Inter*net* Electronic Journal of Molecular Design

July 2002, Volume 1, Number 7, Pages 367–373

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Milan Randić on the occasion of the 70th birthday
Part 3

Guest Editor: Mircea V. Diudea

Quantitative Descriptor for SNP Related Gene Sequences

Ashesh Nandy,[1,2] Papiya Nandy,[3] and Subhash C. Basak[4]

[1] Indian Institute of Chemical Biology, 4 Raja S C Mullick Road, Calcutta 700 032, India
[2] Current address: Environmental Science Programme, Jadavpur University, Calcutta 700032, India
[3] Physics Department, Jadavpur University, Jadavpur, Calcutta 700 032, India
[4] Natural Resources Research Institute, University of Minnesota, Duluth, MN 55811, USA

Received: June 26, 2002; Revised: July 5, 2002; Accepted: July 10, 2002; Published: July 31, 2002

Citation of the article:
A. Nandy, P. Nandy, and S. C. Basak, Quantitative Descriptor for SNP Related Gene Sequences, *Internet Electron. J. Mol. Des.* **2002**, *1*, 367–373, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Quantitative Descriptor for SNP Related Gene Sequences[#]

Ashesh Nandy,[1,2,*] Papiya Nandy,[3] and Subhash C. Basak[4]

[1] Indian Institute of Chemical Biology, 4 Raja S C Mullick Road, Calcutta 700 032, India
[2] Current address: Environmental Science Programme, Jadavpur University, Calcutta 700032, India
[3] Physics Department, Jadavpur University, Jadavpur, Calcutta 700 032, India
[4] Natural Resources Research Institute, University of Minnesota, Duluth, MN 55811, USA

Received: June 26, 2002; Revised: July 5, 2002; Accepted: July 10, 2002; Published: July 31, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (7), 367–373

Abstract

Motivation. The rapid growth of DNA sequence information relating to various gene sequences makes it imperative to have quantitative measures to compare and contrast the different sequences. This is of special relevance to libraries of single nucleotide polymorphic (SNP) genes.

Method. We had presented a scheme earlier for characterizing a given DNA sequence by a numerical descriptor. We had also shown that the sensitivity of the descriptor to small changes in a given sequence renders it suitable for using it as a parameter to index toxicity levels of various chemicals that induce changes in DNAs.

Results. Here we propose that such descriptors can be used to index SNP related genes and show by way of application with simulated changes in human beta–globin gene that very small changes in base composition can lead to identifiable changes in a DNA descriptor.

Conclusions. This scheme identifies the relationship of one member of the SNP gene family with another in terms of exact location and nature of changes in the sequence. Such a scheme will be useful in compilation, identification and distribution of genetic variation data on SNP related gene sequences.

Keywords. SNP genes index; DNA descriptors; graphical representation of DNA; DNA sequence index; molecular design; beta globin gene.

1 INTRODUCTION

The creation of a library of single nucleotide polymorphic genes for gene specific drug targeting holds great promise for new generation of drugs and therapies [1]. Further, the proposal for a compilation of SNP related gene maps for reference and distribution purposes in the public domain will be of additional benefit.

In this context, it would be useful to have a method to quantitatively assess the differences in SNP genes and tag the sequences in a prescribed order in the anticipated large library of SNP related genes. Such a procedure would require some way of characterizing DNA sequences numerically on one or more attributes. Several authors have formulated different schemes to

[#] Dedicated to Professor Milan Randić on the occasion of the 70th birthday.
* Correspondence author; E–mail: anandy43@yahoo.com.

characterize DNA sequences so that members of different gene families and sequences can be identified by unique numbers. DNA characterization can be done through graphical representational methods [2–4] and mathematical invariants [5–11] that have the potential to codify the sequence structure of a DNA into a set of numbers for similarity/dissimilarity comparisons. Raychaudhury and Nandy [5] considered properties of 2D–graphical representation of DNA sequences, for example, parameters arising from mean moments about the graph origin as DNA descriptors, and arrived at one way of characterizing DNA primary sequences. Tarafdar *et al*. [6] calculated a set of fractal dimensions and showed that there were distinct differences between intron and exon sequence parameters, and efforts have also been initiated [7,8] into refining the 2D representation to reduce or remove the degeneracies inherent in such a scheme. Randić, Nandy and Basak [9] worked on a matrix representation based on 2D graphical representation of Nandy [2] and showed that matrix invariants for a family of genes were similar; extension to 3D representation [10] to obtain matrix invariants and leading eigenvalues of (truncated) DNA sequences have been shown to lead to ambitious programs for DNA sequence comparisons. Randić [11] has also formulated a scheme for constructing invariants based on a condensed matrix form of nucleotide doublets of the sequences, and Randić, Guo and Basak [12] considered similar matrices generated from triplets of nucleic acid bases to arrive at invariants for quantitative comparisons of DNAs from various sources, as also Randić and Vračko [13] in their work on quantification of similarities in DNA sequences. These matrix invariants techniques, however, remain very difficult to solve numerically for large numbers of bases and the attempts done to date relate mostly to truncated sequences for such practical reasons [10].

Our initial attempts to formulate a DNA numerical descriptor based on the 2D graphical representation was found to be very sensitive to changes in sequence, providing quantitative estimates of base alterations, deletions and additions [5]. We have recently proposed to use the original DNA descriptors as an index for quantifying the effects of toxic chemicals on DNA sequences [14]. This method provides a good numerical characterization of a DNA sequence, and is capable of handling large sequences with reasonable degree of accuracy. We believe such a descriptor would provide a good basis for quantifying the changes in SNP related genes and thus provide a basis for indexing such genes for ready comparison and tabulation.

2 METHOD

The method is based conceptually on a DNA representative walk in two dimensions [2–4] where, for example, a step is taken in the positive *x*–axis for a guanine in the sequence, a step along the positive *y*–axis for a cytosine, a step along the negative *x*–direction for an adenosine and one along the negative *y*–axis for a thymine [2]. This creates a map of points that depend on the base distribution pattern in the gene and varies with mutational and other changes. We refer to this

particular representation as the ACGT–axes system, reading the base representations clockwise from the negative *x*–axis. Two other orthogonal systems can be drawn: AGCT– and ACTG–axes systems, but in this paper we restrict our attention to the ACGT–axes system to illustrate our hypothesis. Applications of graphical representations methods have clearly demonstrated characteristics of local and global base distributions in a sequence, identified repeat structures and helped in rapid location of coding regions through intron–exon discrimination (see [15] for a review). Graphical methods have been shown also to provide a fast and easy method for study of aspects of comparative genomics, and our studies of gene sequences of the globin, myosin heavy chains, histones, and other gene families have shown through the dispersion in their graphical representations that evolutionary changes produce shifts in base distributions that seem to reflect evolutionary distances [2,16].

In any of these axes systems, defining a weighted mean *x* coordinate value as $\mu_x = \Sigma\, x_i / N$ and mean *y* coordinate value as $\mu_y = \Sigma\, y_i / N$, we can define a graph radius [5]

$$g_R = (\mu_x^2 + \mu_y^2)^{1/2} \tag{1}$$

Because these coordinate values arise essentially from differences, in our preferred coordinate assignments, between the instantaneous totals of A and G and between C and T residues as we move along the sequence, the g_R is very sensitive to small changes in the sequence. The measured difference in g_R, Δ_R, between two graphs differing by a single point mutation can therefore be used as a numerical descriptor for tabulating the SNP changes. In practical applications these numbers can be arrived at from the sequence data alone without having to take recourse to the graphical representations.

In using the DNA difference descriptor described above, one has to keep in mind that the basic descriptors are not necessarily unique. This arises from the inherent degeneracy in describing a higher dimensional object on a two dimensional plane. By associating the four axes with the four nucleotides we have mapped a DNA sequence through a series of points on to the 2D graph, but sequence segments like AG, AGAG, AGAGAG and so on, overlap in an ACGT–axes system and will be represented by a single point on the final plot. This carries over to the computations of the μ_x and μ_y in case of some specific patterns in the distributions of bases; *e.g.*, in the case of two sequences of equal lengths, the contribution of sequence segments GAAG and AGGA to the evaluation of μ_x will be the same ($\Sigma\, x = 0$). Note, however, that (*a*) in case of unequal lengths of the sequence under consideration, the normalization by *N* may be sufficient to produce different results, and (*b*) in the case of single nucleotide alteration, pathological instances like the ones just described will be highly unlikely. Thus we believe that the μ values provide sufficient discriminatory power to make the Δ_R useful as quantitative descriptors for SNP genes.

These Δ_R descriptors are expected to provide a guide to the alterations in a gene sequence as compared to a standard gene sample. In particular, in the case of SNP genes, the point mutations

can occur in any of the four bases, and in any location on the gene. It is possible that the value of the Δ_R for a mutation of a guanine to adenosine may be duplicated in the value of a Δ_R in some other transformation, say cytosine to thymine. It is important therefore to differentiate between the different ranges of Δ_R and we propose to do this by explicitly indicating in superscript the mutation group to which the Δ_R values belong; *e.g.*, for guanine to adenosine mutation table the descriptor would be labeled as Δ_R^{GA}. There can thus be at most 16 such tables for any given gene sequence.

3 RESULTS AND DISCUSSION

We use the beta globin gene (EMBL database Accession Number U01317, ID HSHBB), inclusive of exons and introns, as a sample for testing purposes. The 2D representation of the gene (Figure 1) clearly shows the base distribution characteristics of the sequence that consists of a total of 360 A, 227 C, 296 G and 491 T bases.

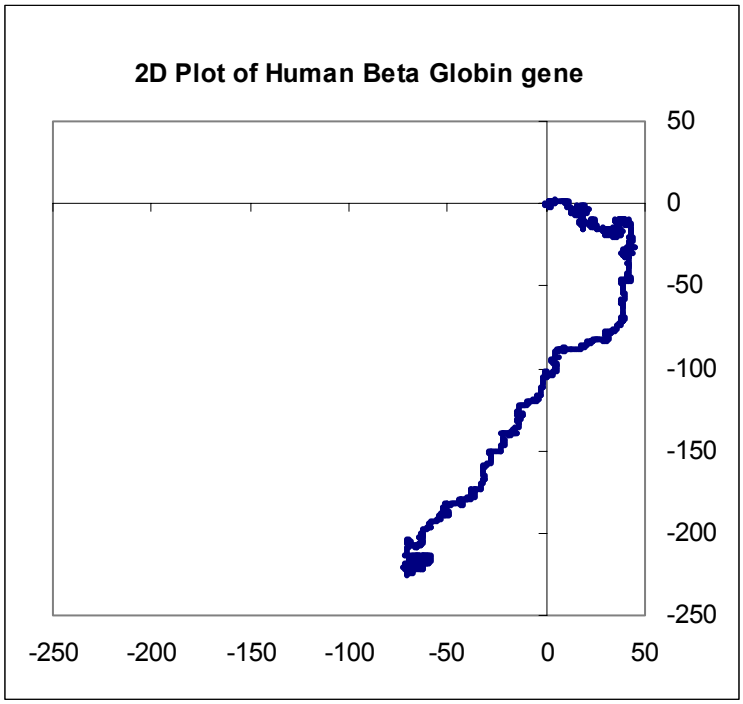

Figure 1. The 2D graphical representation of the beta–globin gene. Axes ACGT reading clockwise from the negative *x*–axis. Total 1424 bases: 360 A, 227 C, 296 G, and 491 T.

We have calculated Δ_R for mutations of a guanine to an adenosine at each position of the guanine in the sequence. The results displayed in Figure 2 for Δ_R^{GA} against guanine number show that the Δ_R^{GA} is position sensitive, starting from its maximum value, for a mutation in the first base position and decreasing to almost zero for the last guanine base position, which is understandable from the impact the change has on succeeding values of the coordinates. Also, large intervals between consecutive guanines produce large differences in Δ_R^{GA}. It is interesting to note the large variation for a guanine to adenosine mutation. On the DNA walk representation mentioned above where *a* and *g* are steps in the negative and positive *x*–directions, respectively, a change from a guanine to

an adenosine implies a retraction in *x*–value by one, but the graph radius can be seen to change appreciably depending on which base position is affected.

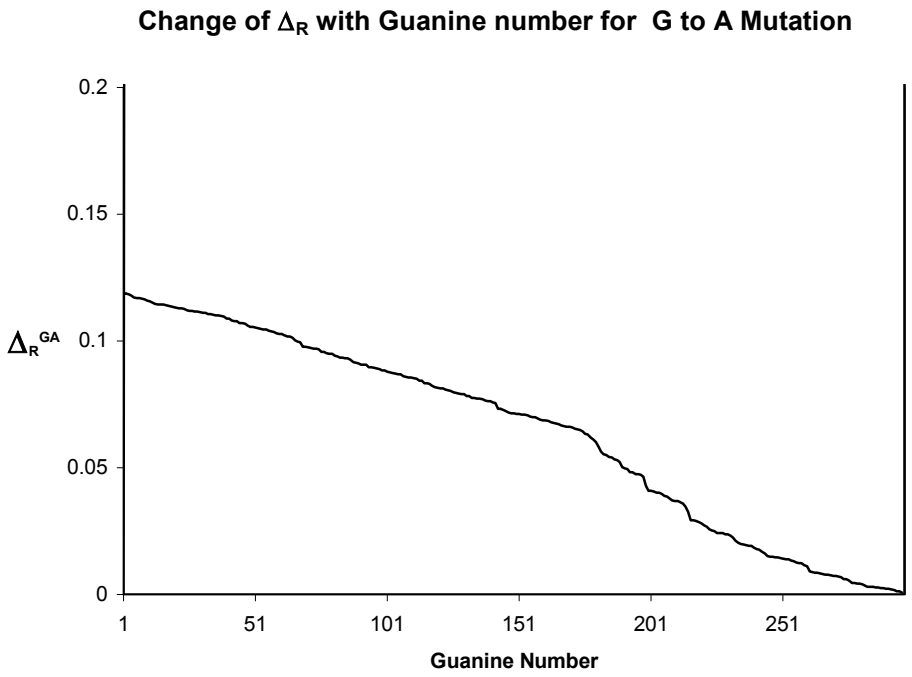

Figure 2. Change of Δ_R with Guanine number for G to A mutations.

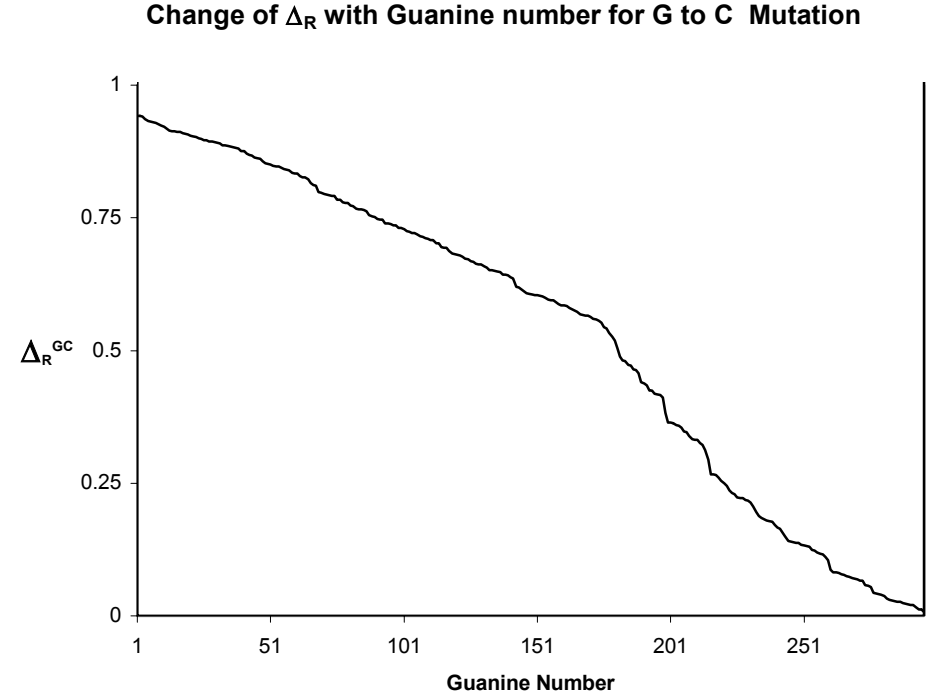

Figure 3. Change of Δ_R with Guanine number for G to C mutations.

Mutation of a guanine to a cytosine, which implies a change in the direction of the step taken from the *x*–direction to the *y*–direction, involves a bigger change than in the previous case. Here also, as seen in Figure 3, there is a different number for Δ_R^{GC} for each position of the guanine.

Thus if two beta globin gene samples differ by a mutation of a guanine to adenosine in the position of the 14th guanine base in the sequence, $\Delta_R^{GA} = 0.11446$ whereas if a third sample sequence has a guanine in the 123rd position mutated to an adenosine the $\Delta_R^{GA} = 0.08091$, allowing a natural ordering of the three sequences. At a sufficient level of accuracy, because the Δ_R are position sensitive, every mutation can be related and identified. Such a method will be helpful in identification and characterization of genetic variation and consequently in tabulating, ordering and recall of SNP related gene sequences.

4 CONCLUSIONS

A numerical scheme based on the DNA descriptor method defined from a 2D graphical representation framework is of sufficient accuracy to distinguish between changes in single nucleotides in a gene. Given a gene for comparison, we could create tables and graphs and index each gene compared to a standard through variations in the appropriate tables of Δ_R. This will be of immense benefit in tabulating and classifying SNP gene libraries.

We would like to mention here that this indexing scheme is one of a continuing series of efforts to provide quantitative descriptors for DNA sequences. Since the original paper by our group on indexing DNA sequences [5], there have been numerous attempts at determining unique descriptors to provide a numerical standard for characterizing DNA sequences [6,9–13], but computable methods have proved to be elusive, whereas the necessity of devising such schemes continue to grow if only to enable efficient search and retrieval procedures for the exploding number of submissions to DNA databases [17].

In this context the method proposed in this paper can be considered as a working hypothesis for SNP genes library, but where, although descriptors based on 2–D representations to index DNA sequences [5] have been proposed earlier also (as a marker for the toxicity of DNA–damaging chemicals [14]), the issue of degeneracy of the graphs and associated indices remains to be resolved even though it may not affect the SNP genes indices. It is to be hoped that the recent proposals of Guo and co–workers [7,8] to remove the degeneracies from the 2D representations may help to overcome this difficulty while still permitting an easily quantifiable and unique set of indices of DNA sequences.

Acknowledgment

We would like to thank the reviewers of this paper for several helpful suggestions that have provided a better perspective on the work undertaken.

One of us (AN) would like to join the editors of this special issue of the Internet Electronic Journal of Molecular Design in honor of Milan Randić to pay his respects to Milan and recall fondly the long evenings of wine and anecdotes together with Subhash Basak, interspersing a very active schedule of research on DNA characterization. That period of work helped to provide new insights into the intriguing and important area of DNA sequences.

5 REFERENCES

[1] *Nature*, **1999**, *398*, 545–546.
[2] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, *Curr. Sci.* **1994**, *66*, 309–314.
[3] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol.* **1986**, *119*, 319–328.
[4] P. M. Leong and S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Applic. Biosc.* **1995**, *11*, 503–507.
[5] C. Raychaudhury and A. Nandy, Indexing scheme and similarity measures for macromolecular sequences, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 243–247.
[6] S. Tarafdar, P. Nandy, S. Sahoo, A. Som, J. Chakrabarti, and A. Nandy, Self–similarity and scaling exponent for DNA walk model in two and four dimensions, *Indian J. Phys.* **1999**, *73B*, 337–343.
[7] X. Guo, M. Randić, and S. C. Basak, A novel 2D graphical representation of DNA sequences of low degeneracy, *Chem. Phys. Lett.* **2002**, *350*, 106–112.
[8] Y. Liu, X. Guo, J. Xu, L. Pan, and S. Wang, Some notes on 2–D graphical representation of DNA sequences, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 529–533.
[9] M. Randić, A. Nandy, and S. C. Basak, On the numerical characterisation of DNA primary sequences, *J. Math. Chem.*, submitted.
[10] M. Randić, M. Vračko, A. Nandy, and S. C. Basak, On 3–D representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235–1244.
[11] M. Randić, On characterization of DNA primary sequences by a condensed matrix, *Chem. Phys. Lett.* **2000**, *317*, 29–34.
[12] M. Randić, X. Guo, and S. C. Basak, On the characterization of DNA primary sequences by triplet of nucleic acid bases, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 619–626.
[13] M. Randić and M. Vračko, On similarity of DNA primary sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599–606.
[14] A. Nandy and S. C. Basak, A simple numerical descriptor for quantifying effect of toxic substances on DNA sequences, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 915–919.
[15] A. Ray, C. Raychaudhury, and A. Nandy, Novel techniques of graphical representation and analysis of DNA sequences – A review, *J. Biosci.* **1998**, *23*, 55–71.
[16] A. Nandy, Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons, *Curr. Sci.* **1996**, *70*, 661–668.
[17] A. Nandy, Recent investigations into global characteristics of long DNA sequences, *Ind. J. Biochem. Biophys.* **1994**, *31*, 149–155.