Descriptors based on information theory for numerical characterization of DNA sequences

R. Jayalakshmi[1], R. Natarajan[2,*], M. Vivekanandan[3] and N. Ganapathy Subramanian[4]

[1]Department of Biotechnology, School of Life Sciences, Bharathidasan University, Tiruchirappalli 620 024, India
[2]Centre for Mathematical Sciences, Arunapuram, Pala 686 574, India, and Department of Chemical Engineering, Lakehead University, Thunder Bay, Ontario, Canada P7B 5E1
[3]Vivekananda Arts and Science College for Women, Thiruchengode, Namakal 637 205, India
[4]Department of Industrial Engineering, Texas Tech University, Lubbock, TX 79409, USA

Descriptors based on information content (IC) are introduced to characterize nucleotide sequences. The descriptors are an extension of Shannon IC and are denoted as IC_r, where $r = 1, 2, \ldots, n$ corresponding to the probability distribution of DNA strings of length 1, 2, etc. Sequence IC (SIC_r) and complementary IC ($CSIC_r$) are also introduced. IC saturates by reaching a maximum after a few orders and the order (string length) corresponding to the maximum IC value for a given sequence depends on the length of the DNA sequence. Effectiveness of the new descriptors in comparing similarity of DNA sequences was evaluated by performing phylogenetic analyses on first exons of 14 β-globin genes, and complete coding sequences of 20 β-globin genes. Dendrograms obtained using the descriptors were comparable to the classification of organisms according to the evolutionary tree. IC_r, SIC_r and $CSIC_r$ could be calculated without much demand for computation time even for very long DNA sequences.

Keywords: Alignment-free sequence comparison, β-globin genes, DNA-descriptors, information theory, numerical characterization, phylogenetic analysis.

SHANNON, an electrical engineer, introduced the concept of information content (IC) in 1948 (ref. 1), and IC for probability distribution is:

$$IC = -\sum p_i \log_2 p_i. \quad (1)$$

The usual convention is whenever $p_i = 0$, $(0)\log_2(0) = 0$. The choice of logarithm base is usually two, and therefore, information is given in bits. Information theory is applied to DNA sequences to find the statistical patterns of nucleotide distribution for studying the genomic evolution[2–4]. Mostly these methods used maximum entropy measures (MEM), and Steinbach[5] used MEM to study the kinetics of binding of carbon monoxide to myoglobin. By applying information theory to the four-letter representation of DNA, its linguistics, IC is obtained based on the probability distribution of four nitrogenous bases for both coding and or non-coding regions[6]. IC derived based on nucleotide distribution, triplet or the amino acid distribution in coding and or non-coding regions of the DNA sequences were used for comparative genome analysis[7–12]. Nandy[13] studied the evolutionary changes of base distributions in gene sequences using Shannon IC and the base proportion index. The base proportion index was calculated using the formula $(C + G - T - A)/(C + G + A + T)$, where A, T, C and G represent the total number of each of the DNA bases in a sequence. These approaches on sequence comparison may be regarded as extensions of the alignment-free comparison methods pioneered by Blaisdell[14–16]. Almeida[17] reviewed some of the alignment-free comparison methods and the statistical procedures to calculate similarity from the numerical descriptors. The composition vector tree (CVTree) method introduced by Qi and co-workers[18–20] is an alignment-free sequence comparison method based on the frequency of amino acid K-strings, oligopeptide content, of complete proteomes. In the realm of chemoinformatics, Basak *et al.*[21] used information theory for the numerical characterization of molecular structures. In their approach, they used the probability distribution of equivalent classes based on the neighbourhood complexity of atoms. A numerical characterization of DNA sequences using the K-string approach and its application in comparison of DNA sequences without alignment is presented here.

The accession numbers of sequences retrieved from the DNA databank (http://www.ncbi.nlm.nih.gov/) along with sequence lengths are given in Table 1. The sequences used in the present study vary widely in length (number of nucleotides) from 93 to 17,000. Several authors[22,23] used a set of first exons of β-globin genes for testing similarity or dissimilarity using invariants developed for the numerical characterization of DNA. In order to compare the results obtained using the approach explained here with that of others, the first exon of the β-globin gene of 14 species was used. In addition to this, complete coding sequence (CDS) of 20 β-globin genes were also used to validate the phylogenetic analysis resulting from the alignment-free sequence comparison method proposed. A computer program was developed in Microsoft Visual Basic 6.0® to calculate the information theoretic indices. The program can calculate the new DNA descriptors in batch mode for a set of sequences. SPSS 16.0® was used to perform statistical analyses such as principal component analysis (PCA) and hierarchical clustering.

Calculation of numerical descriptors based on information theory is explained in the following lines. A DNA sequence is usually represented using the four letters corresponding to the four DNA bases. In order to calculate Shannon IC, the four bases are considered as equivalent

*For correspondence. (e-mail: rnataraj@lakeheadu.ca)

Table 1. Sequences used in the present study with their accession numbers and species names

Accession number	Common name	Binomial nomenclature	Sequence length	
			First exon	Coding sequence
AY260740	Human	*Homo sapiens*	92	444
DQ350619	Goat	*Capra hircus*	86	438
J03642	Opossum	*Didelphis virginiana*	92	444
M15734	Lemur (black)	*Eulemur macaco*	92	444
L17432	Chicken	*Gallus gallus*	92	444
NM_008220	Mouse	*Mus musculus*	92	444
M17084	Rat	*Rattus narvegicus*	92	444
K03256	Rabbit	*Oryctolagus cuniculus*	92	444
X61109	Gorilla	*Gorilla gorilla*	93	364
X00376	Cattle	*Bos tarus*	86	438
Y00501	Frog	*Xenopus tropicalis*	92	444
M61740	Thick-tailed bush baby	*Otolemur crassicaudatus*	92	444
AB063101	Common carp	*Cyprinus carpio*	90	444
DQ352471	Sheep	*Ovis aries*	86	438
AY279119	Titi monkey	*Callicebus torquatus*	–	441
AY279118	Geldi's mormost monkey	*Callimico goeldii*	–	444
AY279117	Black spider monkey	*Ateles paniscus*	–	444
AY279116	Common squirrel monkey	*Saimiri sciureus*	–	444
AY279115	Brown capuchian monkey	*Cebus apella*	–	444
AY279114	Common woolly monkey	*Lagothix lagotricha*	–	444
AY279113	Azara's night monkey	*Aotus azarai*	–	444
U00096	*Escherichia coli* (16s rRNA)	*Escherichia coli*	–	1542
DQ861999	A/chicken/Sudan/2115-12/2006*	N/A	–	1707
AY563044	Soya bean (PEPC)†	*Glycine max*	–	2901
X54252	Round worm (Genome)	*Caenorhabditis elegans*	–	13794
AJ508398	Short-tailed opossum (Genome)	*Monodelphis domestica*	–	17079

†PEPC: Phosphoenolpyruvate carboxylase (C4 plant).
*Avian influenza-A virus, subtype *H5N1* – Haemagglutinin gene (HA).

classes and probability distribution for a hypothetical sequence *atggctatatg* is: $a = 3/12$; $t = 4/12$; $g = 3/12$; $c = 2/12$. IC can be calculated using eq. (1).

$$p_a = \frac{3}{12}\log_2\frac{3}{12};\quad p_t = \frac{4}{12}\log_2\frac{4}{12};$$

$$p_g = \frac{3}{12}\log_2\frac{3}{12};\quad p_c = \frac{2}{12}\log_2\frac{2}{12}$$

$$\mathrm{IC} = -\left[\left(\frac{3}{12}\log_2\frac{3}{12}\right)+\left(\frac{4}{12}\log_2\frac{4}{12}\right)+\left(\frac{3}{12}\log_2\frac{3}{12}\right)+\left(\frac{2}{12}\log_2\frac{2}{12}\right)\right]$$

$$\mathrm{IC} = 2 \times 0.5000 + 0.5284 + 0.4308 = 1.9592.$$

Shannon IC thus obtained may be called the first order IC (IC_1). This is extended and generalized to calculate IC contents of different orders (IC_r) using doublet-codes, triplet-codes, etc. for defining the equivalent classes. Equivalent class considerations of higher order IC are:

order 1: *a t g g c t a t a t c g*

order 2: *at tg gg gc ct ta at ta at tc cg*

order 3: *atg tgg ggc gct cta tat ata tat atc tcg*

order 4: *atgg tggc ggct gcta ctat tata atat tatc atcg*

It may be noted that open reading frame (ORF) is not considered to split a given sequence into number of equivalent classes. If ORF is considered, then some bases may be left out without being include in the calculation of a given order. In the given example, if ORF is used up to IC_4 there is no problem because 12 is divisible by 2, 3 and 4. However, while calculating IC_5 the last two bases have to be left out without being included in the calculation and this amounts to loss of information. On the contrary, if ORF is not considered and the combination of bases is computed recursively from each base, no information will be lost. Thus, for a DNA sequence of length ℓ, number of possible combinations of r-string is $\ell - r + 1$. For the hypothetical sequence given here, $IC_2 = 2.8454$; $IC_3 = 3.1219$; $IC_4 = 3.1699$.

Maximum number of equivalent classes possible for singlet-nucleotides is 4, whereas that for doublet-nucleotides is 16 and thus the maximum value of a combination of r-string is 4^r. This gives the maximum value (upper bound) of the IC_r values as

$$\mathrm{IC}_r \text{ max value } = -\log_2\frac{1}{4^{(r)}} = \log_2 4^r = r \times 2.$$

The idea was extended to the calculation of IC-codons (IC_{codon}) and IC-amino acids (IC_{AA}). For the calculation of IC_{codon}, the equivalent classes and their probabilities were computed based on the triplet codons obtained using ORF. In the case of IC_{AA}, the number of equivalent classes and their distribution are calculated using amino acids that would be formed during the protein synthesis. Calculations for IC_{codon} and IC_{AA} for the hypothetical sequence *atggctatatcg* are illustrated here.

Possible triplets or codons are: *atg gct ata tcg*

Amino acids corresponding to the codons are: methionine (M), alanine (A), isoleucine (I) and serine (S). In the present case $IC_{codon} = IC_{AA} = 2.000$.

However, IC_{codon} need not be equal to IC_{AA} because in case of some amino acids, more than one triplet codon encode their synthesis. For example, isoleucine is encoded from three codons, viz. *att*, *atc* and *ata*.

Two additional measures of information contents, viz. sequence IC (SIC_r) and complementary sequence IC ($CSIC_r$) are also proposed where:

$$SIC_r = \frac{IC_r}{\log_r n_r} \tag{2}$$

$$CSIC_r = \log_2 n_r - IC_r \tag{3}$$

These calculations are similar to ICs introduced by Basak[21] for molecular structures.

IC of order 1–11 (IC_r; $r = 1$ to 11) and the corresponding sequence IC (SIC_r; $r = 1$ to 11) and complementary sequence IC ($CSIC_r$; $r = 1$ to 11) were calculated for the seven sequences varying in length from 93 to 17,079. The seven sequences used were: X61109 *Gorilla gorilla*; AY260740 *Homo sapiens*; U00096 *Escherichia coli*; DQ861999 H5N1 (host: chicken); AY563044 *Glycine max*; X54252 *Caenorhabditis elegans* and AJ508398 *Monodelphis domestica*. Computation of all the descriptors was fast even for the genome corresponding to *Monodelphis domestica* ($n = 17{,}079$) and took less than a minute in a PC with Intel Core2DUO (E4500) 2.20 MHz processor and 1 GB RAM. Maximum value of IC_r increases with increase in order (Max IC = $r \times 2$) due to the increase in the number of equivalent classes of string length r. However, the probability of each equivalent class depends on the sequence length ℓ. Hence, the value of IC saturates after reaching a particular order and beyond this the IC remains unaltered. In the case of SIC, once saturation is reached the maximum value of 1 is attained and beyond this SIC does not increase. From the relation of CSIC it is clear that CSIC reaches zero when SIC is 1 and therefore, CSIC reaches 0 once IC_r reaches the maximum possible value for a given sequence. The trends in IC_r, SIC_r and $CSIC_r$ for sequences of different lengths are given in Figure 1. Hence, the maximum order to which IC should be calculated depends on the sequence length.

Application of the new descriptors for alignment-free comparison of DNA sequences was tested initially using the first exon of 14 β-globin genes and this set of

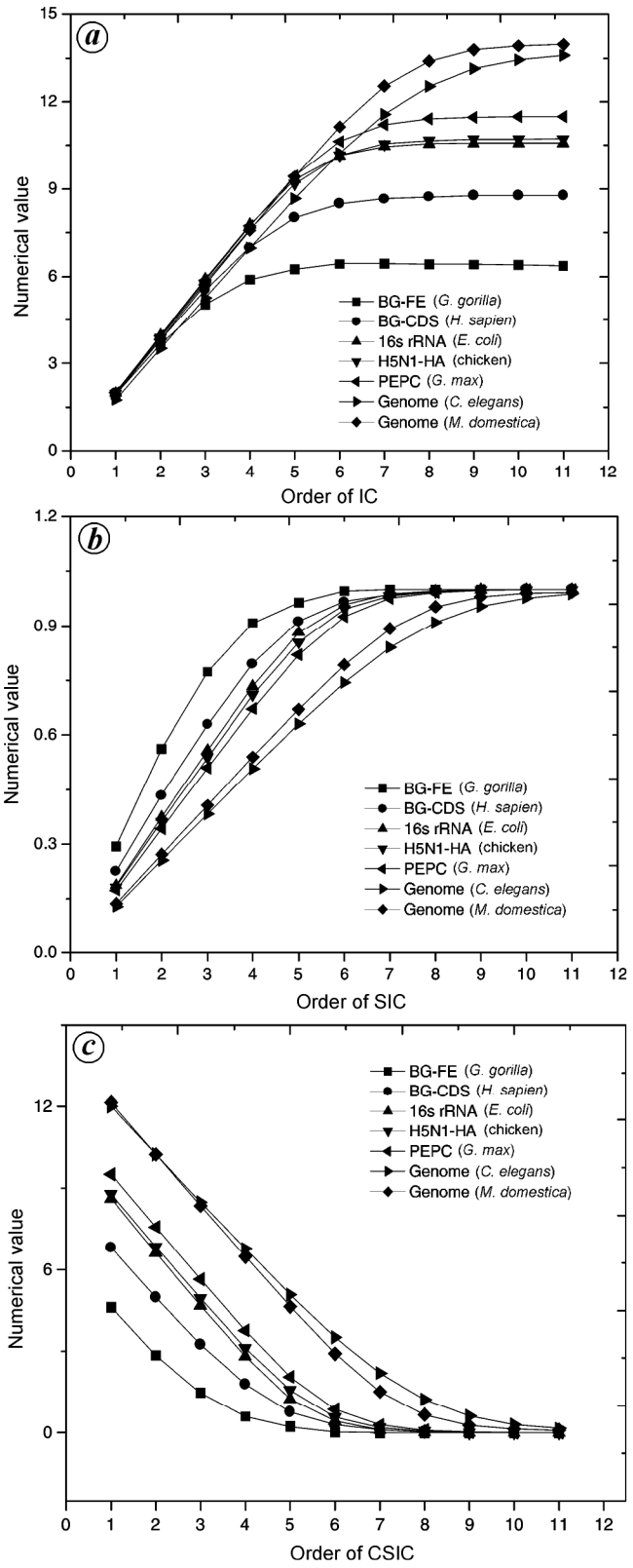

Figure 1. Saturation of information content after reaching a particular order.

sequences was chosen to compare our results with those obtained by other authors[21,22]. IC_r, SIC_r and $CSIC_r$ of orders 1 to 6, and IC_{AA} and IC_{codon} were computed for the 14 sequences. The descriptors thus calculated were cross correlated (computed descriptors and cross correlation can be obtained from the corresponding author). As some descriptors have cross correlations greater than 0.9, they encode redundant information. In situations like these, PCA is a very effective tool for reduction of dimensionality of the data and selection of descriptors that are not highly mutually correlated. PCA is scale dependent and the results are affected by the descriptors, whose scales are several orders higher in magnitude than that of others. Hence, the IC values, input descriptors (D), were scaled using the transformation ($\log_e(D + 1)$). PCA with modified descriptors extracted four principal components with eigenvalue > 1 and they explained 96.5% of the variance among the data (Table 2). Scores of the principal components can be used as the orthogonal variables for phylogenetic analysis or four descriptors that are minimally inter-correlated can be selected using PCA. PCA yields a set of eigenvalues and the corresponding eigenvectors. The elements of eigenvectors can be interpreted as correlation indices; they reflect the degree of association between the ith variable and the jth principal component. The objective of the interpretation is to select one variable to represent each eigenvector; this subset of variables will have low inter-correlation because the eigenvectors are mutually independent, and therefore there should be low inter-correlation between variables that are associated with different eigenvectors. Hopefully, the variables will be selected with the thought of maximizing the variation between the predictor variables selected as the subset and criterion variable. Although the correlation between the predictor variables in the subset and the criterion variables cannot be greater than the correlation between all of the predictor variables and the criterion variable, the difference between the two correlations should not be statistically significant. Thus, PCA should yield a subset of predictor variables that reduces both the data collection and the inter-correlation. The four descriptors, viz. IC_3, IC_{AA}, IC_6 and SIC_1 were selected because each one of them has the highest matrix loading in the respective principal component (IC_3: 0.967, PC1; IC_6: 0.933, PC2; IC_{AA}: 0.606, PC3; SIC_1: 0.376, PC4).

Table 2. Eigenvalues and percentage variance of principal components for β-globin 1st exon (four components with eigenvalue > 1 were extracted)

Component	Eigenvalue	Percentage of variance	Cumulative percentage variance
1	12.581	62.904	62.904
2	4.030	20.150	83.054
3	1.594	7.970	91.024
4	1.096	5.479	96.504

Dendrogram for the 14 sequences (first exon of β-globin gene) were drawn using the linkage between groups. The four predictors for the construction of dendrograms were given in three different formats namely, the principal component scores; the log-transformed predictors ($\log_e(D + 1)$), and MAD-normalized (median absolute deviation) predictors. MAD-normalization is given here:

$$\text{The MAD-normalization is } \frac{X - \text{Median}(X_i)}{\text{MAD}} \tag{4}$$

$$\text{MAD} = \text{Median}_i(|X - \text{Median}(X_i)|) \tag{5}$$

The best dendrogram was obtained when MAD-normalized predictors were used. The dendrogram thus obtained (Figure 2) shows the association of the evolutionarily similar species in the same node. The associations of goat, sheep and cattle; human and gorilla; mouse, thick-tailed bush baby, rabbit and rat indicate that the descriptors-based approach followed identifies the degree of similarity of sequences. Phylogenetic analysis thus carried out is far superior to those reported earlier[23]. Most of the numerical characterization methods attempted to convert the letter sequence into a graph and then to a descriptor. Hence, they may be regarded as secondary descriptors and suffer loss of information during the conversion of a sequence to a graph. In the present approach, the four-letter sequence is directly converted into descriptors using information theory. Hence, they may be regarded as primary descriptors and suffer minimum loss of information in their computation.

In order to test the applicability of the sequence comparison without alignment to larger sequences, CDS of 20 β-globin genes were used. Statistical procedures explained for the first exon of β-globin genes were followed for the 20 CDS. Similar to the study on 14 first exon β-globin genes, PCA extracted four components for 20 CDS also and they explained 98.5% of the total variance (Table 3). The four predictors, viz. IC_4, IC_{AA}, IC_6

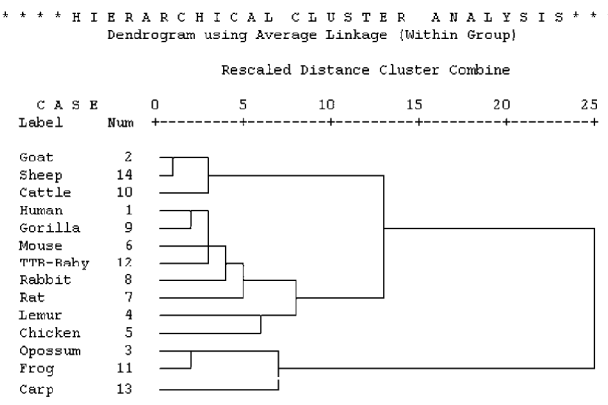

Figure 2. Phylogenetic analysis of 1st exon of β-globin genes (14) using MAD-normalized predictors selected using PCA.

Table 3. Eigenvalues and percentage variance of principal components for β-globin CDS (four components with eigenvalue > 1 were extracted)

Component	Eigenvalue	Percentage of variance	Cumulative percentage variance
1	13.98	69.88	69.88
2	2.83	14.14	84.02
3	1.56	7.80	81.82
4	1.34	6.68	98.50

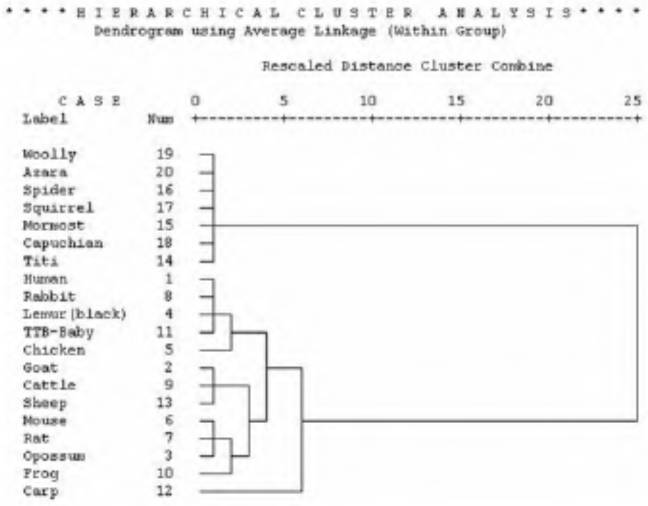

Figure 3. Phylogenetic analysis of complete coding region of 20 β-globin genes using MAD-normalized predictors selected using PCA.

and $CSIC_1$ were selected based on the loadings in the component matrix (IC_4: 0.981, PC1; IC_{AA}: 0.913, PC2; IC_6: 0.527, PC3; $CSIC_1$: 0.815, PC4). In order to perform the hierarchical cluster analysis (phylogenetic analysis) of CDS, the four predictors were used either after log-transformation or after MAD-normalization. The best result was obtained when the predictors selected were MAD-normalized. This was inconsistent with the study discussed here for the 14 first exon β-globin genes. The results of phylogenetic analysis for the CDS are comparable to the classification according to species origin and evolution. However, in the case of phylogenetic analysis of CDS, the gorilla had to be dropped as its sequence length (364) differs very much from the others (438, 441 or 444). This may be considered as one of the limitations of the current approach. The dendrogram obtained (Figure 3) for the 20 species using CDS had four major groups whereas carp and frog were separated as out groups. The organisms that were grouped together in the four clades are: Group 1: the seven monkeys; Group 2: human, rabbit, lemur and thick-tailed bush baby (TTB-baby); Group 3: goat, cattle and sheep; Group 4: mouse, rat and opossum. Each of these groups is connected to an evolutionarily related ancestor. The dendrogram obtained for the first exons of the 14 β-globin genes and that obtained for the CDS did not match exactly. In the dendrogram obtained using CDS, the group containing opossum, rat and mouse (group 4) is separated from the more related group 2 (human, rabbit, lemur and TTB-baby). This might be due to some horizontal transfer of genes, which might not have been reflected when first exons that form only part of the coding sequences were considered. The approach explained here is similar to the *L-tuple* approach discussed in the review[17]. However, in this communication a multidimensional classification approach using a set of descriptors and the use of related statistical procedures such as data reduction (PCA) and normalization using MAD were studied. The evolutionary relationships of genes obtained in the form of dendrograms using the approach described are quite satisfactory as related species were found in the same nodes. The descriptor-based approach explained here used a large pool of descriptors and was expected to encode maximum information with minimum information loss. Moreover, the descriptors explained were shown to be calculated for sequences of lengths 13,794 and 17,079 in a batch process and the computation of descriptors was reasonably fast. Hence, it is possible to extend the descriptor-based alignment-free comparison approach for comparison of whole genomes. However, the present program is developed to run on a desktop personal computer so that sequence comparison could be performed offline without alignment. When we tried to extend this approach for complete genomes containing 100,000 or more nucleotides, we encountered problems due to memory overflow. Owing to this limitation, comparative study with the performance of CVTree could not be included. Moreover, CVTree works well for unicellular organisms such as bacteria and fungi. The alignment-free sequence comparison method proposed can be used to study the geographic origin and spread of infectious diseases[24] such as avian and swine flu.

1. Shannon, C. E., A mathematical theory of communication. *Bell. Syst. Tech. J.*, 1948, **27**, 379–423.
2. Cosmi, C., Cuomo, V., Ragosta, M. and Macchiato, M. F., Characterization of nucleotidic sequences using maximum entropy techniques. *J. Theor. Biol.*, 1990, **147**, 423–432.
3. Frappat, L., Minichini, C., Sciarrino, A. and Sorba, P., Universality and Shannon entropy of codon usage. *Phys. Rev. E.*, 2003, **68**, 061910.
4. Ragosta, M., Cosmi, C., Cuomo, V. and Macchiato, M., An application of maximum entropy techniques to determine homogeneous sets of nucleotidic sequences. *J. Theor. Biol.*, 1992, **155**, 129–136.
5. Steinbach, P. J., Two-dimensional distributions of activation enthalpy and entropy from kinetics by the maximum entropy method. *Biophys. J.*, 1996, **70**, 1521–1528.
6. Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C. K., Simons, M. and Stanley, H. E., Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys. Rev. E*, 1995, **52**, 2939–2950.
7. Dehnert, M., Helm, W. E. and Hntt, M. T., Information theory reveals large-scale synchronisation of statistical correlations in eukaryote genomes. *Gene*, 2005, **345**, 81–90.

8. García, J. A., Alvarez, S., Flores, A., Govezensky, T., Bobadilla, J. R. and José, M. V., Statistical analysis of the distribution of amino acids in *Borrelia burgdorferi* genome under different genetic codes. *Physica A*, 2004, **342**, 288–293.
9. Holste, D., Grosse, I., Beirer, S., Schieg, P. and Herzel, H., Repeats and correlations in human DNA sequences. *Phys. Rev. E*, 2003, **67**, 061913.
10. Li, C. and Wang, J., Relative entropy of DNA and its application. *Physica A*, 2005, **347**, 465–471.
11. Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P. and Zhang, H., An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 2001, **17**, 149–154.
12. Schneider, T. D., Stormo, G. D., Gold, L. and Ehrenfeucht, A., Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 1986, **188**, 415–431.
13. Nandy, A., Investigations on evolutionary changes in base distributions in gene sequences. *Internet J. Mol. Design*, 2002, **1**, 545–548.
14. Blaisdell, B. E., A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Natl. Acad. Sci. USA*, 1986, **83**, 5155–5159.
15. Blaisdell, B. E., Average values of dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. *J. Mol. Evol.*, 1989, **29**, 538–547.
16. Blaisdell, B. E., Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. *J. Mol. Evol.*, 1989, **29**, 526–537.
17. Almeida, J. S. and Vinga, S., Alignment-free sequence comparison – a review. *Bioinformatics*, 2003, **19**, 513–523.
18. Qi, J., Luo, H. and Hao, B., CVTree: a phylogenetic tree construction tool based on whole genomes. *Nucleic Acids Res.*, 2004, **32**, W45–W47.
19. Qi, J., Wang, B. and Hao, B., Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.*, 2004, **58**, 1–11.
20. Chu, K. H., Qi, J., You, Z. and Anh, V., Origin and phylogeny of chloroplast revealed by a simple correlation analysis of complete genomes. *Mol. Biol. Evol.*, 2004, **21**, 200–206.
21. Basak, S. C., Information theoretic indices of neighborhood complexity and their applications. In *Topological Indices and Related Descriptors in QSAR and QSPR* (eds Devillers, J. and Balaban, A. T.), Gordon and Breach Science, The Netherlands, 1999, pp. 563–593.
22. Nandy, A., Harle, M. and Basak, S. C., Mathematical descriptors of DNA sequences: development and applications. *ARKIVOC*, 2006, **ix**, 211–238.
23. Zupan, J. and Randić, M., Algorithm for coding DNA sequences into 'spectrum-like' and 'zigzag' representations. *J. Chem. Inf. Model.*, 2005, **45**, 309–313.
24. Ghosh, A., Nandy, A., Nandy, P., Gute, B. D. and Subhash Basak S. C., Computational study of dispersion and extent of mutated and duplicated sequences of the H5N1 influenza neuraminidase over the period 1997–2008. *J. Chem. Inf. Model.*, 2009, **49**, 2627–2638.

ACKNOWLEDGEMENTS. R.N. thanks the Department of Science and Technology, New Delhi, India, for financial assistance (No. SR/S4/MS: 479/07). We also thank Mr T. M. Anbazhagan, Inflexion Technology, Bangalore, for his help in developing the computer program.

Received 25 March 2009; revised accepted 21 June 2010

Platinum group elements in basic and ultrabasic rocks around Madawara, Bundelkhand Massif, Central India

S. P. Singh[1,*], V. Balaram[2], M. Satyanarayanan[2], K. V. Anjaiah[2] and Aditya Kharia[1]

[1]Department of Geology, Bundelkhand University, Jhansi 284 001, India
[2]National Geophysical Research Institute (CSIR), Uppal Road, Hyderabad 500 007, India

The southern part of the Bundelkhand Massif shows a series of lensoidal bodies of undeformed and unmetamorphosed ultramafics, associated with gabbro/diorite and intrusive into the Bundelkhand Gneissic Complex (BnGC). The ultramafics exposed around Madawara town is characterized by high MgO (26–46 wt%), and low SiO_2 (42–46 wt%), TiO_2 (< 1 wt%) and Al_2O_3 (< 1 wt%). The concentrations of Cr and Ni are high, ranging from 3000 to 6000 ppm and 1500 to 4000 ppm respectively. Geochemical studies of platinum group elements (PGEs) in fertile ultramafic magma developed at great depths showed depleted to undepleted and undersaturated to saturated conditions around Madawara. The rare earth elements patterns show three distinct trends, viz. strongly depleted Eu anomaly, Eu positive and a flat trend. The high values of ΣPGE (~ 700 ppb) especially Ir PGE (IPGE) and Pt (> 100 ppb) suggest that the Madawara ultramafic complex could be a potential PGE prospect. Detailed exploration studies including bore hole drilling should be taken up.

Keywords: Depletion, fertile mantle, fractionation, partial melting, platinum group elements prospect.

THE platinum group of elements (PGEs) (Ru, Rh, Pd, Os, Ir and Pt) is strongly siderophile and chalcophile in character and is a sensitive indicator of partial melting, crystal fractionation and S-saturation conditions of magma[1–3]. PGEs are an economically important group of elements and there are only few localities with PGE deposits in the world, viz. Norilsk–Talnakh in USSR; Stillwater Complex in USA; Bushveld Complex in South Africa, and Sudbury structure and its associated sulphide deposits in Canada[1,4–8]. It has been suggested that sulphur-saturated melts may have relatively high PGE, especially the Pd group of elements[9], but the possibility of PGE deposits in the sulphur undersaturated conditions cannot be ignored[4]. The latter view is mainly related to the oxide facies, which contains high platinum. The oxide facies appears in most of the magma before the appearance of the sulphide facies minerals. High values of PGE have been reported from many mafic and ultramafic terrains of dif-

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