Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
CHAPTER 5

The GRANCH Techniques for Analysis of DNA, RNA and Protein Sequences

Ashesh Nandy*

Centre for Interdisciplinary Research and Education, 404B Jodhpur Park, Kolkata 700068, India

Abstract: The very rapid growth in molecular sequence data from the daily accretion of large gene and protein sequencing projects have led to issues regarding viewing and analyzing the massive amounts of data. Graphical representation and numerical characterization of DNA, RNA and protein sequences have exhibited great potential to address these concerns. We review here in brief several different formulations of these representations and examples of applications to diverse problems based on what this author had presented at the Second Mathematical Chemistry Workshop of the Americas in Bogota, Colombia in 2010. In particular, we note several insights that were gained from such representations, and the applications to the bio-medicinal field.

Keywords: Alignment free sequence comparison, bio-medical applications, chaos generator representation, DNA graphical representation, DNA numerical characterization, DNA sequences global characteristics, DNA visualization, GRANCH applications, GRANCH techniques, identifying potential drug target regions in viruses, lattice graphs in medical applications, long range correlations, long range palindromic sequences, phylogenetic trees, protein graphical representation, protein numerical characterization, purine pyrimidine relationship, QSAR.

INTRODUCTION

My first brush with a DNA sequence, in around 1990, left me totally puzzled: I could not “see” nor get a “feel” of anything noteworthy in the apparent jumble of characters that symbolized a DNA, not the least because I had never studied biology myself. My background was physics, and I began a search for, to me, a more meaningful exposition of the sequence of characters that represented the

*Corresponding author Ashesh Nandy: Centre for Interdisciplinary Research and Education, 404B Jodhpur Park, Kolkata 700068, India; Tel: +91 33 2473 0577; E-mail: anandy43@yahoo.com

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)
Copyright © 2015 Bentham Science Publishers Ltd. Published by Elsevier Inc. All rights reserved.
10.1016/B978-1-68108-053-6.50005-3
DNA sequence. My studies led me to appreciate and anticipate the immense potential opening up with the sequencing of genome length sequences and the concomitant need for rapidly scanning and analyzing DNA sequences for matters of interest [1], and to get excited at the new insights being gained from a global perspective of the DNA sequences: Jeffreys [2] had shown through his Chaos Generator Representation that such sequences had an inherent fractal nature; Peng et al. [3] speculated that DNA sequences had long-range correlations, an observation that raised a storm of papers in very short order; and Voss [4] showed that long range fractal correlations existed in DNA sequences with the degree of correlation varying with evolutionary divergence.

But a close up look at a DNA sequence and how the bases were distributed along it still lacked an appealing representation. Experimenting with various formats I determined that a 2D graphical representation, as explained later in this chapter, was what I could relate to on a purely personal basis. After many graphs of various sequences and discussions with some eminent persons to ensure that such a simple stratagem was not already familiar to cutting edge biologists, I published a paper on it in Current Science (Bangalore) in 1994 [5]. Imagine my consternation when I was informed soon after that Gates had already anticipated such a device, albeit with different axes assignments, way back in 1986 [6], but which seemed to have been in limbo since! A short note had to be published soon after informing of this oversight and explaining the differences although both used Cartesian co-ordinate system to plot the graphs [7].

However, a physics background demands some quantitative appraisal of whatever nature has to offer. I had observed certain similarities and changes in plots of conserved gene sequences of various species, but coming up with some way to measure the changes posed difficulties with these plots of discrete numbers. I had done some number crunching with individual gene segments like introns and exons [8, 9], but now the need was for whole sequences for which we came up with a geometrical interpretation to describe in general a macro-molecular sequence and measure sequence differences. We presented our scheme at the First Indo-US Workshop on Mathematical Chemistry in Shantiniketan, West Bengal, India in 1998 [10] where we reported, as stated in the abstract, that “Geometrisation of macromolecular sequences in the form of a graphical
representation provides one ... technique where the nucleotides in a gene sequence can be viewed as objects in a 4-dimensional space; the method can be extended, in principle, to include, say proteins, in a 20-dimensional space. We have found a reduced 2-dimensional representation of DNA sequences very useful in studies of nucleotide distribution and composition. .... We here propose a new measure of the dispersion of DNA graphs that can be used to quantify the differences between two or more graphs of genes of various organisms .... It also appears that once standardized the proposed scheme may help study molecular phylogeny in evolutionary time scale.”

Although the participants in the Shantiniketan Workshop included stalwarts in the field like Prof. Milan Randić, Prof. Haruo Hosoya, Prof. Paul Mezey and others, our scheme did not seem to evoke any response, not surprising since they apparently did not know about DNA issues. But Prof. Subhash Basak of the University of Minnesota, USA and co-organiser of the Workshop was intrigued enough by our work and its potential to describe DNA sequences through graph invariants to meet me in Kolkata after the workshop to discuss the possibility of using invariants for DNA sequences as descriptors. Subsequently Prof. Basak invited me the following summer to Duluth to carry out further research on DNA mathematical descriptors in his group funded by the Natural Resources Research Institute (NRRI). Prof. Milan Randić and some other distinguished scientists were also invited there to begin to work on DNA descriptors in a project funded by NRRI.

It began with a talk I gave at the University of Minnesota, Duluth about my work on mathematical descriptors of DNAs arising from my graphical representation method. Among the attendees was Prof. Milan Randić who, with Prof. Basak, immediately saw the potential for converting a DNA sequence graph to a matrix and thereby extract numerical invariants which could be a more meaningful way to characterize DNA sequences. We collaborated then on a proposal for a 3D graphical representation and a matrix method for extracting graph invariants for the first exons of beta globin sequences of several species. This was published in 2000 in the Journal of Chemical Information and Computer Science [12] and led very soon to a whole host of papers on the DNA graphical representations and
numerical characterisations and applications of them that continues still as more and more areas keep opening up and a new field of research seems to have begun.

This review is a brief introduction to the readers of this new and exciting field of research on graphical representation and numerical characterization (GRANCH) of bio-molecular sequences, based on the talk I presented at the Second Mathematical Chemistry Workshop of the Americas in Bogota, Colombia, in July 2010 [13]. Some of the various applications made to date using these techniques are also covered briefly, with special emphasis on our recent work that provides a possible approach to anti-viral vaccine design that could be expected to be less susceptible to invalidation through mutational changes in the viral proteins. More details can be found in the several reviews [14-18] and book chapters [19-21] that have appeared on the subject, and of course there are always the original papers. (Note added in proof: See also bibliography in Ref [82].)

GLOBAL CHARACTERISTICS OF DNA SEQUENCES

As sequence data on long stretches of DNA began to become available in the late 1980’s, there arose a problem on how to view them and a curiosity to know whether any systematics lay hidden in the apparently random arrangement of characters representing the bases in the sequence. H J Jeffrey [2] came up with the idea of plotting them in a square grid where the four corners were identified with the four bases A, C, G, T. The algorithm was to start from the center of the square and for the first base plot a point midway between the origin and the home corner of the base. For the second base he started from the point representing the first base and plotted a point midway between it and the home corner of the second base. Continuing in this way filled up the square with a series of points until the entire sequence was plotted. This diagram he called the Chaos Generator Representation (CGR) of the DNA sequence. He noticed that different animal kingdoms showed different patterns – double scoop depletion regions for vertebrates, striped patterns for plant sequences, an apparently random distribution for bacterial genomes. Overall, each sub-square section of the CGR pattern seemed a replica of the whole, i.e. DNA sequences had properties of self-similarity or fractal nature.
The CGR diagrams of various sequences were investigated by several researchers to find different properties of biological interest. Burma et al. [22] showed the structures observed in CGR diagrams arises from skews in base composition and presence of repetitive sequences or specific motifs. Dutta and Das [23] reported that a CGR plot can be reproduced by suitable algorithms by manipulating different combinations of strings of bases with appropriate frequencies. Thus, the double scoop depletion patterns seen in vertebrate CGRs arises from scarcity of CG dinucleotides, and so on. Baranidharan et al. [24] developed quantitative methods to generate similarity/dissimilarity maps of genomic sequences and showed that for certain mitochondrial genomes species wise characteristic features could be seen when nucleotide stretches of 7 or more bases at a time were analysed.

In a slightly different vein, Peng et al. [3] considered the structure of the bases in a DNA sequence and analysed them on the basis of their being pyrimidines (C,T) or purines (A,G) only. On an X-Y graph where the x-axis counted the nucleotide number, they plotted the appearance of the bases in a sequence by taking a step diagonally upwards if it was a purine or downwards if it was a pyrimidine to the next nucleotide number. Plotting the whole sequence step by step in this manner they generated a graph with an irregular up-down structure which they called a “DNA landscape”. By taking subsections of the graph they found that the subsections also looked similar to the up-down structure of the whole, and the same was true of sub-sub-sections and so on showing that the purine-pyrimidine structure of a DNA sequence had self-similarity, which was what Jeffrey had remarked two years ago. Peng et al. then searched for possible correlations by estimating frequencies of different lengths of nucleotide stretches and found that all gene sequences with the mosaic structure of introns and exons had long-range correlations whereas intronless genes did not show this feature. The implications of such an observation, on the face of it, are huge: the beginning of the DNA sequence should, theoretically, be knowing what the end would be like! Such an observation quite naturally led to a storm of papers on the subject until it quittened down after the observation that since DNA sequences are known to elongate by duplicating long stretches of subsequences, it was possible that such sequences showed apparent long range correlations.
Around the same time Voss [4] conducted a rigorous analysis of 25000 DNA sequences with over 50 million bases covering organisms of all classes to search for long range correlations. Using a spectral density function analysis he concluded that “(a) long range fractal correlations exist in DNA sequences, (b) the degree of correlation as measured by a spectral exponent varies systematically with evolutionary category and (c) short range periodicities of period 3 are prominent while other periods, e.g. 9, are also present. The fractal correlations have been seen to extend over long ranges of nucleotide positions, with the smallest for phage and bacteria and extending to over 100,000 bases for the higher classes” [14].

**VISUALIZING THE DNA**

To get a feel for the actual distribution of bases along a DNA sequences you need a more direct graphical depiction than what the abstract representations of Jeffrey [2] or Peng *et al.*, [3] can offer. This problem was addressed many years ago by Hamori and Ruskin [25] with their proposal for a 3-dimensional graphical representation of a DNA sequence. They proposed a hypothetical square on the x-y plane with four corners (NW, NE, SE, SW) identified with the four bases A, C, G, T and the nucleotide number to be counted along the z-axis. Thus for a DNA sequence like ACGGT, one would plot a point on the A-corner at z-coordinate 1, then draw a line to the next base, C in this case, in its corner with z now equal to 2, and so on. For a sequence like ACGTACGTACGT this would generate a spiral around the z-axis; in case there was a preponderance of one base or the other the curve would flow along those corners. These curves the authors called H-curves. Visualizing such a 3D image on the 2D plane of the paper is admittedly difficult. However, the authors suggested that drawing two such curves at slightly different angles would allow stereoscopic vision so that the DNA could be seen in 3D. Taking the bacteriophage M13 as an example they showed that in their representation they could easily identify regions of sharp changes in base composition through visualization that would be difficult to determine from the normal character representation.

This author’s search for a meaningful display of DNA sequence information led him to propose a 2-dimensional graphical representation where the four cardinal
directions are associated with the four bases [5]. The method is to take a walk in the negative x-direction if there is an adenosine in the sequence, in the positive y-direction for a cytosine, positive x-direction for a guanine and the negative y-direction for thymine. Proceeding to walk in succession in the appropriate direction in the order of the bases making up the particular DNA sequence generates a path that visually depicts the arrangement of bases in the sequence. These DNA plots were found to be characteristic of the types of gene sequences and that the same genes from different species showed almost the same pattern. Since we know that specific genes from different species have significant homology, and in fact that is how often new genes are recognized, it is not surprising that their graphical plots will show basically the same shape.

It was found later that Gates [6] had already proposed a similar scheme to depicting gene sequences, although his assignment of bases were different from Nandy’s scheme. A year later Leong and Morgenthaler [26] independently proposed another 2D scheme, where the base assignments were again different from the two just mentioned. On a 2D Cartesian co-ordinate system the assignments of the bases with the cardinal directions in the three schemes are, starting from the negative x-direction and going clockwise, GTCA (Gates), ACGT (Nandy) and CTAG (Leong and Morgenthaler). It is interesting to note that these three axes representations exhaust all possible 2D schemes of this type, and these can be seen to be like 2D projections of the Hamori-Ruskin H-curves.

The 2D plots can be scaled to accommodate from the largest to the smallest DNA sequences depending on the level of detail one wishes to observe. In reference 27 (the illustration (3)) depicts the 73326 thousand base long human beta globin sequence that contains the beta, eta, delta and the gamma globins of less than 2000 bases each, which can individually be plotted on a smaller scale. Plots such as these provide a quick estimation of base composition and distribution along a DNA sequence. An inspection of the human beta globin sequence graph shows that it has two sections that are mainly A-T rich with one part in between that is T-dominated. A plot of a sequence like the chicken myosin heavy chain gene is represented in illustration 2 (loc. cit.) shows that it also is AT-rich; from the angle it makes with the axes, it is evident that the sequence is dominated by larger percentage of T’s than A’s, and likewise one can determine preponderance of
structures like $A_mT_n$ from inspections of such plots. Further applications of these graphs are taken up later in this chapter.

The 2D representations, however, suffer from degeneracy in that nucleotide pairs, like AG or CT in the Nandy scheme, will result in only one step instead of two. Bielińska-Waż et al. [28] have shown that this can be accounted for by a mathematical method of using a weight parameter for each visit to the same location, but a number of researchers has been to propose different ways to represent DNA sequences graphically that reduces or removes this degeneracy. An extensive coverage of these methods can be found in Nandy, Harle and Basak [16], but we may mention here that one of the first proposals to reduce the degeneracy was the scheme of Guo, Randić and Basak [29] where the unit vectors for the four bases were aligned at a small angle to the cardinal directions. Yau et al. [30] used a two-quadrant representation in 2D space where A,G were inclined to the x-axis in the 4th quadrant and T,C were inclined to the x-axis but in the first quadrant, and the nucleotide count was recorded along the x-axis; this generated a DNA graph extending in the positive x-direction and had no degeneracy. He et al., [31] proposed to characterize a DNA sequence by their chemical (amino, keto), structural (purine, pyrimidine) and bond strengths (weak, strong) and plotted this set of three reduced sequences as characteristic curves that extended along the x-axis with nucleotide number thus avoiding degeneracy altogether. Randić proposed several constructs, among them a 4 horizontal line scheme [32] where the four bases were plotted in order of the sequence along four lines parallel to the x-axis and placed unit distance apart where the nucleotide number was again counted along the x-axis, a compact “worm curve” representation [33], four-color maps [34], “spectrum-like” curves [35] among others which reduced or eliminated the degeneracy inherent in the classical 2D approach.

3D and higher dimensional representations have been proposed to more faithfully reproduce the features of a DNA sequence or enable more accurate calculations. Hamori and Ruskin [25] had originally proposed a 5D model where the four bases were plotted in four dimensions and the fifth was for nucleotide count, but since this was difficult to visualize he had moved to the 3D H-curve representation. 3D representations and variations were also proposed by Randić, Vracko, Nandy and Basak [36], and Li and Wang [37] to name a few. A 4D method was proposed by
Chi and Ding [38], a 6D method by Liao and Wang [39] and an 8D method by Liu and Wang [40]. The interested reader can refer to the reviews [e.g. Ref 16] and the literature for details of these interesting developments.

**NUMERICAL CHARACTERIZATION OF DNA SEQUENCES**

Thus, the study of DNA sequences is facilitated in many ways by graphical representation, but making intra- and inter-sequence comparison becomes meaningful when the similarities and differences can be quantified in some manner. The difficulty is that since the graphical plots are composed of a set of discrete points, one has to apply either novel geometrical methods or use graph theory where the points are considered as nodes and the connections between the nodes as edges. We describe below first the geometrical methods and then the graph theoretic methods.

**Geometrical Methods for Numerical Estimations**

Two techniques were devised, one for intra-sequence comparison and another for inter-sequence comparison. For the variations within a sequence arising from the base distribution, we had observed [27] that coding regions of mammalian gene sequences appeared as a dense cluster of points in the 2D graphical representations implying high degree of mixing of the four bases in almost equal proportions, whereas the non-coding regions that were A-T or G-C rich usually appeared as long filaments. We therefore devised a cluster density measurement by enclosing such regions in a square grid and dividing the number of points in the grid by the area of the square. This was complemented by an inverse displacement method and a fractal coefficient method to numerically assess the differences between these two types of regions. Analysis of 386 introns (non-coding regions of a gene) and exons (coding regions) of 35 genes from various species by these measures showed [27] that (a) cluster density of non-coding regions are very small and fall off exponentially rapidly, (b) the cluster density of coding regions grows to about 0.8 per unit area and falls off gradually, (c) exons of evolutionarily later genes have higher cluster densities, (d) cluster densities of intronless genes like the phage M13 genome or the bacteriophage lambda are very low, closely paralleling intron densities and (e) more recent genes show greater
fragmentation and smaller lengths of the exons. The cluster density measure also enabled us to propose a way of predicting protein coding regions in new DNA sequences [41] and was used to analyse the human chromosome 3 contig 7 and predict existence of several genes [42].

Gates [6] had proposed a Manhattan distance computation to compare two or more sequences, but this method is suitable for equal length sequences, whereas gene sequences are not generally of equal lengths. To study similarities and dissimilarities of genes from various species we devised a new and different methodology, which was reported for the first time in the First Indo-US Workshop 1998 [10] and published the following year [11], as mentioned earlier. Since we have in the 2D graphical representation a set of discrete points comprising each gene sequence, we defined a function to describe the sequence as

\[ C_{\text{map}} \sim C(S_0, S_1, S_2, \ldots) \]

where \( S_0 \) is the zeroth-order term representing the coordinates \( x_f, y_f \) of the end points, \( S_1 \) is the linear term representing the first-order moments about the two axes, \( S_2 \) the second-order term representing the variance about the mean, \( S_3 \) the third order term representing the skewness, etc., all of which taken together became a descriptor for the sequence. For the initial presentation we computed the first order moments as weighted center of mass only and defined a graph radius, \( g_R \), the distance of the weighted center of mass from the origin, for each sequence and a \( \Delta g_R \) to estimate the difference between two sequences plotted on the same scale; this scheme gave a reasonable fit to the dispersion of the beta globin genes from various species [11]. Because of the cumulative nature of the sequence plots, differences in base distributions will lead to progressively increasing differences in the plots. Closely related sequences with less mutational changes between them will have smaller \( \Delta g_R \) while unrelated sequences can be expected to lead to larger values of the \( \Delta g_R \). As remarked by the authors, this method could clearly be generalized to apply to the case of protein and other sequences where one may represent the sequences in a multidimensional hyperspace with a view to eventually develop phylogenetic trees. These techniques have been used by several authors (e.g., [43, 48]).
Bielińska-Wąż et al. [28] have computed the moments to various higher orders in a 2D dynamic graph with statistical moments of mass–density distributions as new descriptors. Computing the moments for a set of histone genes, they showed that the larger number of descriptors improved the characterization of the object and different aspects of the DNA could be compared separately while retaining the simplicity of the 2D graphs. Nandy and Nandy [44] showed that the $g_{RS}$ were quite sensitive measures where base composition or base arrangement differences caused the $g_R$ to change and that two or more sequences will not have the same $g_R$ value except in some pathological cases.

**Graph Theoretic Method for Numerical Characterization of DNA Sequences**

The graph theoretic method arose out of deliberations after the first presentation of the 2D graphical representations in Duluth in 1999. The method described in the paper by Randić, Vracko Nandy and Basak [36] was to first represent a DNA sequence graphically in a 3D Cartesian grid and then convert the points to elements of a matrix by computing the ratio of Euclidean distance to graph theoretic distance between all possible pairs of points taken systematically. Matrix methods are well studied and have well recognized properties. The D/D matrix generated by the distance measures was analysed to yield a set of eigenvalues with the leading eigenvalue being taken as invariant of the matrix and therefore of the sequence. Differences between the leading eigenvalues of various gene sequences could then be taken as indicative of their evolutionary distances, although this seminal paper limited itself to computation on the basis of the first exons of 11 beta globin genes only. The interesting point to note is that this paper led to generation of intense interest among researchers and many different ways of representing DNA sequences and computation of evolutionary distances subsequently ensued (see review Nandy et al. [16]). Authors such as Randić et al. [33], Randić [35], He and Wang [45], Song and Tan [46] and many others proposed different ways to graphically represent DNA sequences and convert the plots to mathematical objects, and derive leading eigenvalues as invariants of the sequences. For example, He and Wang [45] reduced the DNA sequences to a set of three sequences based on their structural, chemical and bonding nature and devised a vector of the three leading eigenvalues of the matrices associated with each of the reduced sequences which they proposed as being characteristic
sequences of the original DNA sequence. Distances between two sequences then were computed by determining the distances between the end points of the two vectors. Song and Tan [46] similarly devised a 24-component vector characterizing a sequence, others came up with other ways of computing the inter-sequence distances based on vectors devised out of the matrix eigenvalues. Such matrix invariants from their own representations were used by Liao et al. [47], Wang et al. [43], Liao et al. [48] and others to draw phylogenetic trees for mitochondrial genes, SARS coronavirus genomes, etc. The graph theoretic method, however, does not seem to have been applied so far to determine specific features within a sequence.

**GRAPHICAL REPRESENTATION AND NUMERICAL CHARACTERIZATION OF PROTEIN SEQUENCES**

**Graphical Representation of Protein Sequences**

Developments in the graphical representation and numerical characterization of DNA sequences raised the possibilities of using similar analysis of protein sequences, albeit with difficulty arising from the fact that now we have to contend with 20 amino acids making up a protein chain whereas DNA sequences were made up of only four nucleotides. Although Meeta Rani [49] had shown as early as 1998 the presence of statistical self-affinity, a kind of self-similarity, in protein sequences that implies a fractal nature, graphical representation methods for proteins drew attention with the paper of Randić [50]. The basic idea here was to start with the CGR method of Jeffrey to plot a RNA sequence drawing triangles for every triplets of bases, *i.e.*, the genetic codes, and taking the centers of each such triangle as corresponding to the residue the triplet would code for. Thus starting with the mRNA, this method generates a CGR-equivalent 2D graphical representation for the protein sequence. Randić et al. [51] carried the method further to construct a zigzag curve for the A-chain of human insulin which allows a direct conversion of a protein sequence into a numerical sequence of (x,y) coordinates that can be used subsequently for construction of the graph-theoretic matrices and sequence invariants. The technique was refined to remove some arbitrariness that were inherent in the 2D scheme by converting the 2D graph to a 3D graphical representation where the triplets were assigned to the corners of a tetrahedron structure; although visual inspection of the graphical patterns had to
be discarded in this scheme, the authors claimed that construction of graph invariants in this manner was more accurate and unique.

Randić et al. [52] proposed a Magic Circle representation where the protein sequence graph starts from the centre following the sequence by moving half way towards the corresponding amino acids which are positioned equally spaced on the circumference of a unit circle. The result of the complete execution of the protein sequence within the circle produces a typical graph for a particular protein, except for large protein sequences which are often found to have lesser visual benefits.

Bai and Wang [53] considered the triplet codon concept and using a complex co-ordinate scheme constructed a purine–pyrimidine graph on the left half of the complex plane, with purines (A and G) in the first quadrant and pyrimidines (T and C) in the fourth quadrant. A protein sequence can then be drawn from the triplet codons extending along the x-axis allowing visual inspection of the trends and also from the co-ordinates generate graph-theoretic matrices and their leading eigenvalues as descriptors of the sequences.

Bai and Wang [54] next proposed a 3D graphical representation for protein sequences where the 20 amino acids are represented as end points in a dodecahedron embedded in the 3D space, i.e. each amino acid is represented at one of the vertices of the dodecahedron. This allows construction of a sequence graph following the amino acids in the sequence where each point in the plot can be considered as a node of the graph, from which one can again generate matrices and sequence invariants.

Liao et al. [48] used a 2D graphical representation method to compare 24 coronavirus sequences where the four cardinal directions were associated with particular properties of the amino acids. They classified the 20 amino acids of a protein sequence into four separate groups according to the chemistry of their R groups: amino acids A,V,F,P,M,I,L to the hydrophobic chemical group; amino acids D,E,K,R to charged chemical group; amino acids S,T,Y,H,C,N,Q,W to polar chemical group; and the G amino acid to glycine chemical group. Starting with the nucleotide sequence, this enabled them to construct three 2D graphs (one for
each reading frame) for each gene sequence and compute a distance matrix. In a similar construction, Aguero-Chapin et al. [55] grouped the 20 amino-acids into four categories: acidic, basic, polar and non-polar and assigned the four groups to the four cardinal directions of a Cartesian frame to compute numeric descriptors of 108 sequences of polygalacturonases.

In recent years the field has progressed rapidly to numerically characterize protein sequences for application to different issues. González-Díaz and collaborators have extended these representations to the study of protein sequences [56] and applied to mass spectral data of proteins and protein serum profiles in parasites [57]. Gonzalez-Diaz has found that using different type of numerical indices derived from the protein 2D molecular graphics to perform QSAR studies is simpler than having to work with the protein 3D structures [58]. Integrated QSARs [59] developed using chemodescriptors for ligands and biodescriptors of a molecular entity connect structural information of drug molecules, DNA and RNA sequences or RNA secondary and protein tertiary structures. Basak et al. [60] using a new differential QSAR approach for study of dihydrofolate reductases (DHFR) from multiple strains of Plasmodium falciparum showed that DHFR from the wild strain is substantially different from four mutant strains of their study; this indicated that the protocols indicated in the paper can be used for the development of drugs to combat drug-resistant pathogens arising continuously in nature due to mutations. Nandy et al. [61] showed that their 20D graphical representation of protein sequences (explained later) was useful in generating phylogenetic relationships between sequences without necessity of multiple alignments and for determining conserved surface exposed stretches on viral proteins that could be useful in drug and vaccine designs [62].

We mention in passing that Randić [63], Basak and Gute [64] had developed mathematical techniques for analysis of proteomics data drawing parallels with DNA GRANCH techniques, but we do not go into any details about this topic in this review.

A detailed review of graphical representations of proteins including of proteomics has been made by Randić et al. [65].
APPLICATIONS OF GRANCH TECHNIQUES

Any new technique needs to be tested through applications to real problems and these methods of graphical representation and numerical characterization of biomolecular sequences are no exception. The intense interest which these GRANCH techniques have evoked amongst researchers have led to many and varied applications which shows the wide applicability and great potential of the methods. We cover some of these applications in brief here, with a novel application to anti-virus drug targeting in slightly more detail.

As a natural application of the graphical representation of DNA sequences, consider the visualization of patterns in base arrangements that are otherwise difficult to see in the normal character representation. As already mentioned, Gates [6] had noticed large scale repeats that were revealed by his 2D graphical plots and Nandy [5] showed that conserved genes have shapes on the 2D maps that are similar across species. From a detailed analysis of the graphical plots of families of conserved gene sequences that these altered with evolution such that the constituent bases appear to tend to greater homogeneity in base composition and higher complexity in base composition in the protein coding sequences [66].

Also, visual inspection of the graphical plots can enable new insights into similarities of different stretches of DNA sequences. Larionov et al. [67] had thus found long range palindromes in the mouse and human chromosomes. Nandy, Gute and Basak [68] reported on a stretch of the H5N1 avian flu neuraminidase gene that appeared to be well conserved among the various strains of the avian flu and reported on the possibility of using this site as drug or vaccine target so that these can be effective over many mutational changes (see below). Further observations and numerical computations on over 600 H5N1 neuraminidase sequences showed the wide dispersion and mutations of the gene sequences and especially the possible exchange of structural parts of the genes, which was a new observation for this type of virus [69].

Based on the observations of the plots of several conserved gene sequences, Nandy [70] showed that the base arrangements of these sequences could be conceived as bound by a characteristic function of the instantaneous population of the four bases as one moves along the sequence. Based on spot mutations, Nandy
proposed an equation connecting the instantaneous values of the purine and pyrimidine population asymmetries. It was hypothesized that this may have important consequences for genetic engineering since it implied that stability of engineered gene sequences required these constraints to be followed.

An important issue in molecular biology is identification of protein coding regions in DNA sequences. Nandy showed from the 2D graphical representations that exon and intron regions of mammalian genes showed distinctly different patterns and how these could be used to discriminate between the exons and introns [41]. This method was used by Ghosh et al. [42] to analyse a newly sequenced human chromosome III contig 7 DNA to identify coding regions and predict, using web-based tools, possible genes in the sequence. He, Li and Wang [31] used the numerical characterization of characteristic sequence representation of He and Wang [45] to suggest a protein coding gene finding algorithm specific for the yeast genome and found that the total number of protein coding genes in the yeast genome was 5897, which matches very well with estimates from other methods of 5800–6000. Discrimination between protein coding and non-coding regions was also proposed on an entropy-based approach [71] differentiating the DNA sequence into three subsequences and using Shannon’s formula.

Wiesner and Wiesnerova [72] did an interesting application of GRANCH techniques to study plant germplasm identificators. For their study of multiallelic marker loci from 18 Begonia × tuberhybridas, they used a 2D random walk digitization of the DNA sequences by three transform classes according to the prescription of Bai et al. [73] and derived invariants from the respective matrices to compute sequence similarities and dissimilarities. Principal component analysis done to compare the 18 marker loci to the DNA invariants found statistical correlations between the genetic diversity of the marker loci and the random walk invariants. Based on their results, the authors concluded that “DNA walk representation may function as an efficient pre-scanning procedure, which can predict allele-rich genomic loci as highly informative DNA markers solely using the information from their primary sequence.”

One of the early observations was that these graphical and numerical techniques allowed comparison of DNA and protein sequences without having to do multiple
sequence alignment since here we are dealing with numbers derived from the method rather than having to compare base by base or residue by residue. Almost all proposals of schemes for graphical representations have computed distances between DNA sequences to determine similarities and dissimilarities without multiple sequence alignment and obtained fairly good, though not uniform, results. For example, Liao et al. [47] used a 2D graphical representation proposed by Liao [74] to derive a phylogenetic tree from the elements of a similarity matrix for eleven mitochondrial gene sequences without having to go through any multiple alignment procedure. They constructed a 2x2 covariance matrix of the weighted centers of masses from the co-ordinates of each base of a sequence and computed the Euclidean distance between pairs of sequences to obtain their similarity/dissimilarity matrix.

Liao et al. [48] also investigated the phylogeny of 24 SARS coronavirus genomes by their 2D graphical representations for protein sequences where they could draw three plots for each sequence by considering the three reading frames. These generate three eigenvalues for each sequence which are then used to compute a distance matrix from which they could diagrammatically show the relationships of various strains of the virus. In another exercise, Bai and Wang [75] compared nine different neurocan nerve protein sequences in their 3D dodecahedron representation scheme. A direct comparison of these protein sequences through alignments is difficult since these protein sequences have different lengths. Using 10- and 35-component vectors from their model, they compared the distances between end-points of the vectors corresponding to each of the nine genes and built phylogenetic trees.

Nandy et al. [61] used their 20D representation of protein sequences to compute distances between sequences of the families of globin, the rat and human voltage gated sodium channel alpha subunit and their phylogenetic relationships. It is to be noted that deriving phylogenetic trees from protein sequences is usually a difficult matter when the sequences are of different lengths; but with the GRANCH techniques where D/D and other matrices can be computed for any length sequence and only the eigenvalues compared, the sequence length differences become irrelevant.
Jayalakshmi et al. [76] generalized these methods to compute alignment free sequence comparison using n-dimensional similarity space.

H Gonzalez-Diaz and his group have used 2D graphical methods for extensive work in the bio-medicine field. Based on pseudo-folding Lattice Network (LN) and Star-Graphs (SG) topological indices they proposed two DNA promoter QSAR models to predict promoter sequences in the function regulation of several mycobacterial pathogens [57]. Aguero-Chapin et al. [55] using their reduced four groups of amino acids on a 2D Cartesian co-ordinate framework computed numerical descriptors for 108 polygalacturonases through a Markov model and were able to discriminate between these and other proteins and predict polygalacturonase activity of a new protein.

Comparison of RNA secondary structures are important to understand their catalytic properties. Bai et al. [77] considered a 3D graphical representation of RNA characteristic sequences taken 2 bases at a time to compare similarities and dissimilarities in viral RNAs of nine species. They computed three modular lengths and three phases for each sequence from which they constructed a 6-component vector characteristic for each viral sequence. Two sequences were considered to be similar if their vectors pointed in the same direction and difference between sequences could be quantified by computing the Euclidean distance between the end points of the two vectors: the bigger the distance the less similar the sequence. The resultant difference table showed how methods such as these could be used to do cluster analysis without having to use alignment tools which are time consuming and requires several assumptions. In another instance, Gonzalez-Diaz et al. [78] has computed 2D-RNA coupling numbers by adapting the 2D graphical representation method for DNA sequences.

**Application of GRANCH Technique to Viral Proteins**

A novel application of GRANCH techniques was proposed by Ghosh et al. [62] to determine targets on viral proteins for drug and vaccine design. Viruses are known to mutate very fast and therefore become resistant to drugs and vaccine sin short time scales; the virulence of the avian flu led to an apprehension that it might mutate to a form that would enable human to human transmission of the disease and thus cause widespread infection and possible death as had happened
in the case of the Spanish Flu outbreak in 1918 when millions died. New drugs and vaccines, especially ones that could be readily moved from table to dispensaries were badly needed.

We had already noticed in early 2006 that certain parts of the neuraminidase gene appeared to be fairly well conserved [68]. The neuraminidase, along with hemagglutinin, are surface proteins that enable the viral particles to enter and leave the human cells where they proliferate, and of these the neuraminidase is the preferred target of the currently available drug, Tamiflu. We therefore determined to search the neuraminidase protein for surface residues that were well conserved.

Our procedure was to scan a small stretch of the neuraminidase protein sequences of 600+strains of the H5N1 virus and then slide the window by one base and scan again to calculate the protein graph radius in our 20D representation system. We know that these radii are very sensitive to any changes in the sequence, so equal values of the radii in one stretch over all the strains implied that this stretch was conserved. By covering the entire sequence for all strains we could get a good profile of regions of least variability.

The next step was to determine which parts of the sequences were surface exposed. There are several on-line engines available to scan a sequence and assign parameters to predict the degree of probability that certain portions were surface exposed. Matching these predictions with the hard facts we had on low variability we were able to identify six regions in the neuraminidase protein that were surface exposed and largely stable to mutational changes. These included the peptide we had identified earlier as being exceptionally stable. However, in a recent report on influenza virus RNA structure [79], it has been noted that the structures seen in the crystalline form may be one of several structural forms in vivo and confirmation will need to be experimentally determined.

The results of the analysis on the H5N1 neuraminidase protein sequence was published in 2010 [62]. Subsequently we have done a similar study on the VP7 protein of the rotavirus, a mainly tropical disease responsible for causing deaths to over half a million children every year. We identified four regions on the VP7 which appeared to be surface situated and quite stable. Our findings were reported
at the 2nd Mathematical Chemistry Workshop at Bogota, Colombia in 2010 [13] and the Indian Biophysics Conference, Delhi 2011 [80].

PROBLEMS AND PROSPECTS OF THE GRANCH TECHNIQUE

While a number of applications have shown the usefulness of the GRANCH approach to analyzing DNA, RNA and Protein sequences, this remains as yet a nascent field where many issues need to be looked into and problems resolved for the potential to be well realized. An early indication of some of these areas was outlined some years ago [81], but they are worth recapitulating along with some more issues that may bear scrutiny.

The intense interest in this field of graphical representation and numerical characterization of bio-molecular sequences have led to proposals for a vast array of models for depicting the sequences, some real and some virtual, more for DNA sequences, less for protein and RNA sequences. This has almost become an intellectual sport, with new ideas being propounded on regular basis, generally without a proper rationale for yet another method or critical comparison with earlier proposals. What appears to be lost in the process is the target: How useful are these representations to the practicing biologist? Critical to this issue is the problem of determining the domains of applicability of the various representations if different, i.e., which model is best suited to address which classes of problems. As of now, the vast majority of proposals have addressed themselves to comparisons of similarity and dissimilarity, but as we have seen in the previous section, the issues that we can address and which biologists need answers for are more varied.

From the applications made to date, the 2D graphical representations where the sequence data are easily viewed have generated the most interest. Even aside for the global characteristics revealed by the investigations of Jeffreys [3] and Peng et al. [3], the particular patterns of intron and exon segments [9] or characteristic curves of He et al. [31] have led to models to predict protein coding regions, determination of long-range palindromes [67], identification of target segments for vaccine development for viral proteins [62] and determination of allele-rich genomic loci for plants [72] among other applications have been based on 2D
representation schemes. Hamori had identified regions of sharp changes in base compositions from his 3D H-curves [25], but for almost all other 3D, 4D and higher dimensionality representations applications have been restricted to sequence similarities and generation of phylogenetic trees.

The mathematical technique involved in generating the descriptors and characterizers for DNA sequences are still at a preliminary level. While the first moments in the geometrical method for generating descriptors have generally yielded reasonable results in comparing intra- and inter-genic sequences, attempts to calculate higher moments to increase the accuracy and effectiveness of these descriptors have only lately begun [28, 82]. The leading eigenvalues from the Euclidean and graph theoretic distance ratios matrix have so far been used mainly to compute inter-sequence distances; given the rigorous mathematics of matrix mechanics, it may be worthwhile to try and extend the applications to other areas. For the benefit of users of these methods, it would be useful to have a comparison of the geometrical and graph theoretic models to determine at what level the two could give comparable results.

In the case of 2D graphical representations using Cartesian co-ordinates, we had seen that gene sequences take characteristic shapes [5]. This raises the possibility that some day we could create an Atlas of Gene Sequences where samples of each gene would be depicted and the descriptor parameters listed for easy reference and rapid visual identification. We have described quantitatively the gross features of the graphical plots in the 2D representations by using the first moments in a geometrical method [10, 11]; better descriptors can be determined through higher order moments [28] to quantify the curvature, skewness and other properties. These, and the leading eigenvalues from the graph theoretic approach, could be considered as a list of parameters describing the sequence, akin to the quantum numbers that are used to describe elementary particles. Such a scheme then provides a method to electronically store, retrieve and compare data between various sequences more efficiently, especially with a view to quickly scan newly sequenced DNA, RNA and proteins to determine the genes and functions.

We have considered the moments calculated from the geometrical approach to 2D graphical representations as numerical “descriptors” of the DNA sequence and
taken tentative steps to enhance the number of descriptors of a sequence by computing higher order moments to more completely describe the sequences. In the matrix method applied to different varieties of graphical representation, leading eigenvalues arising from the matrices have been taken as “invariants” of the sequence in the strict mathematical sense. However, the concept of invariants derived from these matrix methods of numerically characterizing DNA sequences may require some modification to account for the fact that DNA sequences constantly change due to mutations in the bases. The vast majority of these changes do not affect the functioning of the protein or the enzyme coded by the gene due to synonymous mutations in the coding segments or in the non-coding part; e.g. in the case of intronless gene like the neuraminidase of the avian flu H5N1 we had found [69] 447 out of total 682 sequences prevalent over the period 1997 to 2008 had undergone mutations in one or more bases in the gene, but even then, all of these variants coded for a functioning flu neuraminidase protein. For a beta globin gene, the common standard example of most graphical representation schemes, a sample from one person may differ by a base or more from the next person due to mutational changes. Determining an “invariant” from one sample sequence of these genes, while being mathematically precise, may not adequately express or characterize a gene sequence from a practical point of view. Perhaps a biologically more relevant measure would be a sampling of several such sequences and from them to compute an average eigenvalue with a standard deviation and derive a numerical to characterize the gene. In fact, in the absence of a sensitivity analysis or a standard deviation, it would be difficult to accept that the computations through leading eigenvalues of distances between several sequences that are only a few percentage points apart could be statistically meaningful. The descriptors are no exception either.

Once these basic issues are attended to, the GRANCH techniques can become a useful tool in the medicare field. Since the computations of the numerical descriptors/characteristics are quite simply done, they can be incorporated into the DNA sequencing schemes so that there will be automatic computations of, e.g., $g_R$ and $p_R$ values which would enable the physician to immediately ascertain the presence of any harmful genetic disorders; Huntington’s potential to degenerate into a disease for the patient, or some similar genetic problem areas could be easily read out
as the genome is sequenced provided we know the characteristic locus and have a standard genome, for example the readout for a normal person from the family, available for comparison. The viral application already discussed in detail in the previous section could be automated and extended to other viruses and bacterial genomes to promote new generation of drugs and vaccines. The researches of Gonzalez-Diaz [55, 58, 78] and Basak [59, 60] are already pointers to new directions.

Many potential application areas remain to be explored. Since the numerical descriptors mentioned previously are seen to be quite sensitive to changes in base composition and distribution, the potential exists to devise schemes to index various aspects related to the bio-molecular sequences. Initial attempts have been made to index toxic chemicals that have damaging effects on DNA sequences [83], and to index SNP gene sequences measured against some standard sequences [84]. However, these need to be refined and made more useful for the confidence to be generated for their use in laboratory situations.

One area that requires in depth study is how to address non-contiguous sequence segments. For example, in the case of epitopes, it is found that there could be continuous epitopes and discontinuous epitopes; in the latter case the folded protein brings residues from different parts of the amino acid sequence close together, which then become sites for the antibodies to act upon. The methods delineated so far for $g_R$ and $p_R$ or leading eigenvalue evaluation require contiguous span of the bases or residues for the numbers to be calculated. One way to circumvent this difficulty is to work on small segments of the sequence at a time as had been done in Ref. [62]. However, this is time consuming and inefficient, and more improved methods to be able to focus on regions of interest and calculate a minimum number of the parameters could offer better rewards.

CONCLUSION

In summary it is apparent that graphical representation and numerical characterization of molecular sequences hold far-reaching potential of rapidly analyzing the sequences to extract numerous information. It opens up new ways to look at these sequences, and to gain new insights such as long range palindromes, fractal properties and intra-purine intra-pyrimidine relationships not seen by any
other means. It allows one to compute many aspects of biological and medicinal interest and provide novel methods of tackling old problems; we have seen examples of gene identification, analysis of evolutionary trends and generation of phylogenetic trees, identification of conserved sites on viral proteins for drug and vaccine targeting, predict promoter sequences and new properties of polygalacturose proteins, among others and many possibilities remain unexplored, or barely scratched. Still, from plants to viruses, from mammalian genes to mitochondrial genomes, a varied series of applications have been formulated. Although many issues doubtless remain yet such as handling non-contiguous stretches of bases and residues like discontinuous epitopes, it is apparent that the GRANCH techniques hold a lot of promise for a new direction in molecular analysis.

ACKNOWLEDGEMENTS

I would like to express my gratitude to the Department of Science and Technology, Government of India for funding my research in its initial stages and to the Natural Resources Research Institute (NRRI), University of Minnesota, Duluth, USA for sponsoring my first trip to the NRRI in 1999. We are also thankful to the Consortium for Bioinformatics and Computational Biology, University of Minnesota, for funding the project awarded to Subhash Basak entitled "Numerical Characterization of DNA Sequences and Investigations into Structural aspects of Gene Additions and Mutations" that supported my second trip to NRRI in 2005 to do further research on the mathematical characterization of DNA sequences.

CONFLICT OF INTEREST

The author confirms that this chapter contents have no conflict of interest.

ABBREVIATION

GRANCH = Graphical Representation and Numerical Characterization

REFERENCES

[1] Nandy, A. Recent investigations into characteristics of long DNA sequences. *Ind. J. Biochem. Biophys.*, 1994, 31, 149-155.
Jeffrey, H. J. Chaos game representation of gene structure. *Nucleic Acids Res.*, 1990, 18, 2163–2170.

Peng, C.-K.; Buldyrev, S.V.; Goldberger, A.L.; Havlin, S.; Sciortino, F.; Simons, M.; Stanley, H.E. Long range correlation in nucleotide sequences. *Nature (London)*, 1992, 356, 168-170.

Voss, R. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.*, 1992, 68, 3805-3808.

Nandy, A. A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. *Curr. Sci.*, 1994, 66, 309–314.

Gates, M.A. A Simple way to look a DNA. *J. Theor. Biol.*, 1986, 119, 319–328.

Nandy, A. Graphical representation of long DNA sequences. *Curr. Sci.*, 1994, 66, 821.

Nandy, A. Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons. *Curr. Sci.* 1996, 70, 661–668.

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

Raychaudhury, C.; Nandy, A. Indexation Schemes and Similarity Measures for Macromolecular Sequences. Paper presented at the Indo-US Workshop on Mathematical Chemistry, Shantiniketan, West Bengal, India, December 1998.

Nandy, A. Indexing scheme and similarity measures for macromolecular sequences. *J. Chem. Inf. Comput. Sci.* 1999, 39, 243–247.

Randić, M.; Vracko, M.; Nandy, A.; Basak, S.C. On 3–D representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci.* 2000, 40, 1235–1244.

Nandy, A. A decade and more of GRANCH: Novel analysis of DNA and Protein sequences through Graphical Representation and Numerical Characterization techniques. Second Mathematical Chemistry Workshop of the Americas, Universidad de los Andes, Bogota, Colombia, July 19 - 24, 2010.

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

Berger, J.A.; Mitra, S.K.; Carli, M.; Neri, A. Visualization and analysis of DNA sequences using DNA walks. *J. Franklin Instt.*, 2004, 341, 37–53.

Nandy, A.; Harle, M.; Basak, S.C. Mathematical descriptors of DNA sequences: development and applications. *Arkivoc*, 2006, 9, 211-238.

Nandy, A.; Basak, S.C. New Approaches to Drug-DNA Interactions Based on Graphical Representation and Numerical Characterization of DNA Sequences, *Current Computer-Aided Drug Design*, 2010, 6, 283-289.

Ghosh, A.; Nandy, A. Graphical representation and mathematical characterization of protein sequences and applications to viral proteins. *Adv. in Prot. Chem. and Struc. Biol.*, 2011, 83, 1-43.

Chang, H.T. DNA Sequence Visualization. In *Advanced data mining technologies in Bioinformatics*, Hui-Huang Hsu Ed.; Idea Group Publishing: Hershey, PA, USA, 2006; pp 63-84.

Chun Li; Zhi-fu Wang, Z-f. Charcatrerizations of DNA Primary Sequences. In *Trends in DNA Research*, Woods, C.R. Ed.; Nova Science Publishers, Inc.: New York, USA, 2006; pp 123-144.

Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, Methods and Principles in Medicinal Chemistry, Mannhold, R.; Kubinyi, H.; Folker, G. Series Eds.; Wiley-VCH,Weinheim, Germany, 2009, Vol 41.
[22] Burma, P.K.; Raj, A.; Deb, J. K; Brahmachari, S.K. Genome analysis: A new approach for visualisation of sequence organisation in genomes. *J. Biosci., 1992, 17*, 395-411.

[23] Dutta, C.; Das, J. Mathematical characterisation of chaos, game representation: New algorithms for nucleotide sequence analysis. *J. Mol. Biol., 1992, 228*, 715-719.

[24] Baranidharan, S.; Sankaranarayanan, B.; Brahmachari, S.K. Chaos game representation of similarities and differences between genomic sequences. *Int. J. Genome Res., 1994, 1*, 309-319.

[25] Hamori, E.; Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem., 1983, 258*, 1318–1327.

[26] Leong, P.M.; Morgenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Applic. Biosci., 1995, 11*, 503–507.

[27] Nandy, A. Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons. *Curr. Sci. 1996, 70*, 661–668.

[28] Bielińska-Wąž, D.; Nowak, W.; Wąž, P.; Nandy, A.; Clark, T. Distribution moments of 2D-graphs as descriptors of DNA sequences. *Chem. Phys. Lett., 2007, 443*, 408–413.

[29] Guo, X.; Randić, M.; Basak, S.C. A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chem. Phys. Lett., 2001, 350*, 106-112.

[30] Yau, S. T.; Wang, J.; Niknejad, A.; Lu, C.; Jin, N.; Ho, Y. DNA sequence representation without degeneracy. *Nucleic Acids Res., 2003, 31*, 3078.

[31] He, P.; Li, C.; Wang, J. Finding Protein Coding Genes in the Yeast Genome Based on the Characteristic Sequences, *Internet Electron. J. Mol. Des. 2005, 4*, 613–624,

[32] Randić, M.; Vracko, M.; Lers, N.; Plavsic, D. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett., 2003, 371*, 202–207.

[33] Randić, M.; Vracko, M.; Zupan, J.; Novic, M. Compact 2-D graphical representation of DNA. *Chem. Phys. Lett., 2003, 373*, 558.

[34] Randić, M.; Lers, N.; Plavsic, D.; Basak, S. C.; Balaban, A. T. Four-color map representation of DNA or RNA sequences and their numerical characterization. *Chem. Phys. Lett., 2005, 407*, 205–208.

[35] Randić, M. Spectrum-like graphical representation of DNA based on codons. *Acta Chim. Slov. 2006, 53*, 477–485.

[36] Randić, M.; Vracko, M.; Nandy, A.; Basak, S. C. On 3–D representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci. 2000, 40*, 1235–1244.

[37] Li, C.; Wang, J. On a 3-D Representation of DNA Primary Sequences. *Combinatorial Chem. & High Throughput Screening 2004, 7*, 23.

[38] Chi, R.; Ding, K. Novel 4D numerical representation of DNA sequences. *Chem. Phys. Lett., 2005, 407*, 63.

[39] Liao, B.; Wang, T. Analysis of Similarity/Dissimilarity of DNA Sequences Based on Nonoverlapping Triplets of Nucleotide Bases. *J. Chem. Inf. Comput. Sci., 2004, 44*, 1666.

[40] Liu, Y. Z., Wang, T. M. Vector representations and related matrices of DNA primary sequence based on L-tuple. *Math. Biosci., 2010, 227*, 147–152.

[41] 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.

[42] Ghosh, S.; Roy, A.; Adhya, S.; Nandy, A. Identification of New Genes in Human Chromosome 3 Contig 7 by Graphical Representation Technique. *Curr. Sci. 2003, 84*(12), 1534-1543.
[43] Wang, W.; Liao, B.; Wang, T.; Zhu, W. A Graphical Method to Construct a Phylogenetic Tree. *International J. of Quant. Chem.*, 2006, 106, 1998–2005.

[44] Nandy, A.; Nandy, P. On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models. *Chem. Phys. Lett.*, 2003, 368, 102–107.

[45] He, P.A.; Wang, J. Characteristic Sequences for DNA Primary Sequence. *J. Chem. Inf. Comput. Sci.*, 2002, 42, 1080–1085.

[46] Song, J.; Tang, H. A new 2-D graphical representation of DNA sequences and their numerical characterization. *J. Biochem. Biophys. Methods* 2005, 63, 228.

[47] Liao, B.; Tan, M.; Ding, K. Application of 2-D graphical representation of DNA sequence *Chem. Phys. Lett.*, 2005, 414, 296-300.

[48] Liao, B.; Liu, Y.; Li, R.; Zhu, W. Coronavirus phylogeny based on triplets of nucleic acids bases. *Chem. Phys. Lett.*, 2006, 421, 313–318.

[49] Rani, M. Dynamics of protein evolution. *J Biosc.* 1998, 23(1), 47-54.

[50] Randić, M. 2-D Graphical representation of proteins based on virtual genetic code. *SAR & QSAR* 2004, 15, 147-157.

[51] Randić, M.; Zupan, J.; Balaban, A.T. Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.*, 2004, 397, 247–252.

[52] Randić, M.; Butina, D.; Zupan, J. Novel 2-D graphical representation of proteins. *Chem. Phys. Lett.*, 2006, 419, 528–532.

[53] Bai, F.; Wang, T. A 2-D graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.*, 2005, 413, 458–462.

[54] Bai, F.; Wang, T. On graphical and numerical representation of protein sequences. *J. Biomol. Struct. Dyn.* 2006, 23, 537–546.

[55] Aguero-Chapin, G.; Gonzalez-Diaz, H.; Molina, R.; Varona-Santos, J.; Uriarte, E.; Gonzalez-Diaz, Y. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L., *FEBS Letters* 2006, 580, 723-730.

[56] Aguero-Chapin, G.; Varona-Santos, J.; de la Riva, G. A.; Antunes, A.; Gonzalez-Vlla, T.; Uriarte, E.; Gonzalez-Diaz, H. Alignment-free prediction of polygalacturonases with pseudofolding topological indices: experimental isolation from Coffea arabica and prediction of a new sequence. *J. Proteome Res.* 2009, 8, 2122–2128.

[57] Gonzalez-Diaz, H.; Prado-Prado, F.; Ubeira, F. M. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. *Curr. Top. Med. Chem.* 2008, 8, 1676–1690.

[58] Gonzalez-Diaz, H.; Perez-Montoto, L. G.; Duardo-Sanchez, A.; Paniagua, E.; Vazquez-Prieto, S.; Vilas, R.; Dea-Ayuela, M.A.; Bolas-Fernández, F.; Munteanu, C.R.; Dorado, J.; Costas, J.; Ubeira F.M. Generalized lattice graphs for 2D-visualization of biological information. *J. Theor. Biol.* 2009, 261, 136–147.

[59] Basak, S.C.; Mills, D.; Gute, B. D.; Natarajan, R. Predicting pharmacological and toxicological activity of heterocyclic compounds using QSAR and molecular modeling. In *QSAR and Molecular Modeling Studies of Heterocyclic Drugs I*, Gupta, S. P. Ed.: Springer-Verlag, New York, 2006, pp 39-80.

[60] Basak, S.C.; Mills, D.; Hawkins, D. M. Characterization of dihydrofolate reductases from multiple strains of Plasmodium falciparum using mathematical descriptors of their inhibitors. *Chem. Biodivers.*, 2011, 8, 440–453.
[61] Nandy, A.; Ghosh, A.; Nandy, P. Numerical Characterization of Protein Sequences and Application to Voltage-Gated Sodium Channel Alpha Subunit Phylogeny. *In Silico Biol.* 2009, 9, 77-87.

[62] Ghosh, A.; Nandy, A.; Nandy, P. Computational analysis and determination of a highly conserved surface exposed segment in H5N1 avian flu and H1N1 swine flu neuraminidase. *BMC Structural Biology* 2010, 10, 6 doi:10.1186/1472-6807-10-6.

[63] Randić, M. On Graphical and Numerical Characterization of Proteomics Maps. *J. Chem. Inf. Comput. Sci.* 2001, 41, 1330-1338

[64] Basak, S. C.; Gute, B. D. Mathematical biodescriptors of proteomics maps: background and applications. *Curr. Opin. Drug Discov. Devel.*, 2008, 11, 320–326.

[65] Randić, M.; Zupan, J.; Balaban, A.T.; Vikic-Topic, D.; Plavsic, D. Graphical Representation of Proteins. *Chem. Rev.* 2011, 111, 790–862.

[66] Nandy, A. Investigations on Evolutionary Changes in Base Distributions in Gene Sequences, *Internet Electron. J. Mol. Des.* 2002, 1, 545–558, http://www.biochempress.com.

[67] Larionov, S.; Loskutov, A.; Ryadchenko, E. Chromosome evolution with naked eye: palindromic context of the life origin. *Chaos* 2008, 18, 013105.

[68] Nandy, A.; Basak, S.C.; Gute, B.D. Graphical representation and numerical characterization of H5N1 avian flu neuraminidase gene sequence. *J. Chem. Inform. Model.* 2007, 47, 945-951.

[69] Ghosh, A.; Nandy, A.; Nandy, P.; Gute, B.D.; 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.

[70] Nandy, A. Empirical relationship between intra-purine and intra-pyrimidine differences in conserved gene sequences. PLoS ONE 2009, 4, e6829.

[71] Li, C.; Wang, J. Relative entropy of DNA and its application. *Physica A* 2005, 347, 465-471.

[72] Wiesner, I.; Wiesnerová, D. 2D random walk representation of *Begonia × tuberhybrida* multiallelic loci used for germplasm identification. *Biologia Plantarum* 2010, 54(2): 353-356.

[73] Bai, F-L.; Liu, Y-Z.; Wang, T-M. A representation of DNA primary sequences by random walk. *Math. Biosci.*, 2007, 209, 282-291.

[74] Liao, B. A 2D graphical representation of DNA sequence. *Chem. Phys. Lett.*, 2005, 401, 196–199.

[75] Bai, F.; Wang, T. On graphical and numerical representation of protein sequences. *J. Biomol. Struct. Dynamics*, 2006, 23(5), 537-545.

[76] Jayalakshmi, R.; Natarajan, R.; Vivekanandan, M.; Natarajan, G.S. Alignment-Free Sequence Comparison Using N-Dimensional Similarity Space, *Current Computer-Aided Drug Design*, 2010, 6, 290-296.

[77] Bai, F.; Zhu, W.; Wang, T. Analysis of similarity between RNA secondary structures. *Chem. Phys. Lett.*, 2005, 408, 258–263.

[78] Gonzalez-Diaz, H.; Aguero-Chapin, G.; Varona, J.; Molina, R.; Delogu, G.; Santana, L.; Uriarte, E.; Podda, G. 2D-RNA-Coupling Numbers: A new computational chemistry approach to link secondary structure topology with biological function. *J. Comput. Chem.*, 2007, 28, 1049-1056.
[79] Gultyaev, A.P.; Fouchier, R.A.M.; Osthoorn, R.C.L. Influenza Virus RNA Structure: Unique and Common Features, *International Rev. Immunol.*, 2010, 29, 533–556.

[80] Ghosh, A.; Chattopadhyay, S.; Chawla-Sarkar, M.; Nandy, P.; Nandy, A. Characterization of Conserved Regions in Rotaviral VP7 Proteins: A Graphical Representation Approach towards Epitope Prediction. In: *7th Asian Biophysics Association (ABA) Symposium & Annual Meeting of the Indian Biophysical Society (IBS)*. Indian Habitat Center, New Delhi, India: ABAIBS-2011; 2011: 88.

[81] Nandy, A. Mathematical Characterization of DNA Sequences - Towards New Directions, CP963, Vol 1, *Computational Methods in Science and Engineering, Theory and Computation: Old Problems and New Challenges*, G. Maroulis and T. Simos, Ed. 2007, pp. 596-602.

[82] Bielińska-Wąż, D. Graphical and numerical representations of DNA sequences: statistical aspects of similarity. *J. Math. Chem.*, 2011, 49, 2345-2407.

[83] Nandy, A.; Basak, S.C. Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences, *J. Chem. Infor and Comput Sc* 2000, 40, 915-919.

[84] Nandy, A.; Nandy, P.; Basak, S.C. Quantitative Descriptor for SNP Related Gene Sequences. *Internet Electron. J. Mol. Des.*, 2002, 1, 367–373