes from the human β-globin cluster on chromosome 11 (*f*) The orangutan α-globin gene map

essentially *G*-rich part, following which both have a small *T*-rich segment, with minor characteristic differences between the β and the d globins, then a long *AT*-rich segment like the ε globins but comparatively smoother. Figure 2 also shows a characteristic α-globin representation where the graph shows strong *C*, *G*-richness. The ζ-globins (not shown) produce a pattern qualitatively similar to the α globins but with long stretches induced by the excess intronic components.

The similarity of the β globin gene patterns and the large intron content of the globin genes make it possible to compare their evolutionary history in our representation. In sequence regions that play specific roles as in gene expression or bending around histones to form nucleosomes, evolutionary pressures have constrained nucleotide sequences[9], but traces of evolutionary changes may be more evident in other parts of the sequence. We may thus expect patterns of conserved gene sequences to be generally similar, differing only due to effects of evolutionary changes retained in the nonspecific regions. All the β-globin type genes are seen to have a characteristic shape of a flat horizontal *G*-rich head comprised of the first two exons and the first intron (epsilon excepted), followed by a long *A*, *T*-rich tail comprising the long second intron and the third exon, which may be interpreted as being due to the common evolutionary origin of these genes. In the case of the two γ-globins where the protein coding regions differ only by a single amino acid, the variations are small and separation may have been very recent. By the same token, however, the α-globin genes show a distinctly different shape, due perhaps to the ancient divergence of the α- and β-branches of the globin genes.

We find these characteristic patterns to be conserved for a wide variety of mammalian globins (Table 1), leading to the possibility that an unknown variety of

Table 1. Mammalian globin gene of different species with visually similar global patterns for the same type of gene All comparisons are for the genes without flanking regions The EMBL DNA sequence identification numbers are given for reference

Globin type	Species with visually similar graph	EMBL ID
Alpha	Goat adult α-1-globin	CHHBAI
	Horse BI α-1 globin	ECHBA22
	Mouse α-globin	MMAGL1
	Rhesus monkey α-globin	MMHBA
	Orangutan adult α-1-globin	PPHBA02
Zeta	Chimpanzee ζ-globin	PTAZGLO
	Human ζ-globin on α-globin cluster on chromosome 16	HSHBA1
	Horse ζ-globin (BI allele)	ECZGL1
	Horse ζ-globin (BII allele)	ECZGL2
Beta	Human β-globin cluster on chr 11	HSHBB
	Macaque cynomolgus β-globin	MCBGLOG
	Lemur (brown) β-globin	LMHBB
Gamma-globins	Human G-γ globin, A-γ-globin from β-globin cluster on chromosome 11	HSHBB
	Gorilla fetal A-γ-globin	GGAGGLOG
Epsilon	Human β-globin cluster on chr 11	HSHBB
	Bovine ε (2) β-globin	BTEBGL2
	Bovine ε (4) β-globin	BTEBGL4
	Goat ε-1 β-globin	CHEBGL1
Delta	Human β-globin cluster on chr 11	HSHBB
	Tarsius syrichta δ-globin	TSHBD
	Spider monkey δ-globin	AGHBD

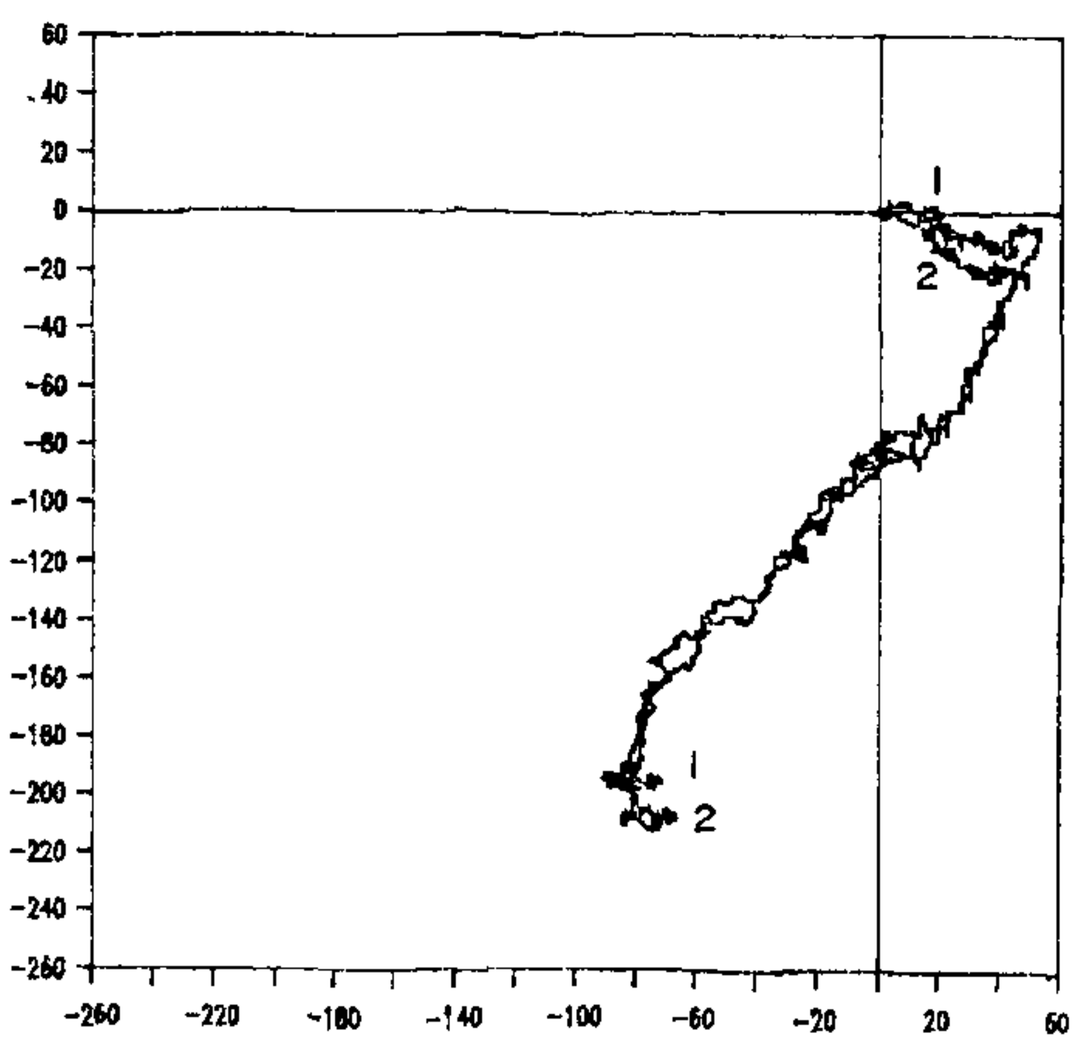

Figure 3. Superposition of DNA sequence maps of δ-globin genes of human and African spider monkey

globin gene can be classified based on its graphical representation in the present scheme. This can be illustrated by observing that a superposition of the δ-globin gene sequences from the human β-globin cluster on chromosome 11 and that of the spider monkey (*A. geoffroyi*) shows almost complete overlap (Figure 3). The small differences that exist between the representations of the same conserved gene between various species can be expected to have arisen from evolutionary developments and this feature can be exploited to give a qualitative idea of gene divergence and phylogenetic relationships. The superposition of α-globin genes of goat, horse, rhesus monkey and orangutan (Figure 4) shows the closest homology between rhesus monkey and orangutan, next is the goat, with horse showing marked differences. In evolutionary time scale, the horse family including tapir, pig, camel, etc. is believed to have diverged in the early palaeocene period about 60 million years ago, whereas the divergence of the bovids, monkeys and apes are comparatively more recent. We also note that the four maps appear to be identical in the *N*-terminal proximal region. This may be due to the fact that this region is covered by the first two exons along with the short first intron; since evolutionary divergence is expected to be restrained in exons and less so in introns, shifts in the maps can be expected to be more pronounced in the region containing the large, second intron.

The characteristic shapes of different globins also bring up the possibility of using such features for a rapid visual search of homologous zones in megabase sequences that are being rapidly accumulated and whose analyses present formidable problems[13,14]. Using fast digital computers one can contemplate moving a window along the sequence plotted using the present technique to identify possible homologies with sequences with specific shapes. A similar search along the

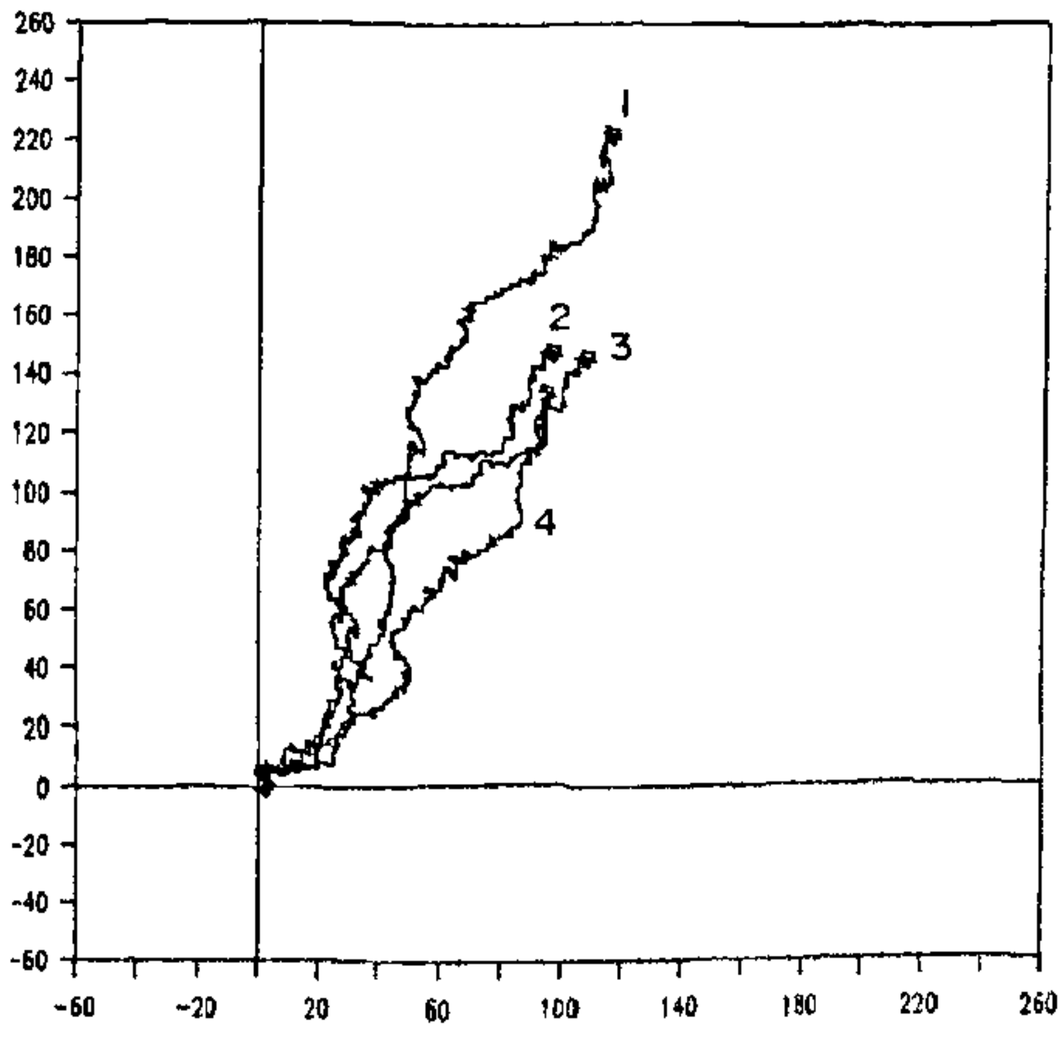

Figure 4. Superposition of DNA sequence maps of α-globin genes excluding the flanking regions of 1 horse 2 orangutan, 3 rhesus monkey, 4 goat

human β-globin section on chromosome 11 does reveal only two locations that appear to have the characteristic vertical shape of the γ-globins; the other globins are relatively harder to identify, but can be made possible using sophisticated pattern recognition techniques.

We may note here that when matching the patterns of two homologous sequences we would be looking for a close rather than an exact match since the base distribution along the sequences will be close but not necessarily identical, and this may reduce Hamori and Ruskin's[1] H-curve method's effectiveness. In the Hamori–Ruskin model, the three-dimensional curve obtained is open and stretched according to the length of the DNA and therefore makes pattern recognition and matching that much more difficult. In the current technique since we plot differences between pairs of nucleotides along the two axes, the pattern is compressed while at the same time it can show sharp variations in the slope of the graph as the differences between the bases swing through zero values. On the other hand, minor variations in base composition cause only slight changes to the overall pattern, making the task of its recognition comparatively easier, while major variations in base composition over a significant length of the DNA produce distinctly different patterns as exemplified in the case of the globin genes discussed above.

It is instructive also to compare the plots of gene sequences in the present model with the diagrams obtained in the chaos generator (CGR) model of Jeffrey[5]. CGR diagrams are not very informative for few hundreds or thousands of bases, unlike the current plotting technique which can be used for tens of bases to megabases of any length to provide characteristic features according to the scale of the plot. CGRs of gene families such as the globins show similar depletion zone patterns but there are no noticeable differences between the different types of β-globin genes. However, where there are sharply contrasting regions of nucleotide abundances, both the CGR and the present model can be used to indicate the differences, with the present model having the advantage of visually identifying regions of differing abundances directly.

In summary, we may state that we have presented a novel graphical method for visual display of nucleotide distribution patterns in DNA sequences and have demonstrated the usefulness of the method by taking the globin gene family as an example; study of other gene families is in progress. From the studies presented here we may conclude that this method can be of use in identifying groups of nucleotides that may be related evolutionarily as in the case of the β- and δ-globins. The method can also be used for visual global homology which we have seen exists in the case of highly conserved sequence families such as the globin genes; where such characteristic shapes can be identified this leads to the possibility of rapid homology search on megabase sequences that are fast becoming part of the global sequence databases. While this graphical technique bears some similarity to the H-curves of Hamori and Ruskin[1], it is a two-dimensional graph that is easier to construct and visualize; at the same time, the major advantage accruing out of the present method is that since this plots the differences of pairs of bases along the two axes, it is very sensitive to significant changes in the base composition. However, in representing a 4-independent parameter object like a gene sequence on a two-dimensional plane there will be shortcomings; in this case parallel maps with *C*, *G* and *G*, *T* axes interchanged may lead to more information. In this context we may mention the extensions to the chaos generator diagrams proposed by Burma *et al.*[6] and submit that the current method is simple to implement, extends sequence analyses techniques in different ways and readily complements the various methods in existence for understanding the nature of the distribution of nucleotides along a gene sequence.

1. Hamori, E. and Ruskin, J., *J. Biol. Chem.*, 1983, **258**, 1318–1327.
2. Hamori, E., *Nature*, 1985, **314**, 585.
3. Lathe, R. and Findlay, R., *Nature*, 1984, **311**, 610.
4. Hayashi, K. and Munakata, N., *Nature*, 1984, **310**, 96.
5. Jeffrey, H. J., *Nucl. Acids Res.*, 1990, **18**, 2163–2170.
6. Burma, P. K., Raj, A., Deb, J. K. and Brahmachari, S., *J. Biosci.*, 1992, **17**, 395–411.
7. Peng, C-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M. and Stanley, H. E., *Nature*, 1992, **356**, 168–170.
8. Voss, R., *Phys. Rev. Lett.*, 1992, **68**, 3805–3808.
9. Nussinov, R., *Comput. Appl. Biosci.*, 1991a, **7**, 287–293.
10. Nussinov, R., *Comput. Appl. Biosci.*, 1991b, **7**, 295–299.
11. Gojobori, T., Moriyama, E. N. and Kimura, M., *Methods Enzymol.*, 1990, **183**, 531–550.
12. Lewin, B., *Genes*, Wiley Eastern Ltd, New Delhi, Ch. 21, 1986.
13. Maddox, J., *Nature*, 1992, **357**, 13.
14. Erickson, D., *Sci. Am.*, 1992 July, **266**, 128.

ACKNOWLEDGEMENTS. I thank Dr S. Adhya for many helpful discussions and a careful reading of the manuscript and the Centre for Cellular and Molecular Biology, Hyderabad for access to the EMBL DNA database. I thank the referees for helpful suggestions.

Received 1 December 1992, revised accepted 10 December 1993

Comparative study of some methods of oxidation of plant materials for elemental analysis

C. L. Arora and M. S. Bajwa

Department of Soils, Punjab Agricultural University, Ludhiana 141 004, India

Four different methods were used to dissolve eleven different plant materials for simultaneous estimation