Simple Numerical Descriptor for Quantifying Effect of Toxic Substances on DNA Sequences

A. Nandy[†] and S. C. Basak*

Natural Resources Research Institute, University of Minnesota, 5013 Miller Trunk Highway,
Duluth, Minnesota 55811

Received September 9, 1999

Many chemicals are known to be toxic to living organisms, inducing mutations and deletions at the chromosomal and genetic level. One of the tasks in risk assessment of genotoxic chemicals is to devise a simple numerical descriptor which may be used to quantify the relationship between chemical dose and the effect on the genetic sequences. We have developed numerical descriptors to characterize different DNA sequences which are especially useful in sequence comparisons. These descriptors have been developed from a graphical representational technique that enables easy visualization of changes in base distributions arising from evolutionary or other effects. In this paper we propose a scheme to use these descriptors as a label to help quantify the potential risk hazard of chemicals inducing mutations and deletions in DNA sequences.

INTRODUCTION

The deleterious effects of many chemicals and newly synthesized compounds on human and environmental health is of serious concern. Many of these chemicals are known to pass through cell barriers and cause mutations and deletions in DNAs. Recent studies have demonstrated how many common chemicals cause such effects: exposure to common environmental chemicals such as nitropyrenes present in diesel exhausts cause mutations and homologous recombinations in DNAs leading to carcinogenesis;[1,2] some polycyclic aromatic hydrocarbons from coal burning for industry and home heating form DNA adducts that have been shown to act as transplacental carcinogens and developmental toxicants[3] or induce mutations at the GC and the AT base pairs of the *hrpt* genes;[4] other chemicals such as ethylnitrosourea and ethyl methanesulfonates have been shown to induce mostly transition types of mutations in DNAs leading to chromosomal aberrations.[5] A carbonyl compound, acrolein, present in the environment as commonly used industrial chemicals, natural products, environmental contaminants and products of endogenous metabolism in human beings, has been found to cause mutations and intrastrand cross-links between guanine residues,[6] and similar effects of many other compounds are known in the literature (see, e.g., refs 7 and 8). DNA damage is also induced by excesses of heavy metals such as Rh[9] and Cu(II),[10,11] which preferentially induce depletion of guanine residues. Table 1 gives a brief list of some of the data available in recent literature on effects of chemical substances on DNA sequences.

One of the prime tasks in risk assessment of these and other chemicals and ions is to define one or more numerical descriptors of the chemical dose and the measured effect. Much of the data to date, however, consist of measures of types of mutations and deletions observed in specific genes at various levels of chemical dosages, and much of it is order of magnitude indications of genetic risk.[8] While some chemicals would induce mutations and deletions at sites with specific base pair combinations, others could lead to oxidative damages and mutations at random at intragenic and intergenic segments including point mutations and small deletions. Techniques of unbiased measures of such alterations in a DNA sequence from a set of numerical descriptors would be essential in assessing, in a universal and standard manner, the risk potential of such chemicals and form a vital link in integrating pharmacokinetics and mutational studies.

In this paper we outline such a measure arising from descriptors of DNA sequences of any specified length and show that small changes due to random point mutations or deletions in such sequences can be quantified for scaling purposes. It has developed out of a technique for graphical representation of DNA sequences but can now be done rapidly and accurately using computer programs bypassing the graphical stage altogether.

METHOD

The fundamental basis of our proposed quantitative descriptor is analysis of base distribution in a sequence by taking a running account of compositional differences in pairs of bases, e.g. intra-purines and intra-pyrimidines, as we read down the sequence from the 5′- to the 3′-end. This is most easily visualized in terms of a two-dimensional graphical representation described below. Since the method depends on small differences between the numbers of bases present in the sequences, it is very sensitive to small changes in base composition and distribution patterns.

The method of representing DNA sequences graphically using a two-dimensional Cartesian coordinate system has been explained elsewhere.[12,13] The shapes of these DNA

* To whom correspondence should be addressed. E-mail: sbasak@wyle.nrri.umn.edu.

[†] On leave from: Indian Institute of Chemical Biology, 4 Raja S C Mullick Road, Calcutta 700 032, India. E-mail: anandy43@yahoo.com.

10.1021/ci990117a CCC: $19.00 © 2000 American Chemical Society
Published on Web 06/23/2000

Table 1. Effects of Different Chemicals on DNA Sequences (Recent Studies)[a]

		mutation composition			
			substitutions (%)		
chemical	DNA sample	deletions (%)	transitions	transversions	refs and remarks
acrolein	SupF gene	24	21	55 (∼GC to TA)	4
ethylnitrosourea	lacZ	5	43 (∼GC to AT)	52 (∼AT to TA)	5
ethylmethanosulfonate	lacZ	8	74 (∼GC to AT)	18 (GC to TA)	5
heavy metals—Rh	oligomeric DNA duplexes	100 (5′-G deleted in 5′GG-3′ doublets)			9 (long-range electron transfer)
5-nitroimidazoles	Bacteroides fragilis			100 (majority C to G, CG to AT)	7
1,3-butadiene	Various—in mice, rat, humans				8 genetic hazard exists at permitted concns mutation data not available
polycyclic aromatic hydrocarbons	hprt gene	∼25		∼55	4

[a] Notes: The "∼GC to AT" implies that the majority transitions are of the GC to AT type, etc. Acrolein is one of the a,b-unsaturated carbonyl compounds present in the environment. Nitroimidazoles, Metronidazole and dimetridazole are used in treatment of intraabdominal, pulminory, and brain abscesses and other diseases. 1,3-Butadiene is widely used in the petroleum industry.

graphs depend on the base distribution in the sequence. The plot of a typical representation is generated by moving one step in the positive *x*-direction for a guanine (G) in the sequence, the negative *x*-direction for an adenosine (A), the positive *y*-direction for a cytosine (C), and the negative *y*-direction for a thymine (T), the succession of such steps producing a graphical shape characteristic of the sequence. This essentially plots the progressive differences in the instantaneous individual totals of guanine and adenosine along the *x*-axis (i.e. $n_G - n_A$) and of cytosine and thymine along the *y*-axis (i.e., $n_C - n_T$); two other sets of axes can be similarly defined for a complete representation, but we use the one described here as the default axes system. We have shown[12] that for conserved genes such plots are shape similar thereby making identification of a new sequence of the gene family possible rapidly and easily by visual inspection alone; elsewhere we have shown that one can read off base preferences and local abundances directly from the shape of these graphs[14] or identify coding and noncoding regions of the sequences.[15] Changes in base distribution and composition induce changes in the visual plots of the DNA sequences; for the same genes for different species we have noticed systematic drifts in the sequence pattern which have been attributed to evolutionary changes.[16]

Differences in the plots of a family of genes can be quantitatively assessed.[17] This method consists essentially of defining a set of moments of the graph points around the origin of the plot. In the first order we define quantities $\mu_1(x)$ and $\mu_1(y)$ which are the sum of the *x*- and *y*-coordinate values of each point averaged by the total number of points in the distribution. One can then define a graph radius for each plot

$$g_R = [(\mu_1(x))^2 + (\mu_1(y))^2]^{1/2}$$

and correspondingly a distance measure between two graphs:

$$d(s,s') = [(\mu_1(x) - \mu_1(x'))^2 + (\mu_1(y) - \mu_1(y'))^2]^{1/2}$$

where *s* and *s*′ represent the two graphs. We have observed[17] that small differences in DNA sequences arising out of base mutations and deletions manifest themselves in observable changes in g_R and *d*. We propose to use the g_R as one numerical descriptor of a sequence and deviation from g_R, Δg_R, as a measure of the changes in a sequence as a consequence of genotoxic effects of chemicals. For greater precision, one could also use a set of $\mu_1(x)$, $\mu_1(y)$, and g_R as numerical descriptors of a DNA sequence.

RESULTS AND DISCUSSIONS

As a preliminary exploration of this technique, we have used the complete human β globin gene sequence (from the HSHBB sequence of the EMBL DNA database rel 31), inclusive of the introns and exons, as the control sequence. This has a total of 1424 bases consisting of 444 (31.2%) bases in the coding regions and 980 (68.8%) in the noncoding part. Plot 1 in Figure 1a shows the graphical representation of this gene starting from exon 1 through introns 1 and 2 to exon 3. Intron 1 is G-rich and shows a horizontal shift to the right; intron 2 has a T-rich part in the initial stages, represented on our graph as an almost vertical drop, and then a long stretch of TA repeats that move the graph generally in a southwesterly direction ending with exon 3 represented as a small region of a dense cluster of points. Exons 1 and 2 are also represented as (less dense) clusters of points unlike the long runs of the introns; we have elsewhere[15] exploited this characteristic difference between intron and exon representations as a means for determining protein coding regions in new sequences.

With regard to the problem at hand, we simulated the effects of Rh and Cu(II) toxicity on a DNA by performing programmatically random deletions of several guanines in the sample sequence. Such deletions will tend to alter the $\mu_1(x)$ in the default representation with a bias toward negative *x*-values (because of a higher percentage of adenosines in the altered sequence) while leaving the $\mu_1(y)$ unchanged and will consequently alter the graph radius. Graphically, the reductions in the number of guanines will make the plot shift to the left in the default reference frame, and the shift will be greater for a greater degree of deletions effected. This is evident visually from a low value of 5% deletions in the complete sequence (Figure 1a). The values of Δg_R for different numbers of guanine deletions are plotted in Figure 1b.

In the case of mutations, the graph radius is quite sensitive to small changes and to specific base positions affected. A

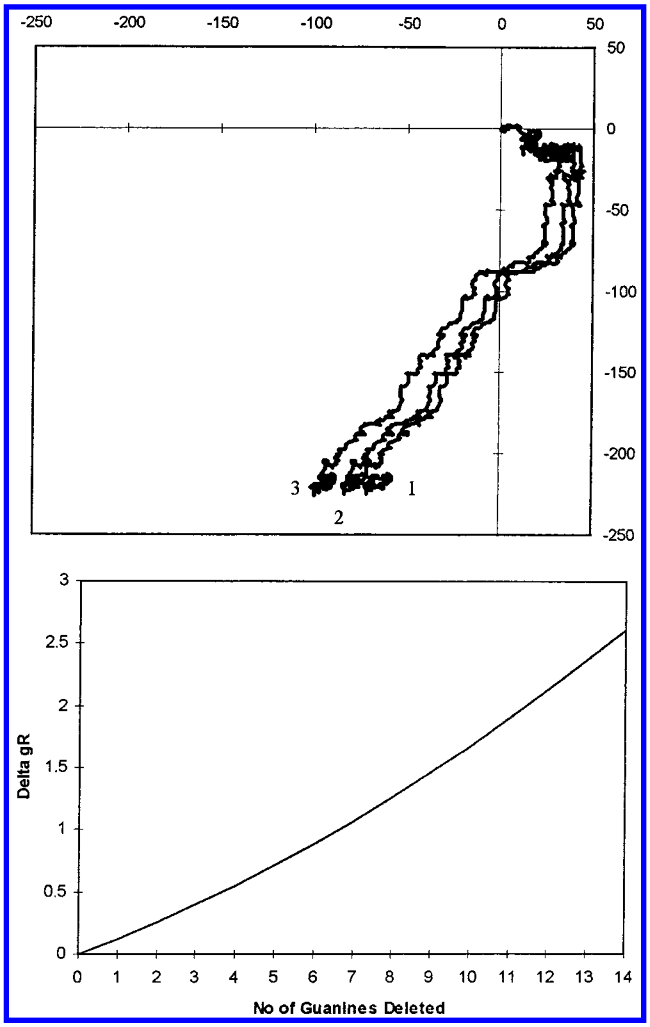

Figure 1. (a, top) Human β globin gene and its model modifications plotted in the two-dimensional representation system. Axes as explained in the text. Plot 1 is for the normal human β globin gene complete with exons and introns. Plot 2 is for the same gene with 5% random depletion in guanine residues. Plot 3 is the same gene with 10% depletion in guanine bases. (b, bottom) Plot of changes in graph radius (Δg_R) against guanine number for deletion of guanines in positions 1−14.

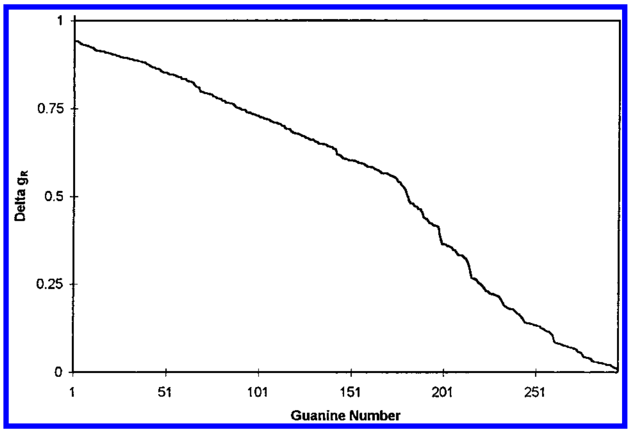

Figure 2. Plot of changes in graph radius (Δg_R) against the guanine number for mutation (G to C) of single guanine to cytosine at various positions.

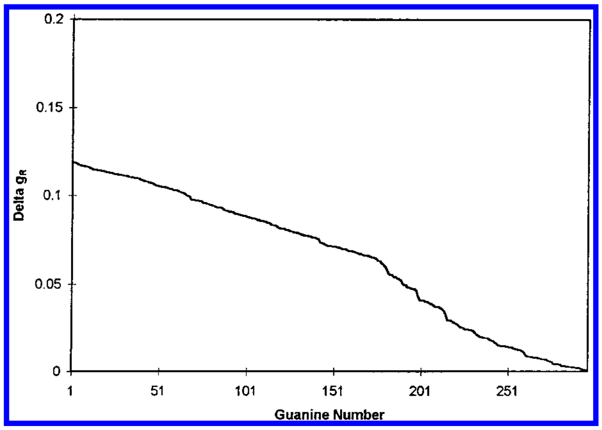

Figure 3. Plot of changes in graph radius (Δg_R) against the guanine number for mutation (G to A) of single guanine to adenosine at various positions.

mutation in the first position, reading from the 5′-end, effects the maximum change while a mutation in the last base has the least effect; this is easily understood from the fact that the change in the first position alters the coordinate value of each subsequent point all the way to the last base and thus affects the value of μ_1 much more than would be the case for mutation of the last base. (The argument remains true when read from the 3′-end and as long as one is consistent in one's convention; here we use the common convention of reading from the 5′-end.) Figure 2 shows Δg_R plotted against the guanine number for mutation of one guanine to cytosine in each position of the guanine in the complete sequence of the human β globin gene. It is interesting to note that Δg_R has a unique value for each position, and, as can be expected, the value goes down to almost zero for the last guanine (the kink seen in the curve occurs at a large gap between successive guanines). Mutations of guanine to adenosine will produce smaller amount of changes in Δg_R since this is a change occurring exclusively in the x-direction and lead to a contraction or expansion of the general curve, whereas the previous mutations produced a change in direction of the plot in our default axes system. Figure 3 shows the variation of Δg_R with guanine number for mutation of a single guanine to adenosine. We have noted elsewhere[18] that Δg_R can therefore be used as quantitative descriptors for indexing single nucleotide polymorphic genes.

In the present case of indexing as a measure of risk assessment for toxicity, the sensitivity of Δg_R raises the question of adequate knowledge of the exact location of the toxic damage. Since any random mutation or deletion could arise from the genotoxic effects, it would be preferable to average over the entire range of values of Δg_R over the chosen DNA segment to arrive at an acceptable index value for purposes of comparative assessment. For example, for the case of mutation of one guanine to adenosine, the average value of Δg_R is 0.064 while that for the case of guanine to cytosine is 0.537, and an index for the two types of causative chemicals that produce just this level of mutation could be written in thousandths as 64 or 537.

In the case of multiple base mutations also this trend of different values of Δg_R for mutations at different base positions will hold true: e.g., mutations of three guanines to cytosines will cause maximum deviation from g_R when the mutations occur in the first three guanines (Δg_R = 2.789 76 compared to the unmutated gene), and the change will be least when the mutations take place in the last three guanines (Δg_R = 0.031 41 compared to the unmutated gene). Multiple mutations will therefore create a field of values for

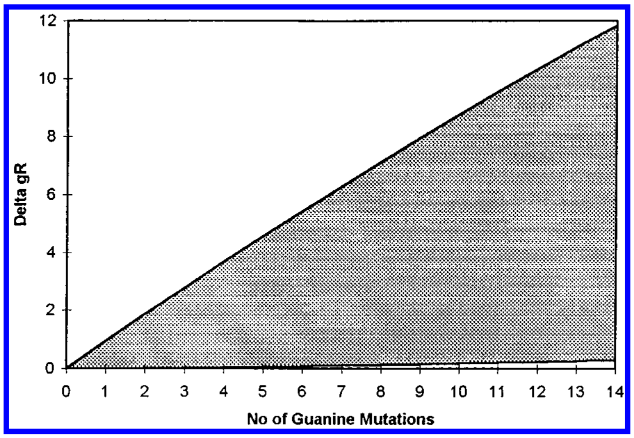

Figure 4. Plot of changes in graph radius (Δg_R) against the number of guanines mutated for G to C mutations. The upper line is the highest value and the bottom line the lowest value of Δg_R for a given number of mutations.

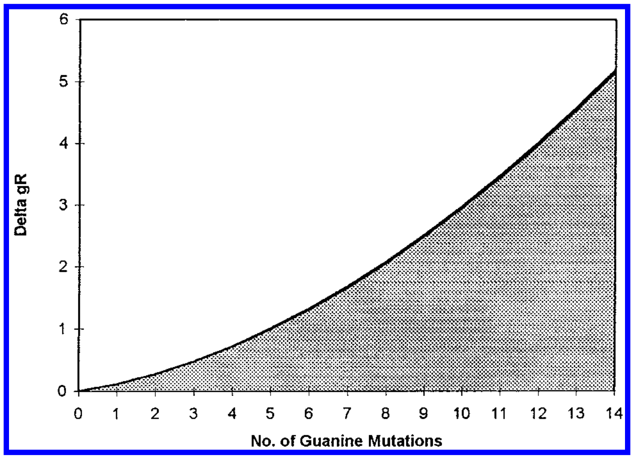

Figure 5. Plot of changes in graph radius (Δg_R) against the number of guanines mutated for G to A mutations. The upper line is the highest value and the bottom line (not visible on this graph range) the lowest value of Δg_R for a given number of mutations.

Δg_R, the maximum for a specific number of mutations being the value realized from mutations in the first of those bases. These maximum values will thus form an envelope as shown in Figure 4, and a lower bound will be created by the minimum values of Δg_R; all values between these two boundaries will relate to the different bases in the sequence that can be mutated for any specified number of mutations. Figure 5 shows similar data for the various degrees of G to A possible mutations.

While we have discussed these effects on the hypothesis of G to C and G to A mutations, these results can be generalized to mutations in any base combinations also. For example, in the case of genetic mutations induced by high levels of toxic chemicals where more than one base can be affected, e.g. mutations of the type GC to AT shown in Figure 6, which occurs in the case of the ethylnitrosourea and ethyl methanosulfonate types of compounds, one can determine the value of Δg_R from a sample sequence exposed to a standard dosage and use that value as an index for measuring the least number of mutations that can be generated from such a number. From Figure 6, for example, it can be seen that a Δg_R of 10 implies that the number of corresponding mutations will be five GC doublets or more.

Thus an experimental measure of Δg_R for a given dose of a toxic chemical can lead to association of an index value that will permit easy gradation of chemicals on levels of toxicity. Each toxin will affect a DNA in its own unique ways: some by deleting a preferred base, some by causing random mutations in one or more preferred bases. The usefulness of an index such as Δg_R arises from associating one number with each dosage level of each chemical providing an easy path to associating risk with dosage without having to enumerate which base and how many are mutated or deleted. Δg_R thus enables a normalization approach to risk assessment of genotoxic chemicals where no other such measure is readily available.

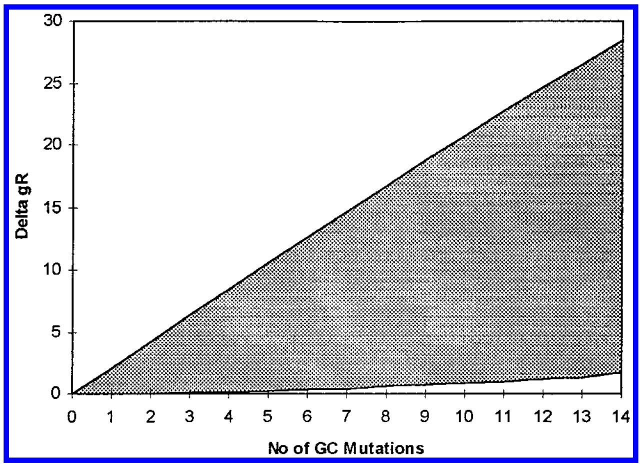

Figure 6. Plot of changes in graph radius (Δg_R) against the number of GC to AT mutations. The upper line is the highest value and the bottom line the lowest value of Δg_R for a given number of mutations.

Note that the method is not dependent on the type of DNA sequence used; while for some chemicals specific DNA segments will be susceptible to damage, for others damages can occur in any of the coding or noncoding segments as for example in case of Cu(II) and Rh induced damages. The indexing can be done for all these cases with respect to any standard sequence segment chosen.

CONCLUSION

Thus we see that the concept of graph radius in a graphical representation of a DNA sequence can be extended to make quantitative estimation of any changes in the sequence. This observation indicates that it is possible to consider using such quantitation as an index of the intensity of the effects in the case of changes arising out of effects of genotoxic chemicals. As of now, however, we are restricted by the paucity of experimental data to only indicating the use of Δg_R as a possible index; experimental work so far are generally in the nature of inquiries into the kinds of changes induced in DNA sequences by genotoxic chemicals, whereas building up a quantitative index would require controlled experiments relating dosage and the extent of DNA damage.

Our work has shown that Δg_R, the change in g_R, is a very sensitive indicator of changes in a sequence arising out of base depletions and mutations. This provides us therefore a numerical descriptor of the alterations in base distribution and composition of DNA sequences and can be used to compare with any standard or control sequence. Δg_R, therefore, averaged over its relevant range of values, can be used as a numerical descriptor to provide a measure of the genotoxic effects of chemicals such as oxidants such as Rh and Cu(II), or acrolein, ethyl methanosulfonate, or any other chemicals whose effect on DNA sequences can occur in a random manner and therefore can affect any part of the DNA whether coding or noncoding. In the case of genotoxins that affect specific genes or base combinations, the Δg_R will need to be calculated for those specific genes only, and there the sensitivity of the measure can be exploited to provide an indicator of the genotoxicity level of the chemicals.

ACKNOWLEDGMENT

This paper is contribution number 268 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by Grant F49620-96-1-0330 from the United States Airforce. We acknowledge gratefully the help received from Ms. Denise Mills for literature search and Mr. Brian Gute for the computer setups. A.N. would also like to thank all members of the NRRI for the kind hospitality and courtesy shown during his stay at the Institute there.

REFERENCES AND NOTES

(1) Beland, F. A. How do chemicals in diesel engine exhaust damage DNA? Health Effects Institute Research Report No. 46; Health Effects Institute: Cambridge, MA, 1991.

(2) Maher V.; Bhattacharya, N. P.; Chia-Miao, Mah. M.; Boldt, J.; Yang, J.−L.; McCormick, J. J. Relationship of nitropyrine-derived DNA adducts to carcinogenesis; Health Effects Institute Research Report No 55; Health Effects Institute: Cambridge, MA, 1993.

(3) Whyatt, R. M.; Santella, R. M.; Jedrichowsky, W.; Garte, S. J.; Bell, D. A.; Ottman, R.; Gladek-Yarborough, A.; Cosma, G.; Young, T.−L.; Cooper, T. B.; Randall, M. C.; Manchester, D. K.; Perera, F. P. Relationship between ambient air pollution and DNA damage in Polish mothers and new-borns. *Environ. Health Perspect.* **1998**, *106* (Suppl. 3), 821−826.

(4) Zanesi, N.; Mognato, M.; Pizzato, M.; Viezzer, C.; Ferri, G.; Celotti, L. Determination of hprt mutant frequency and molecular analysis of T-lymphocyte mutants derived from coke-oven workers. *Mutat. Res.* **1998**, *412*, 177−186.

(5) Suzuki, T.; Hayashi, M.; Wang, X.; Yamamoto, K.; Ono, T.; Myhr, B. C.; Sofuni, T. A comparison of the genotoxicity of ethylnitrosourea and ethyl methanesulfonate in lacZ transgenic mice (Muta Mouse), *Mutat. Res.* **1997**, *395*, 75−82.

(6) Kawanishi, M.; Matsuda, T.; Nakayama, A.; Takebe, H.; Matsui, S.; Yagi, T. Molecular analysis of mutations induced by acrolein in human fibroblast cells using supF shuttle vector plasmids, *Mutat. Res.* **1998**, *417*, 65−73.

(7) Trinh, S.; Reysset, G. Mutagenic action of 5-nitroimidazoles: In vivo induction of GC → CG transversion in two Bacteroides fragilis reporter genes. *Mutat. Res.* **1998**, *398*, 55−65.

(8) Pacchierotti, F.; Adler, I.−D.; Anderson, D.; Brinkworth, M.; Demopoulos, N. A.; Laehdetie, J.; Osterman-Golkar, S.; Peltonen, K.; Russo, A.; Tates, A.; Waters, R. Genetic effects of 1,3-butadiene and associated risk for heritable damage. *Mutat. Res.* **1998**, *397*, 93−115.

(9) Hall, D. B.; Holmlin, R. E.; Barton, J. K. Oxidative DNA damage through long-range electron transfer. *Nature* **1996**, *382*, 731−735.

(10) Richard, H.; Daune, M.; Schreiber, J. P. *Biopolymers* **1973**, *12*, 1.

(11) Foerster, W.; Bauer, E.; Schitz, H.; Akimenko, N. M.; Minchenkova, L. E.; Evolokimov, Y. M.; Varshakovsky, Y. M. *Biopolymers* **1979**, *18*, 625.

(12) Nandy, A.; A New Graphical Representation and Analysis of DNA Sequence Structure: I. Methodology and Application to Globin Genes. *Curr. Sci.* **1994**, *66* (4), 309−314.

(13) Ray, A.; Raychaudhury, C.; Nandy, A. Novel Techniques of Graphical Representation and Analysis of DNA Sequences−A Review. *J. Biosc.* **1998**, *23* (1), 55−71.

(14) Nandy, A.; Nandy, P. Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication. *Current Sci.* **1995**, *68* (1), 75−85.

(15) Nandy, A. Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *Comput. Appl. Biosci.* **1996**, *12* (1), 55−62.

(16) Nandy, A. Graphical Analysis of DNA Sequence Structure: III. Indications of Evolutionary Distinctions and Characteristics of Introns and Exons. *Current Sci.* **1996**, *70* (7), 661−668.

(17) Raychaudhury, C.; Nandy, A. Indexing Scheme and Similarity Measures for Macromolecular Sequences. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243−247.

(18) Nandy, A.; Nandy, P.; Basak, S. C. Quantitative Descriptor for SNP Related Gene Sequences. Manuscript in preparation.

CI990117A