Chapter 9
Alignment-Free Analyses of Nucleic Acid Sequences Using Graphical Representation (with Special Reference to Pandemic Bird Flu and Swine Flu)

Ashesh Nandy, Antara De, Proyasha Roy, Munna Dutta, Moumita Roy, Dwaipayan Sen, and Subhash C. Basak

Abstract The exponential growth in database of bio-molecular sequences have spawned many approaches towards storage, retrieval, classification and analyses requirements. Alignment-free techniques such as graphical representations and numerical characterisation (GRANCH) methods have enabled some detailed analyses of large sequences and found a number of different applications in the eukaryotic and prokaryotic domain. In particular, recalling the history of pandemic influenza in brief, we have followed the progress of viral infections such as bird flu of 1997 onwards and determined that the virus can spread conserved over space and time, that influenza virus can undergo fairly conspicuous recombination-like events in segmented genes, that certain segments of the neuraminidase and hemagglutinin surface proteins remain conserved and can be targeted for peptide vaccines. We recount in some detail a few of the representative GRANCH techniques to provide a glimpse of how these methods are used in formulating quantitative sequence descriptors to analyse DNA, RNA and protein sequences to derive meaningful results. Finally, we survey the surveillance techniques with a special reference to how the GRANCH techniques can be used for the purpose and recount the forecasts made of possible metamorphosis of pandemic bird flu to pandemic human infecting agents.

A. Nandy · A. De · P. Roy · D. Sen
Centre for Interdisciplinary Research and Education, Kolkata, India

M. Dutta · M. Roy
Dinabandhu Andrews College, Kolkata, India

S. C. Basak (✉)
Department of Chemistry and Biochemistry, University of Minnesota Duluth, Duluth, MN, USA
e-mail: sbasak@nrri.umn.edu

S. Singh (ed.), *Synthetic Biology*, https://doi.org/10.1007/978-981-10-8693-9_9

Keywords DNA/RNA sequence · BLAST · Numerical characterisations of DNA/RNA/protein sequences · Graphical representation and numerical characterisation (GRANCH) · Alignment-free graphical representation methods · Genes and genomic sequences · Molecular sequence similarity/dissimilarity · Beta-globin sequences · Mathematical descriptors · Graph theory · Graph-theoretical (topological) distance · Euclidean distance · Adjacency matrix · Distance matrix · D_E/D_G matrix · Graph invariant · L/L matrix · M/M matrix · 3D plot · Protein sequences · Spanish influenza · H1N1 flu virus · Pandemics · H5N1 avian flu · Negative-sense strand RNA virus · Hemagglutinin (HA) · Neuraminidase (NA) · Sialic acid residue · Antigenic drift · Antigenic shift · Mutations through recombination · Flu vaccines · Surveillance of flu · Ethics in surveillance

9.1 Introduction

Since the availability of large nucleic acid sequence data in the late 1980s, followed by the subsequent exponential growth of such data in the databases, the need for supportive tools to store, view, retrieve and analyse such sequences became and continues to be of urgent necessity. These tasks require methods to characterise the information contained in these sequences with a view to compare and contrast different sequences to determine their functions and place in the biological domain. The symbolic DNA sequence representation has been the legacy mainstay of such analyses carried over from the oligonucleotide sequence days, falling far short of the requirements of the broad sweeping overview of local and global base distributions, multiple sequence similarities and dissimilarities, molecular evolution and other studies necessary for an overarching assessment of the relationship of any DNA/RNA sequence to the biome through the accumulated data in the nucleic acid databases.

Sequence alignment has been the primary tool of choice for such comparison of generic sequences. However, in the conventional methods, the time complexity of sequence alignment varies as $\sim O(l^2)$ for two sequence segments of length l (Kobori and Mizuta 2015). In the realm of genomic sequences, this clearly consumes an enormous amount of computation time, which grows even worse when multiple alignments are necessary. The reason behind such sequence alignment is the fact that mutational changes in nucleic acid sequences through individual nucleotide mutations and rearrangements in nucleotide distributions brought on by various biological processes are reflected in the DNA/RNA sequences we observe today. This constitutes one of the main considerations facing biologists trying to understand the variations between different sequences of the same genera and their evolution (Qi et al. 2011).

Numerical characterisation of the DNA/RNA sequences is one of the main tools in undertaking this task, and a number of methods were developed to analyse specific attributes of DNA sequences. For example, the dot plot (Gibbs and McIntyre 1970) is used to visualise identical sequence segments between two proteins by the diagonal runs in a matrix of the constituents of the two sequences. Needleman and Wunsch (1970) used dynamic programming techniques to align protein or nucleotide sequences; another dynamic programming scheme was later proposed by Smith and Waterman (1981) (Smith et al. 1985), for improved scoring local sequence alignment, but the computational complexity and time requirements have limited its use. For protein secondary structure prediction, Chou and Fasman (1974a, b) developed an empirical technique, which, however, has been superseded by new machine learning techniques. In particular, dynamic programming techniques that analyse DNA sequences in terms of discrete small segments to eventually merge into the full sequence require time commitments of *~O(l^k)*, where *k* is the number of sequences to be aligned and is expensive in the use of computing resources, whereas fast Fourier transform (FFT) requiring resource time of *~O(l log l)* is only slightly better (Hanson, thesis 2003). Multiple alignment fast Fourier transform (MAFFT) with a simplified scoring system to reduce the computational overhead (Katoh et al. 2002), discrete Fourier transforms (Yin et al. 2014) and wavelet transforms (Dodin et al. 2000) that displays regularity in patterns in DNA sequences have been some of the other methods deployed to extraction of information of interest from the DNA databanks. BLAST (Basic Local Alignment Search Tool) is one of the most popular search tools for local alignments; it uses a heuristic algorithm that takes shortcuts to identify, very rapidly, sequences in a library that most closely resemble the query sequence (Altschul et al. 1990, 1997).

A more general approach that permits local and global sequence comparisons, generation of phylogenetic trees, numerical characterisations of DNA/RNA/protein sequences, quantitative estimation of sequence divergences and other properties and rational design of peptide vaccines is offered by novel, alignment-free schemes of graphical representation and numerical characterisation of biomolecular sequences (see review Nandy et al. 2006). The basic procedure in all these methods is to identify a vector with each type of nucleotide and plot each vector successively until the end of the sequence, thus drawing out a trajectory in the space of the model; associating a numerical value with the curve results in a descriptor of the sequence. This allows for quantitative comparisons between sequences and forms a powerful tool for analysis of similarities and dissimilarities between genes on a local and global scale. These novel approaches dispense with the need for alignments, which, for large sequences, become less reliable due to divergence of species and subsequent rearrangements (e.g. reversal, transposition or block exchange) occurring over evolutionary time scales; corrections introduced to take care of these divergences depend upon significant homologies between the sequences and the genes to be compared and thus become restrictive (Qi et al. 2011).

However, as we discuss later, these alignment-free graphical representation methods are not without their problems either. One of these is related to computation of the sequence descriptors. In geometric approaches (Raychaudhury and Nandy 1999; Liao and Ding 2006; Bielinska-Waz et al. 2007), computational closure in limited time is possible; however, in some of the approaches where sequence descriptors are defined through matrices of distances between the bases or amino acids of the sequence (Randic et al. 2000a, b; Song and Tang 2005; Wang and Zhang

2006; Li et al. 2016; Randic et al. 2004; Bai and Wang 2006), generating a final value for large sequences remains a significant problem, and approximations have to be resorted too. In some instances, defining a suitable matrix to characterise and an invariant to associate with a gene sequence presents some difficulties (Qi et al. 2011).

In this chapter, we present a concise summary of a selection of graphical representation methods to demonstrate the basics of this class of approaches of numerically characterising a biomolecular sequence. There have been a very large number of methodologies proposed to quantitatively estimate the differences between selected groups of genes and genomic sequences (see review Nandy et al. 2006). Extensions of these approaches have been made to take into account protein sequences (see review Randic et al. 2011) where the problem is compounded by the fact that the basic building blocks in protein sequences are constituted by 20 amino acids (Nandy et al. 2009) in comparison to nucleotide sequences where the basic building blocks constitute four bases; in terms of the methodologies, we consider a few in a short section and refer the readers to the original papers for greater details.

These graphical representation methods are found ready for acceptance in research. The visual representation of base distribution in DNA sequences was used by Larionov et al. (2008) to compare the chromosomal DNAs of human and mouse to find several examples of palindromic runs in almost the same manner in the two species. Wiesner and Wiesnerova (2010), in an innovative application of 2D graphical techniques to multiallelic marker loci from *Begonia* x *tuberhybrida* Voss, suggested that the DNA walks could be correlated with genetic diversity to predict new allele-rich loci and improve new DNA germplasm identificators. Applications of the graphical representation and numerical characterisation techniques were made by Liao and his group (Liao et al. 2005, 2006) to analyse the genomic sequences of the viruses in SARS (severe acute respiratory syndrome) pandemic of 2003 and generate their phylogenetic relationship tree. Humberto Gonzalez-Diaz and his group applied the underlying concepts in an analysis of protein sequences for drug discovery research through numerical parameters analogously to quantitative structure-activity relationship (QSAR) topological indices (Estrada and Uriarte 2001; González-Díaz et al. 2007), determination of mass spectral data of proteins, toxicoproteomics, toxicity prediction and diagnosis of cancer patients (Cruz-Monteagudo et al. 2008; Gonzalez-Diaz et al. 2008a, b; Aguero-Chapin et al. 2009).

In a number of interesting applications of the original 2D graphical representation of DNA/RNA sequences, one of the current authors of this chapter showed that introns and exons of mammalian genes differed significantly in base distribution characteristics (Nandy 1996a), that these properties can be used to determine coding regions in newly identified DNA sequences (Nandy 1996b) and that mutational changes in genetic sequences are guided by some restrictions which indicate an intricate relationship between intra-purine and intra-pyrimidine content of the sequences (Nandy 2009), among others.

A number of applications have been made over the years of some of these graphical representation and numerical characterisation methods, molecular sequence similarity/dissimilarity and phylogenetic relationships being among the ones with the most abiding interest. For example, Randic et al. (2000a, b) and Liao and Ding (2006) used a novel 3D graphical representation model to examine the base distributions in the first exons of beta-globin genes of 11 species giving results that matched closely with those obtained by other graphical representation methods. In another approach, Zhang et al. (2010) used a spectral representation of DNA sequences defining a 24-component vector as sequence descriptor and illustrated their method using complete beta-globin sequences of seven species. Li et al. (2016) introduced a novel method of considering dinucleotides in a 3D graphical representation model and obtained phylogenetic trees of 4 datasets: the coding sequences of the beta-globin genes of 18 species, the mitochondrial cytochrome oxidase subunit I (COI) genes of 9 butterflies, the S segments of 32 hantaviruses (HVs) and 70 complete mitochondrial genomes, all of which gave evolutionary relationships that matched with standard data.

In a slightly different application area, we refer in some detail to the applications of graphical representations and numerical characterisation schemes to viral genetics, focusing our attention more on the influenza virus in view of the several pandemics that it has caused, but we do mention in passing other viral pandemics also. It will be seen that some of the issues related to vaccine design and drug discovery are more easily and readily approached using graphical techniques, and these approaches with their visual representations also give fresh insights into base distribution and composition characteristics. These have led to determination of vaccine targets in the virions, tracking viral progress across geographical and time boundaries, computation of phylogenetic relationships and possible regulation of base composition and distribution in DNA/RNA sequences. Brief discussions of these applications and methods of surveillance of viral developments and progress are covered in another section of this chapter.

9.2 Graphical Representation and Numerical Characterisation (GRANCH)

Graphical representation and numerical characterisation (GRANCH) of nucleotide and amino acid sequences provides a very powerful tool for basic biological research. Mathematical descriptors based on graphical representations are used as a compact method for sequence comparison. The novelty, utility and low complexity of these representations have spawned a wide variety of approaches, some of which have seen many applications. A selection of such methods is given here in order to elucidate the many ways in which this technique can be formulated.

While most applications to date have been done in graphical representations in 2D and 3D spaces, several authors (e.g. Tang et al. 2010) have proposed representations in higher orders of dimensionality. However, there are some advantages and disadvantages to them; in a 4D graphical representation of nucleic acid sequences, while overlaps and intersections of the sequence curve with itself can be avoided, graphical visualisation and the ability to directly compare two sequences are lost which, on the other hand, are easily achieved in 2D or 3D plots. Liao and Wang (2004) proposed a 6D representation, while Randic et al. (2005) suggested a novel four-colour map representation.

Although graphical representation provides visual clues to the characteristics of base distribution in a sequence and enables visual comparison of similarities and dissimilarities between two or more sequences, quantitative estimation of such characteristics or differences between sequences is required for any significant comparisons. All authors of graphical representations have, therefore, recommended sequence descriptors to quantify sequence similarity/dissimilarity through the use of such descriptors. There are two approaches to define mathematical descriptors: geometrical and graph-theoretical methods.

9.2.1 Quantitative Estimation Methods

Various numerical descriptors for the numerical characterisation of sequences have been formulated during the past few decades. A particular descriptor maps the set of sequences (*S*) into the set of real numbers (*R*). In some cases, a vector may be extracted instead of a single number. Given below is a representative sample of the important and widely used descriptors known thus far. However, it should be noted that the list is not exhaustive.

9.2.1.1 Geometrical Methods

This method was first described by Raychaudhury and Nandy (1999), where they used the graphical representation of DNA sequence on a 2D rectangular grid, as explained below, using the (x, y) coordinate to derive the descriptor values. The first-order moments (μ_x, μ_y) and a graph radius g_R are defined for each sequence as:

$$\mu_x = \frac{\sum x_i}{N}, \mu_y = \frac{\sum y_i}{N} \text{ and } g_R = \sqrt{{\mu_x}^2 + {\mu_y}^2}$$

where (x_i, y_i) are the coordinates of each point on the plot and N is the total number of bases in the segment. g_R is the base distribution index which is dependent on the specific positions of the bases in a given sequence and has been found to be characteristic of the specific sequence base distributions. Thus, if two sequences yield the same g_R value, the two sequences will be found to be identical for all practical purposes (Nandy and Nandy 2003). Now, if μ_1 and μ_2 refer to two different DNA sequences, then the quantity Δg_R defined by:

$$\Delta g_R = \sqrt{(\mu_{1x} - \mu_{2x})^2 + (\mu_{1y} - \mu_{2y})^2}$$

provides an estimation of the difference between the two sequences. Thus, g_R and Δg_R are very important measures of sequence composition and distribution.

9.2.1.2 Graph-Theoretical Methods

In this method, DNA/RNA sequence is represented and characterised by a two-step process: representation of the sequences using graphs and characterisation of the graphs by graph invariants. A graph invariant is a graph-theoretic property which is preserved by isomorphic graph (Harary 1969; Janežič et al. 2007). A graph G is defined as an ordered pair consisting of two sets, V and G=(V(G), R), where V(G) represents a finite nonempty set of points and R is a binary relation defined on the set V(G). The elements of V are called vertices, and the elements of R also symbolised by E(G) or E are called edges. Such an abstract graph is commonly visualised by representing elements of V(G) as points and by connecting each pair (u, v) of elements of V(G) with a line or edge if and only if (u, v)ϵ R. The vertex, v, and edge, e, are incident with each other, as are u and e. Two vertices in G are called adjacent if (u, v)ϵ R, i.e. they are connected by an edge. A walk of a graph is a sequence beginning and ending with vertices in which vertices and edges alternate and each edge is incident with vertices immediately preceding and following it. A walk of the form v_0, e_1, v_1, e_2, . . ., v_n joins vertices v_0 and v_n. The length of a walk is the number of edges in the walk. A graph G is connected if every pair of its vertices is connected by a path. The distance (u, v) between vertices u and v in G is the length of the shortest path connecting u and v.

In a molecular graph corresponding to DNA/RNA sequences, V represents the set of nucleic acid bases (A, T/U, G, C) and E represents the set of bonds connecting the adjacent bases. For any pair of bases (i, j) in the sequence, (i, j)ϵ R, they are either connected (adjacent) or not. Such a graph may be represented by an adjacency matrix A= $\{\alpha_{i\,j}\}$ where:

$$a_{\text{ij}}\begin{cases} = 1 & \text{if } i, j \text{ are connected} \\ = 0 & \qquad\quad \text{otherwise} \end{cases}$$

Now the graph-theoretical distance (topological) D_G and the Euclidean distance D_E between the vertices are measured, and a particular type of matrix, D_E/D_G, is formed from which the leading eigenvalues are calculated. The leading eigenvalue is used to characterise a sequence, while the difference between the eigenvalues forms an estimate of the similarity/dissimilarity between the sequences. Thus, the leading eigenvalues of D_E/D_G matrix and the associated eigen matrices are considered to be descriptors of the DNA sequences.

To clarify the process of graphical representations and the differences between the multiple approaches, a short sequence of 10 bases, ATGAACACTC, is used as a standard example to generate plots using each of the methods mentioned here. The sequence denotes the stretch of first ten bases of the fourth segment of the influenza genome and the hemagglutinin (HA) gene, belonging to influenza A virus (A/Shenzhen/SP139/2014(H7N9)).

9.2.2 Graphical Representation and Numerical Characterisation Methods for Nucleotide Sequences

9.2.2.1 2D Graphical Representations

2D Rectangular Plot

A simple 2D Cartesian coordinate system is used to represent the nucleic acid sequences. Gates (1986), Nandy (1994) and Leong and Morgenthaler (1995) independently proposed the four cardinal directions on a 2D grid to represent the four bases. According to Nandy, starting from the origin, a point is plotted by moving one step in negative x-direction if the base is an adenine (A) and in the opposite direction if the base is a guanine (G) and a walk of one step in positive y-direction if the base is a cytosine (C) and in the opposite direction if it is a thymine (T) (or uracil, U). Plotting the points successively in this manner draws a graph of the sequence of bases. Gates and Leong and Morgenthaler proposed a similar graphical representation but with different assignments of the bases to the cardinal directions. The graph according to the Nandy prescription is given in Fig. 9.1.

Table 9.1 represents the coordinates, the centre of mass of all the points and the graph radius g_R of the sequence as described in Sect. 9.2.1.1 (Raychaudhury and Nandy 1999).

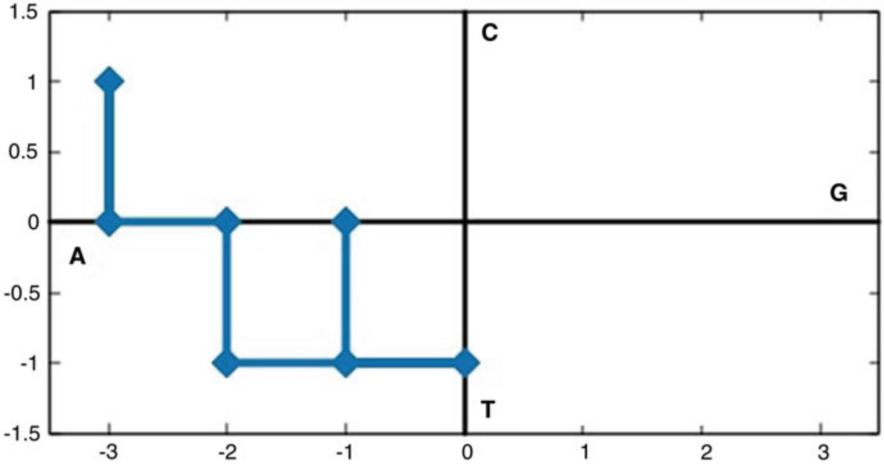

Fig. 9.1 Nandy plot of ATGAACACTC

Table 9.1 The co-ordinates, the centre of mass of all the points, and the graph radius g_R of the sequence as per Nandy rectangular plot

Base	Coordinates		Centre of mass		
	X	Y	μ_x	μ_y	g_R
A	−1	0	−1.9	−0.2	1.91
T	−1	−1			
G	0	−1			
A	−1	−1			
A	−2	−1			
C	−2	0			
A	−3	0			
C	−3	1			
T	−3	0			
C	−3	1			

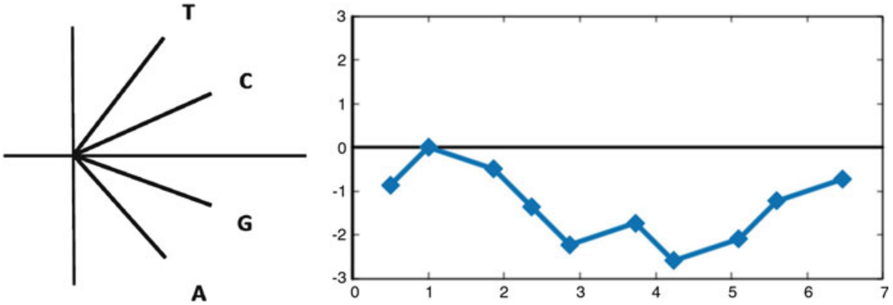

Fig. 9.2 Yau plot coordinate system and the graph of the sample sequence ATGAACACTC

Yau Plot

Although the 2D rectangular plots use a simple technique to easily visualise the sequence, it involves visual degeneracy that leads to an apparent loss of genetic information where repeated sequences, such as GAGAGAGAGAG, show overlapping paths. To overcome this, several authors proposed different modifications to the method which would reduce degeneracy (see review Nandy et al. 2006). Yau et al. (2003) proposed using two quadrants instead of four; cytosine (C) and thymine (T) are assigned to the first quadrant and adenine (A) and guanine (G) to the second quadrant, with the coordinates of the four bases defined as $\left(1/2, -\sqrt{3}/2\right)$ for A, $\left(\sqrt{3}/2, -1/2\right)$ for G, $\left(\sqrt{3}/2, 1/2\right)$ for C and $\left(1/2, \sqrt{3}/2\right)$ for T (Fig. 9.2).

While they did not prescribe a precise method for computing descriptors for the plots arising from their graphical method, we can, analogously to the Raychaudhury and Nandy (1999) method, define a centre of mass and a graph radius as descriptors. The results for our sequence in the Yau et al. (2006) plot are given in Table 9.2.

Table 9.2 The co-ordinates, centre of mass of all points and the graph radius as per the Yau et al plot

Base	Coordinates		Centre of mass		g_R
	X	Y	μ_x	μ_y	
A	0.5	−0.866	3.372	−1.336	3.627
T	1	0			
G	1.866	−0.5			
A	2.366	−1.366			
A	2.866	−2.232			
C	3.732	−1.732			
A	4.232	−2.598			
C	5.098	−2.098			
T	5.598	−1.232			
C	6.464	−0.732			

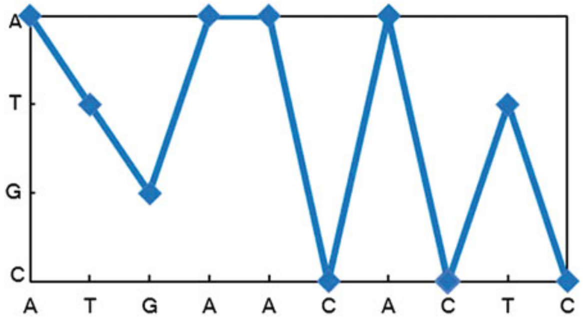

Fig. 9.3 Randic 2D plot for the sequence ATGAACACTC

Randic 2D Plot

There are many 2D approaches which do not involve the Cartesian coordinate system to represent nucleic acid sequence. Randic et al. (2003) proposed four horizontal equidistant parallel lines labelled as A, T, G and C from top to bottom. To plot the nucleic acid sequence, the bases are labelled along the x-axis, and straight lines are drawn from individual points on the four parallel lines according to the bases occurring in the sequence (Fig. 9.3).

A matrix method is used for determining a descriptor by constructing an M/M matrix for the given sequence ATGAACACTC. The off-diagonal entries of the M/M matrix are determined by dividing the Euclidean distance between two vertices of the zigzag curve by the graph-theoretical distance, i.e. the number of edges, between the two vertices. Here, if we take the first base A and the sixth base C of our sequence as an example, then the corresponding matrix element is obtained by taking the ratio of Euclidian distance between the first base (A) and the sixth base (C), i.e. $\sqrt{34}$, to the number of edges between A and C, which is 5 here. Proceeding in this way, we obtain the M/M matrix elements as shown below and the leading eigenvalue as the descriptor of the sequence (Tables 9.3, 9.4 and 9.5).

The leading eigenvalue in this case turns out to be 12.2388.

Table 9.3 The D_E matrix in Randic 2D model

	A	T	G	A	A	C	A	C	T	C
A	0	1.4142	2.8284	3	4	5.831	6	7.6158	8.0623	9.4868
T		0	1.4142	2.2361	3.1623	4.4721	5.099	6.3246	7	8.2462
G			0	2.2361	2.8284	3.1623	4.4721	5.099	6.0828	7.0712
A				0	1	3.6056	3	5	5.099	6.7082
A					0	3.1623	2	4.2426	4.1231	5.831
C						0	3.1623	2	3.6056	4
A							0	3.1623	2.2361	4.2426
C								0	2.2361	2
T									0	2.2361
C										0

Table 9.4 The D_G matrix in Randic 2D model

	A	T	G	A	A	C	A	C	T	C
A	0	1	2	3	4	5	6	7	8	9
T		0	1	2	3	4	5	6	7	8
G			0	1	2	3	4	5	6	7
A				0	1	2	3	4	5	6
A					0	1	2	3	4	5
C						0	1	2	3	4
A							0	1	2	3
C								0	1	2
T									0	1
C										0

Table 9.5 The M/M matrix in Randic 2D model

	A	T	G	A	A	C	A	C	T	C
A	0	1.4142	1.4142	1.0000	1.0000	1.1662	1.0000	1.0880	1.0078	1.0541
T		0	1.4142	1.1181	1.0541	1.1180	1.0198	1.0541	1.0000	1.0308
G			0	2.2361	1.4142	1.0541	1.1180	1.0198	1.0138	1.0102
A				0	1.0000	1.8028	1.0000	1.2500	1.0198	1.1180
A					0	3.1623	1.0000	1.4142	1.0308	1.1662
C						0	3.1623	1.0000	1.2019	1.0000
A							0	3.1623	1.1181	1.4142
C								0	2.2361	1.0000
T									0	2.2361
C										0

Song-Tang Plot

In a slight variation, Song and Tang (2005) proposed a three-horizontal line method where the central line represents two bases and the two peripheral lines denote each of the remaining bases. Counting of the nucleotides is done along the x-axis, and the

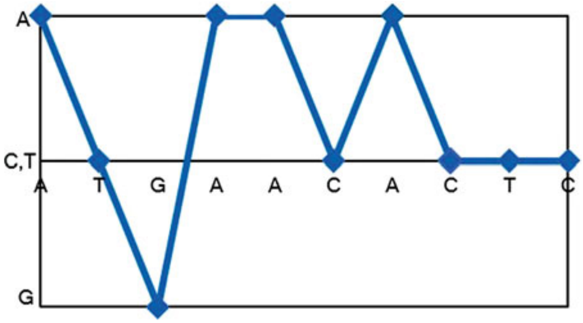

Fig. 9.4 One plot as per Song and Tang's representation of the sequence ATGAACACTC

bases are plotted and connected by straight lines as in the case of the Randic 2D plot. In the case of DNA primary sequences, the four bases A, C, G and T can be grouped as per their characteristics:

- Purine R = (A, G) and pyrimidine Y = (C, T)
- Amino M = (A, C) and keto K = (G, T)
- Weak H-bond W = (A, T) and strong H-bond S = (C, G)

Keeping one of such groups, such as purine, as central line, and cytosine (C) and thymine (T) as the peripheral lines, we can plot a graph. Repeating the same process for the other groups, we get a total of six (${}^4C_2$) combinations of plots. For each of them, the two peripheral lines can also be exchanged. In total, 12 (6×2) plots can be obtained.

In one such graph, where cytosine (C) and thymine (T) are assigned to the central line, adenine (A) to the uppermost line and guanine (G) to the lowermost line, the plot for the sequence ATGAACACTC is shown in Fig. 9.4.

Here, the matrix elements and the descriptor are obtained by using the Randic 2D approach as discussed earlier. Like in the Randic 2D plot, two new matrix relations, as defined below, M/M and L/L, are constructed. The leading eigenvalues are calculated from the M/M and L/L matrices which form descriptors of the corresponding DNA sequence. The matrix elements are formed as follows:

$$\begin{aligned}(\mathrm{ED})_{\mathrm{ij}} &= \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}\\ (M/M)_{\mathrm{ij}} &= (\mathrm{ED})_{\mathrm{ij}}/(\mathrm{GD})_{\mathrm{ij}} i \neq j\\ (M/M)_{\mathrm{ij}} &= 0\end{aligned}$$

$$\begin{aligned}&(\mathrm{PD})_{\mathrm{ji}} = (\mathrm{PD})_{\mathrm{ij}} = (\mathrm{ED})_{\mathrm{ii}+1} + (\mathrm{ED})_{i+1,i+2} + \ldots, + (\mathrm{ED})_{j-1,j} i < j\\ &(L/L)_{\mathrm{ij}} = (\mathrm{ED})_{\mathrm{ij}}/(\mathrm{PD})_{\mathrm{ij}} \quad i \neq j\\ &(L/L)_{\mathrm{ij}} = 0\end{aligned}$$

where ED, GD and PD are the Euclidian distance matrix, graph-theoretical distance matrix and path distance matrix, respectively (Tables 9.6 and 9.7).

The leading eigenvalue value in this case turns out to be 10.0552.

Table 9.6 M/M matrix

	A	T	G	A	A	C	A	C	T	C
A	0	1.4142	1.4142	1.0000	1.0000	1.0200	1.0000	1.0102	1.0078	1.0062
T		0	1.4142	1.1180	1.1180	1.0000	1.0200	1.0000	1.0000	1.0000
G			0	2.2361	1.4142	1.0541	1.1180	1.0200	1.0138	1.0102
A				0	1.0000	1.1180	1.0000	1.0308	1.02000	1.0138
A					0	1.4142	1.0000	1.0541	1.0308	1.0200
C						0	1.4142	1.0000	1.0000	1.0000
A							0	1.4142	1.1180	1.0541
C								0	1.0000	1.0000
T									0	1.0000
C										0

Table 9.7 L/L matrix

	A	T	G	A	A	C	A	C	T	C
A	0	1	1.2649	0.6708	0.7310	0.7405	0.6161	0.7278	0.7524	0.7729
T		0	1	0.6125	0.6800	0.7898	0.7870	0.7601	0.7871	0.8086
G			0	1	0.8740	0.6800	0.7374	0.6818	0.7174	0.7460
A				0	1	0.9262	0.7836	0.7864	0.8168	0.8398
A					0	1	0.7071	0.7453	0.7864	0.8168
C						0	1	0.7071	0.7836	0.8284
A							0	1	0.9262	0.9262
C								0	1	1
T									0	1
C										0

In this method the leading eigenvalues of M/M and L/L matrices are used as DNA descriptors. These numerical parameters will facilitate the comparison of two DNA sequences. A vector can be constructed to characterise a DNA sequence from these data obtained from different characteristic curves. For example, if we consider only L/L matrix and since two symmetrical characteristic curves have the same L/L matrices, we can construct a six-component vector as a DNA descriptor by using the leading eigenvalues of these matrices associated with six characteristic curves.

Wang-Zhang Plot

In the Wang-Zhang method (Wang and Zhang 2006), there are three configurations, namely, non-A, non-C and non-G. A binary method is employed where the presence or absence of bases in the sequence is assigned 0 and 1 according to the specific configuration, for example, 0 for A and 1 for the others if it is a non-A plot. A similar protocol is implemented for the non-C and non-G plots. The graphs for the sequence ATGAACACTC are shown in Fig. 9.5a–c.

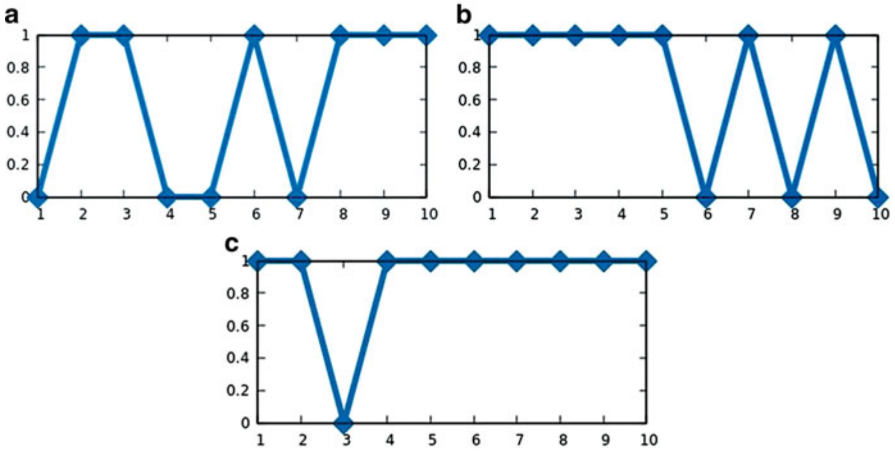

Fig. 9.5 (**a**) Non-A, (**b**) non-C and (**c**) non-G

Table 9.8 The co-ordinates and the leading eigenvalues calculated as per the Wang-Zhang model

	Coordinates			
Base	Non-A	Non-C	Non-G	Leading eigenvalues

A	0	1	1	Non-A = 0.3201
				Non-C = 2.1605
				Non-G = 4.1949

T	1	1	1	
G	1	1	0	
A	0	1	1	
A	0	1	1	
C	1	0	1	
A	0	1	1	
C	1	0	1	
T	1	1	1	
C	1	0	1	

In this model, the L/L matrices are constructed first from the three 2D graphs shown in Fig. 9.5 as a typical example to characterise the DNA sequence. Then the eigenvalues are computed from the matrices, and lastly, $\lambda_{\text{non-N}}$ are calculated which is $\lambda_{\text{non-N}}$ = maxeig (maximal eigenvalue) + mineig (minimal eigenvalue), where N = A, G, C. Proceeding in the same way, we obtain Table 9.8 for our sequence ATGAACACTC with the leading eigenvalues tabulated.

Yu-Hua Yao, Xu-Ying Nan and Tian-Ming Wang Plot

Yao et al. (2006) proposed a new 2D graphical representation of sequences, grouped as defined in section "Song-Tang plot", viz. W-S, M-K and R-Y. Here, we have plotted our sequence ATGAACACTC based on their system where the coordinates for the three curves are:

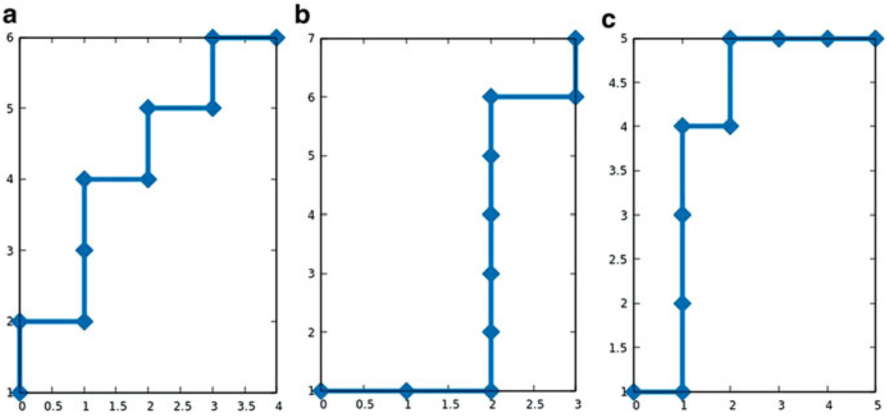

Fig. 9.6 (**a**) W-S curve, (**b**) M-K curve, (**c**) R-Y curve

Table 9.9 W-S curve

	Coordinates		Centre of mass		
Base	X	Y	μx		

The value of ρ for W-S curve = 1.4849

$$W-S\,\text{curve}\begin{cases} x=G_i+C_i \\ y=A_i+T_i \end{cases}$$
$$M-K\,\text{curve}\begin{cases} x=G_i+T_i \\ y=A_i+C_i \end{cases}$$
$$R-Y\,\text{curve}\begin{cases} x=C_i+T_i \\ y=A_i+G_i \end{cases}$$

The degeneracy is completely avoided in such representations. For our sample sequence, we get the plots as shown in Fig. 9.6a–c.

In this method, two sets of DNA descriptors are obtained. One is graph radius, subtending an angle θ with the x-axis, and the other is relative departure ρ which represents the difference between contents of bases of strong H-bonds and weak H-bonds in DNA sequence. Taking W-S curve as an example, the bisector is 50%(A, T)−50%S(G, C) line. A mean distance between the points in the DNA curve and this bisector is the new descriptor (Tables 9.9, 9.10 and 9.11):

Table 9.10 M-K curve

Base	Coordinates		Centre of mass		
	X	Y	μx	μy	gR
A	0	1	1.9	3.6	4.0706
T	1	1			
G	2	1			
A	2	2			
A	2	3			
C	2	4			
A	2	5			
C	2	6			
T	3	6			
C	3	7			

The value of ρ for M-K curve = 1.3435

Table 9.11 R-Y curve

	Coordinates		Centre of mass		
Base	X	Y	μ_x	μ_y	g_R
A	0	1	2.0	3.5	4.0311
T	1	1			
G	1	2			
A	1	3			
A	1	4			
C	2	4			
A	2	5			
C	3	5			
T	4	5			
C	5	5			

The value of ρ for R-Y curve = 1.0607

$$\begin{aligned}\rho &= 1/N\sum_{i=1}^{N}\sqrt{(x_i - i/2)^2 + (y_i - i/2)^2}\\ &= \sqrt{2}/2N\sum_{i=1}^{N}|x_i - y_i|\\ &= \sqrt{2}/2N\sum_{i=1}^{N}\sqrt{|(A_i + T_i) - (G_i + C_i)|}\end{aligned}$$

Li and Ji TB Curve

In the 2D graphical method reported by Ji and Li (2006), let $X=X_1X_2...X_n$ be a DNA primary sequence with n bases, and define a homomorphic map φ_1 by $\varphi_1(X) = \varphi_1(X_1)\varphi_1(X_2)...\varphi_1(X_n)$ as:

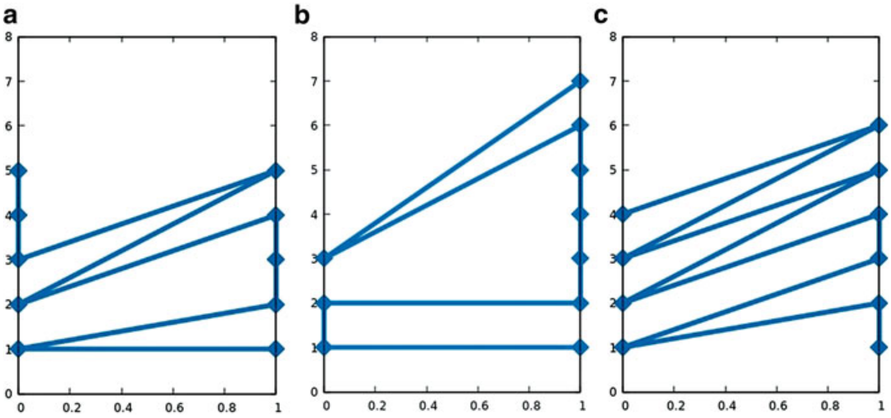

Fig. 9.7 (**a**) R-Y curve, (**b**) M-K curve, (**c**) W-S curve

Table 9.12 ED matrix

	A	T	G	A	A	C	A	C	T	C
A	0	1	1	2	3	1.4142	4	2.2360	3.1622	4.1231
T		0	1.4142	2.2360	3.1622	1	4.1231	2	3	4
G			0	1	2	1	3	1.4142	2.2360	3.1622
A				0	1	1.4142	2	1	1.4142	2.2360
A					0	2.2360	1	1.4142	1	1.4142
C						0	2.2360	1	2	3
A							0	2.2360	1.4142	1
C								0	1	2
T									0	1
C										0

$$\varphi_1(X_i) = \begin{cases} 1(1, R_i) \text{ if } X_i \in \mathrm{R} \\ 0(0, Y_i) \text{ if } X_i \in \mathrm{Y} \end{cases}$$

$$\text{where } (i = 1, 2, \ldots, n)$$

where R_i (Y_i) is the cumulative occurrence of the numbers of bases $\in R(Y)$ in the first i bases. The DNA sequence is mapped into a series of nodes $P_i^{'}s$. Connecting the adjacent nodes, an R-Y curve is obtained. In a similar fashion, the other two maps are defined, and the M-K and W-S-TB curves for the sequence ATGAACACTC are obtained (Fig. 9.7a–c).

Here, three matrices, namely, ED matrix, M/M matrix and L/L matrix, are formed as before from which the leading eigenvalues are calculated. These are the descriptors of the corresponding DNA sequence. The formation of the matrix elements has already been discussed in section "Song-Tang plot" (Tables 9.12, 9.13 and 9.14).

Table 9.13 M/M matrix

	A	T	G	A	A	C	A	C	T	C
A	0	1	0.5000	0.6667	0.7500	0.2828	0.6667	0.3194	0.3953	0.4581
T		0	0.7071	0.7453	1.0541	0.2000	0.6872	0.2857	0.3750	0.4444
G			0	0.3333	0.6667	0.2000	0.5000	0.2020	0.2795	0.3514
A				0	0.3333	0.2828	0.3333	0.1429	0.1768	0.2484
A					0	0.4472	0.1667	0.2020	0.1250	0.1571
C						0	0.3727	0.1429	0.2500	0.3333
A							0	0.3194	0.1768	0.1111
C								0	0.1250	0.2222
T									0	0.1111
C										0

Table 9.14 L/L matrix

	A	T	G	A	A	C	A	C	T	C
A	**0**	1	0.4142	0.5858	0.6797	0.2127	0.4076	0.1856	0.2433	0.2935
T		0	1	0.4142	0.7144	0.1770	0.4679	0.1810	0.2490	0.3065
G			0	1	1	0.2361	0.4055	0.1468	0.2103	0.2718
A				0	1	0.4370	0.3126	0.1158	0.1468	0.2103
A					0	1	0.1852	0.1852	0.1158	0.1468
C						0	1	0.1852	0.3125	0.4055
A							0	1	0.4370	0.2361
C								0	1	1
T									0	1
C										0

Zhao, Qi and Yang 2D Plot (Based on Nucleotide Triplets)

Zhao et al. (2015) proposed a 2D graphical representation method on the basis of nucleotide triplets (codons), as opposed to individual bases, in the DNA coding sequence strand. There are 64 codons including the 3 termination codons. Let $S = s_1 s_2 \ldots s_n$ represent a random DNA sequence. On the basis of standard genetic codes, each codon in the sequence can be defined with the aid of a mapping ϕ:

$$\phi(\ s_is_{i+1}s_{i+2})\begin{cases}(i,0)\text{if } s_is_{i+1}s_{i+2}=\text{TAA, TAG, TGA}\\(i,1)\text{if } s_is_{i+1}s_{i+2}=\text{ATG}\\(i,2)\text{if } s_is_{i+1}s_{i+2}=\text{GCT, GCC, GCA, GCG}\\(i,3)\text{if } s_is_{i+1}s_{i+2}=\text{CGT, CGC, CGA, CGG, AGA, AGG}\\(i,4)\text{if } s_is_{i+1}s_{i+2}=\text{AAT, AAC}\\(i,5)\text{ if } s_is_{i+1}s_{i+2}=\text{GAT, GAC}\\(i,6)\text{if } s_is_{i+1}s_{i+2}=\text{TGT, TGC}\\(i,7)\text{ if } s_is_{i+1}s_{i+2}=\text{CAA, CAG}\\(i,8)\text{ if } s_is_{i+1}s_{i+2}=\text{GAA, GAG}\\(i,9)\text{if } s_is_{i+1}s_{i+2}=\text{GGT, GGC, GGA, GGG}\\(i,10)\text{ if } s_is_{i+1}s_{i+2}=\text{ATT, ATC, ATA}\\(i,11)\text{if } s_is_{i+1}s_{i+2}=\text{CAT, CAC}\\(i,12)\text{if } s_is_{i+1}s_{i+2}=\text{TTA, TTG, CTT, CTC, CTA, CTG}\\(i,13)\text{ if } s_is_{i+1}s_{i+2}=\text{AAA, AAG}\\(i,14)\text{ if } s_is_{i+1}s_{i+2}=\text{TTT, TTC}\\(i,15)\text{ if } s_is_{i+1}s_{i+2}=\text{CCT, CCC.CCA, CCG}\\(i,16)\text{if } s_is_{i+1}s_{i+2}=\text{TCT, TCC, TCA, TCG, AGT, AGC}\\(i,17)\text{if } s_is_{i+1}s_{i+2}=\text{ACT, ACC, ACA, ACG}\\(i,18)\text{if } s_is_{i+1}s_{i+2}=\text{TGG}\\(i,19)\text{if } s_is_{i+1}s_{i+2}=\text{TAT, TAC}\\(i,20)\ \textit{if}\ s_is_{i+1}s_{i+2}=\text{GTT, GTC, GTA, GTG}\end{cases}$$

where $i = i^{th}$ is the base of sequence S. Applying this technique to our sequence ATGAACACTC, the resulting plot set is {(1, 1), (2, 0), (3, 8), (4, 4), (5, 17), (6, 11), (7, 17), (8, 12)}, which gives a 2_D curve termed as the considering codon degeneracy curve or CCD curve (Fig. 9.8).

Here the CCD curve is associated with a 21-dimensional characteristic vector V.

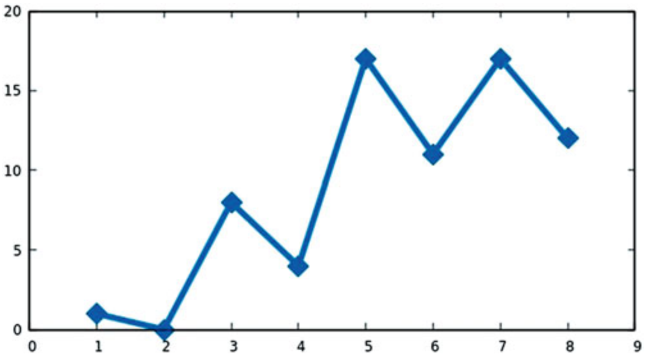

Fig. 9.8 CCD curve of the sequence ATGAACACTC using the method proposed by Zhao et al. (2015)

$$\begin{aligned}
v_0 &= 0,\\
v_1 &= \underbrace{1+1+\cdots+1}_{n_1},\\
v_2 &= \underbrace{2+2+\cdots+2}_{n_2},\ \ldots,\\
v_{20} &= \underbrace{20+20+\cdots+2}_{n_{20}}\\
V &= (v_0, v_1, \ldots, v_{20})
\end{aligned}$$

where the parameters n_1, n_2, …, n_{20} indicate the sequential trinucleotides of the mapping ϕ, respectively. For example, for the formula:

$$v_2 = \underbrace{2+2+\cdots+2}_{n_2}$$

v_2 is the summation of the y-coordinate in the CCD curve which equals to 2. The n_2 denotes the number of the trinucleotides, GCT, GCC, GCA and GCG. Extending the characteristic vector V, they propose two characteristic vectors V_a and V_b for an evolutionary distance computing scheme whose formula is:

$$\begin{aligned}
D(V_a, V_b) &= \frac{1 - V_a V_b / ||\ V_a||\ ||V_b||}{2}\\
&= \frac{1 - \sum_{i=1}^{21} v_a(i) \text{ x } v_b(i) / \sqrt{\sum_{i=1}^{21} (v_a(i))^2 \text{x} \sum_{i=1}^{21} (v_b(i))^2}}{2}
\end{aligned}$$

The distance of two CCD curves is represented by the distance $D(V_a, V_b)$ of two vectors, V_a and V_b, which can be used to quantify the evolutionary distance between sequences S_a and S_b. The two DNA sequences would be relatively similar if the D (V_a, V_b) was small. The smaller the distance, the more similar (relatively) are the two sequences.

For our sample sequence ATGAACACTC, the descriptor, vector V_a, can be defined as:

$$V_a = (0, 1, 0, 0, 4, 0, 0, 0, 8, 0, 0, 11, 12, 0, 0, 0, 0, 17, 0, 0, 0, 0),$$

whereas, for another sequence, ATGGTGCACC, the descriptor, V_b, is:

$$V_b = (0, 1, 2, 0, 0, 0, 6, 0, 0, 9, 0, 11, 0, 0, 0, 0, 0, 17, 18, 0, 20)$$

The distance can then be computed as:

$$D(V_a, V_b) = 0.26995$$

The small distance value implies that the two sequences are relatively similar.

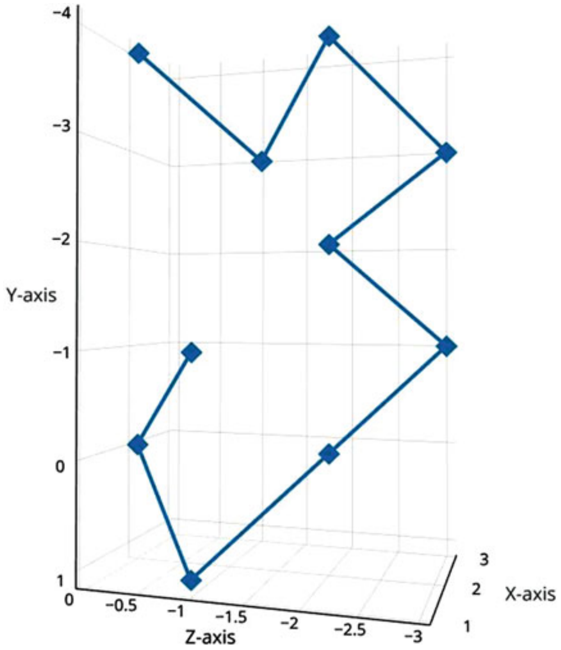

Fig. 9.9 3D plot as per Randic et al.'s representation of the sequence ATGAACACTC

9.2.2.2 3D Graphical Representations

Randic 3D Plot

The 3D graphical representation of DNA sequence was proposed by Randic et al. (2000). They associated the 2D method with a 3D graph by assigning each of the four bases to the corners of a tetrahedron where the points are as follows: (+1, −1, −1) for A, (−1, +1, −1) for G, (−1, −1, +1) for C and (+1, +1, +1) for T. The graph is plotted by placing the first base say A (for our mini sequence ATGAACACTC) at the corner position, i.e. at (+1, −1, −1), and the next base, i.e. T at (2, 0, 0). Proceeding in this manner, we obtain a plot of the sequence ATGAACACTC as shown in Fig. 9.9.

In this graphical plot of our mini sequence ATGAACACTC, a D_E/D_G matrix (from the Euclidean pairwise distance matrix D_E and a pairwise graph-theoretical distance matrix D_G of all the points on the graph) is constructed from which the leading eigenvalues are calculated to evaluate the descriptor of the sequence. In the following table, the coordinates of the bases of the sequence and the elements of the D_E/D_G matrix constructed from these coordinates are shown. Each matrix element is calculated as follows: Taking the (1,6) element, i.e. the first base A to the sixth base C as an example, the Euclidian distance is calculated from the coordinates shown in Table 9.15 and is divided by the minimum distance between two consecutive points

Table 9.15 The co-ordinates as per Randic 3D plot

Base	Coordinates		
	X	Y	Z
A	1	−1	−1
T	2	0	0
G	1	1	−1
A	2	0	−2
A	3	−1	−3
C	2	−2	−2
A	3	−3	−3
C	2	−4	−2
T	3	−3	−1
C	2	−4	0

Table 9.16 D_E/D_G matrix in Randic 3D model

	A	T	G	A	A	C	A	C	T	C
A	0	1	0.5774	0.3333	0.4082	0.2000	0.3333	0.2736	0.2041	0.2128
T		0	1	0.5774	0.6383	0.4083	0.5033	0.4303	0.2736	0.2887
G			0	1	1.000	0.6383	0.2000	0.6000	0.4303	0.4286
A				0	1	0.5774	0.6383	0.5774	0.3830	0.4303
A					0	1	0.5774	0.6383	0.4082	0.5033
C						0	1	0.5774	0.3333	0.4082
A							0	1	0.5774	0.6382
C								0	1	0.5774
T									0	1
C										0

which is √3. The division by the minimum distance is performed in order to normalise the distance scale, so that the Euclidian distance between adjacent vertices equals 1 and not √3 (due to taking the side of the cube to be 1). Now this value is divided by the number of edges before the sixth base, i.e. 5 here. Thus, the value obtained is = (√3/√3)/5 = 0.200 (Table 9.16).

The descriptor value, i.e. the leading eigenvalue, in the Randic 3D model for this sequence is 5.38566.

Li et al. 3D Plot

A novel 3D representation of DNA sequences was proposed by Li et al. (2016). If the bases are represented by a set {A, G, C, T}, they graphically characterised a sequence in terms of nucleotide pairs by taking two combinations of the set in terms of a multiset $\{\infty.\ A, \infty.\ G, \infty.\ C, \infty.\ T\}$. The possible number of such pairs can be ten (Table 9.17).

Let V be a tetrahedron with its centre represented by $O(0, 0, 0)$. The bases A, C, G and T are plotted at the vertices $V_1(1, 1, 1)$, $V_2(-1, -1, 1)$, $V_3(1, -1, -1)$

Table 9.17 Two combinations of multiset $\{\infty.\ A, \infty.\ G, \infty.\ C, \infty.\ T\}$

Bases	A	G	C	T
A	{A, A}	{A, G}	{A, C}	{A, T}
G		{G, G}	{G, C}	{G, T}
C			{C, C}	{C, T}
T				{T, T}

Table 9.18 The co-ordinates of the Li et al 3D plot

Point	Dinucleotide	X	Y	Z
1	AT	0	1	0
2	TG	0	1	−1
3	GA	1	1	−1
4	AA	1.5774	1.5774	−0.4226
5	AC	1.5774	1.5774	0.5774
6	CA	1.5774	1.5774	1.5774
7	AC	1.5774	1.5774	2.5774
8	CT	0.5774	1.5774	2.5774
9	TC	−0.4226	1.5774	2.5774

and $V_4(-1,\ 1,\ \ -1)$, respectively. For all the line segments formed between each pair of vertices, a midpoint is regarded as follows:

- M is the midpoint of AC.
- K is the midpoint of GT.
- R is the midpoint of AG.
- Y is the midpoint of CT.
- W is the midpoint of AT.
- S is the midpoint of CG.

Ten directions $\underset{\mathrm{OA}}{\rightarrow}, \underset{\mathrm{OC}}{\rightarrow}, \underset{\mathrm{OG}}{\rightarrow}, \underset{\mathrm{OT}}{\rightarrow}, \underset{\mathrm{OM}}{\rightarrow}, \underset{\mathrm{OK}}{\rightarrow}, \underset{\mathrm{OR}}{\rightarrow}, \underset{\mathrm{OY}}{\rightarrow}, \underset{\mathrm{OW}}{\rightarrow}, \underset{\mathrm{OS}}{\rightarrow}$ are obtained from which ten-unit vectors are produced as:

$$r_A = \underset{\mathrm{OA}}{\rightarrow} / \left\| \underset{\mathrm{OA}}{\rightarrow} \right\|, r_C = \underset{\mathrm{OC}}{\rightarrow} / \left\| \underset{\mathrm{OC}}{\rightarrow} \right\| \ldots \ \text{and so on.}$$

They associated each of the two combinations with the ten-unit vectors as:

$$\begin{aligned}&\{A, A\} \leftarrow r_A, \{A, G\} \leftarrow r_R, \{A, C\} \leftarrow r_M, \{A, T\} \leftarrow \ \ r_W,\\&\{G, G\} \leftarrow r_A, \{G, C\} \leftarrow r_S, \{G, T\} \leftarrow r_K,\\&\{C, C\} \leftarrow r_C, \{C, T\} \leftarrow r_Y, \{T, T\} \leftarrow \ \ r_T\end{aligned}$$

To plot our sequence ATGAACACTC with this method, we take two nucleotides at a time. Commencing from the origin, we move to the first dinucleotide AT, r_W and arrive at P_1, the first point on the 3D curve. Proceeding from P_1, we move towards the dinucleotide TG, r_K and reach the second point P_2. Progressing in this manner (Table 9.18), we get the curve as shown in Fig. 9.10.

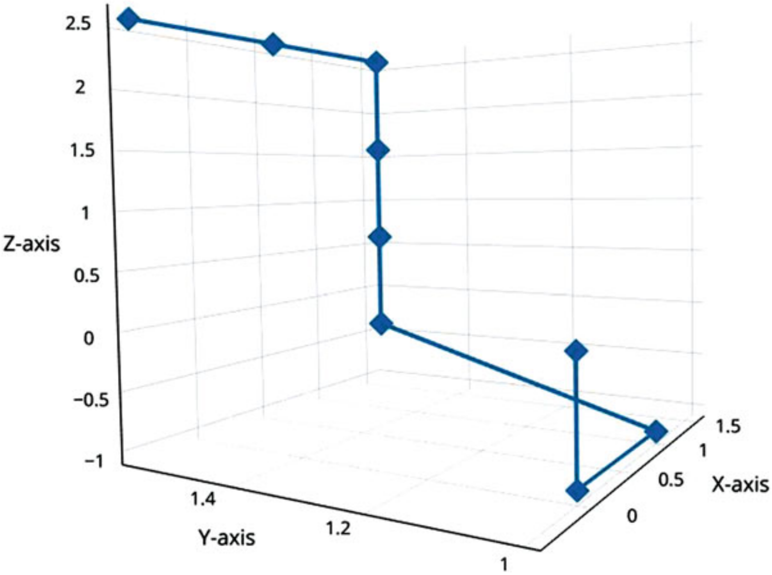

Fig. 9.10 3D dinucleotide plot method of Li et al. of the sequence ATGAACACTC

A common way to find the descriptor of DNA/RNA sequence is to construct a matrix from which the leading eigenvalue is calculated that serves as a descriptor for that sequence. But in this method, a different technique is introduced to calculate the DNA descriptor which is called ‘piecewise function’. A 3D representation having n vertices, in order of their appearance in the curve, is partitioned into K parts; each part is termed as a cell. Therefore, the i^{th} cell:

$$\underset{U_{i-1}U_i}{\xrightarrow{\qquad}} = (x_i - x_{i-1}, y_i - y_{i-1}, z_i - z_{i-1}\ )$$

where $U_0 = (0, 0, 0)$

A DNA sequence can be numerically characterised by a $3K$ dimension vector V_{tp}:

$$\begin{aligned} V_{\text{tp}} = \big(&x_1 - x_0, x_2 - x_1, \ldots, x_k - x_{k-1}, \\ &y_1 - y_0, y_2 - y_1, \ \ldots, y_k - y_{k-1}, \\ &z_1 - z_0, z_2 - z_1, \ldots, z_k - z_{k-1}\big) \end{aligned}$$

which in the case of our sample sequence will be:

$$\begin{aligned} V_{\mathrm{tp}} = (&0, 1, 0.5774, 0, 0, 0, -1, -1,\\ &0, 0, 0.5774, 0, 0, 0, 0, 0,\\ &-1, 0, 0.5774, 1, 1, 1, 0, 0) \end{aligned}$$

Now, for any two sequences S_a, S_b, if their descriptor values are $a = (a_1, a_2, \ldots, a_{3K})$ and $b = (b_1, b_2, \ldots, b_{3K})$,respectively, their similarity can be calculated by the Euclidean distance given by:

$$d(a, b) = \sqrt{\sum_{j=1}^{3k} \left(a_j - b_j\right)^2}$$

whereby, the smaller the $d(a, b)$, the more similar are the two DNA sequences.

9.2.2.3 Comparative Study of Results from Different Graphical Methods

The following table contains the list of values of the descriptors prescribed for each of the methods described above for comparative study based on the coding sequences of four human globin genes. Some of the result sets had to be trimmed to accommodate in the list. Brief descriptions are given below for the results listed of some of the methods to understand the computation.

Yao-Nan-Wang plot (section "Yu-Hua Yao, Xu-Ying Nan and Tian-Ming Wang plot"): Three curves had been proposed as part of this method (i.e. RY, MK and WS). Ideally each curve is described by a 3D vector (of θ, g_R and ρ), among which only the results for the WS curve have been listed.

Li-Ji plot (section "Li and Ji TB curve"): The method proposes three curves the same as in Yao-Nan-Wang one. Only the leading eigenvalues for the L/L matrices corresponding to the RY family are shown in here.

Zhao-Qi-Yang plot (section "Zhao Qi Yang 2D plot (based on nucleotide triplets)"): The method proposes a 21-dimensional vector, of which the 0th set, given no weight, always ends up with a value of zero. Thus, it can be considered as a 20-dimensional vector, effectively. Values for only positions 1st, 2nd, 19th and 20th are listed for representational purpose.

Li et al. plot (section "Li et al. 3D plot"): This method prescribes a unique descriptor using 'piecewise function'. It depends mainly on the average number of bases per sequence in a given dataset (in our case, this is 440.25) and a factor 'k' (logarithmically dependent on the aforementioned 'average' factor, which is ~4 in our case,). As per the proposed representation, the length of the representative vectors for our dataset would have been 12 for each sequence. Hence, for the sake of simplicity and a biting space crunch in hand, we did away with the critical representation and presented the endpoints of the prescribed 3D curve as a numerical descriptor for each sequence instead (Table 9.19).

Table 9.19 Sequence descriptors and sample parameters computed for four human globin genes for all the GRANCH methods discussed above

	Alpha-globin HBA1-mRNA (NM_000558)	Beta-globin-HBB-mRNA (NM_000518)	Delta-globin-HBD-mRNA (NM_000519)	Gamma-globin HBG1-mRNA (NM_000559)	Remarks
Nandy plot section	47.310	31.864	29.092	11.132	g_R value
Yau plot section	158.900	157.402	155.728	155.263	g_R value
Randic 2D plot section	430.861	445.751	445.723	445.992	D/D – leading eigenvalue
Song-Tang plot section	429.202	444.274	444.300	444.380	D/D – leading eigenvalue
Wang-Zhang plot section	22.840168	29.7082743	27.8319458	24.365207	L/L – leading eigenvalue
Yao-Nan-Wang plot section	[0.296, 94.274, 46.555]	[0.614, 82.260, 20.930]	[0.697, 80.733, 14.521]	[0.736, 80.213, 10.986]	[θ, g_R, ρ] of WS curve
Li-Ji plot section	88.328	78.169	66.411	35.411	L/L – leading eigenvalue for RY curve
Zhao-Qi-Yang plot section	[4,90…57,460]	[7,70…95,640]	[10,70…95,620]	[11,62…76,420]	1st, 2nd, 19th and 20th position values of the 20-dimensional vector
Randic 3D plot section	101.177	61.562	59.945	46.591	D/D – leading eigenvalue
Li et al. plot section	[−23.97, −103.064, 22.053]	[−3.762, −38.939, −41.309]	[0.97, −28.743, −52.66]	[15.083, −24.011, −6.691]	The 3D coordinates of the endpoints for the prescribed 3D curve

9.2.3 Graphical Representation and Numerical Characterisation Methods for Protein Sequences

9.2.3.1 Introduction

Graphical representation schemes for protein sequences came about as a natural consequence of the development and applications of graphical representations of nucleotide sequences. In contrast to that of nucleic acid sequences, graphical representation of protein sequences is complicated, given there are 20 amino acid residues as the basic units of proteins, while nucleic acids can be defined by only 4 bases. Initial efforts were made to portray the amino acids in terms of four properties so they could be represented in a 2D coordinate system like for the DNA sequences and later extended to multidimensional representations (see review Randic et al. 2011). We proceeded to simplify the representation intuitively by projecting the primary sequence in a 20-dimensional Euclidean space which can be used to understand the phylogenetic relationship, provide mutational insight and identify conserved regions for groups of proteins. We briefly mention just this one method in this chapter on nucleic acid research for completeness.

9.2.3.2 Design of the 20D Graphical Representation Algorithm

A 20-dimensional Cartesian coordinate system is used to model the protein sequences. The 20 amino acids are assigned to 20 axes in the Cartesian framework (Nandy et al. 2009); all the axes are equivalent, and the choice of association of amino acids with the axes can be made arbitrarily but once made will remain constant for the duration of the exercise. The computation starts from the origin and moves a step in the appropriate axis direction for the first amino acid in the sequence, another step in the respective axial direction for the next amino acid in the sequence and so on. Tracing the path for the occurrences of all the amino acids in the sequence in order results in a series of points in the abstract 20D space generating a 20D curve. An index for the protein sequence can be considered as a vector from the origin of the coordinate system to the endpoint of the graph. In order to eliminate the disadvantage that different distributions of the residues with the same composition in the sequences will lead to degeneracy in these vectors, the algorithm is designed to choose the characteristics of a sequence by a weighted centre of mass approach, first used for DNA sequence (Raychaudhury and Nandy 1999) and defined by:

$$\mu_1 = \frac{\sum x_1}{N}, \mu_2 = \frac{\sum x_2}{N}, \ldots\ldots\ldots, \mu_{20} = \frac{\sum x_{20}}{N}$$

where the $x_i{}'$s are the coordinate values of each point on the curve and N, a normalisation factor for the μ_is, is the number of amino acids in the protein sequence. Using these weighted averages, a protein graph vector $\overrightarrow{p_R}\,(\mu_1, \mu_2, \mu_3, \ldots\ldots\ldots, \mu_{20})$ and a protein graph radius p_R can be obtained:

$$p_R = \sqrt{\left(\mu_1^2 + \mu_2^2 + \ldots\ldots\ldots + \mu_{20}^2\right)}$$

Identical protein sequences have the same p_R values, which change with any alteration in the amino acid composition and distribution in the sequence. Δp_R is used to compare the differences between two protein sequences quantitatively as the Euclidean distance between two protein graph radii:

$$\Delta p_R = \sqrt{\left\{\left(\mu_1 - \mu_1'\right)^2 + \left(\mu_2 - \mu_2'\right)^2 + \ldots\ldots\ldots \left(\mu_{20} - \mu_{20}'\right)^2\right\}}$$

For multiple sequences, this method allows the comparison of pairwise differences, irrespective of sequence lengths. Thus, it can generate a distance matrix that could be used to construct phylogenetic trees, the advantage being that it does not require any multiple sequence alignments nor make any other model-dependent assumptions (Nandy 2009).

The 20D algorithms have been applied in phylogenetic analysis (Nandy 2009), mutational analysis of protein sequences on conserved surface accessible regions and drug target finding in influenza neuraminidase (Ghosh et al. 2010), rotavirus VP7 (Ghosh et al. 2012), Zika virus (Dey et al. 2017a) and others (Dey et al. 2016).

9.3 Applications of the New Methodology

We have briefly recounted in the introduction the various applications that have been made of alignment-free comparisons of DNA/RNA/protein sequences using the new methodology of graphical representation and numerical characterisation techniques. An area of particular concern that has seen many applications has been viral pandemics; specifically, in the twenty-first century, successive waves of epidemics and pandemics have swept across the globe starting with SARS (severe acute respiratory syndrome) pandemic of 2002–2003 that killed over 8000 people, the swine flu of 2009 (>100,000 deaths), the Ebola epidemic of 2014–2016 in West Africa (over 15,000 deaths) and the Zika virus pandemic that struck the Americas in 2015–2017 (>200,000 deaths). Understanding the characteristics of the viral spread and genetics would thus appear of urgent necessity. In this section we concentrate on the influenza pandemics as one of the most virulent killer diseases and summarise some of the specific applications of graphical representation methods to understand the character of the influenza virus, especially in relation to the pandemic varieties.

9.3.1 A Brief History of Flu Epidemiology

Viral outbreaks and epidemics have been recorded in historical archives from as far back as 400 BCE (Martin and Martin-Granel 2006). There have been written records of influenza outbreaks since the fifteenth century. The first cases of influenza in North America were possibly brought on by the European settlers in the late 1400s. The sixteenth century witnessed a flu pandemic spread from Russia to Europe via Africa (Pyle 1986), while England, Ireland and the USA were affected by three prominent outbreaks in the seventeenth century (Knobler et al. 2005).

The twentieth century saw one of the most horrific viral disasters of any era, viz. the 1918 'Spanish influenza'. This was a type A H1N1 virus that brought about an onslaught of worldwide pandemic, spreading to far corners of the world, including the Arctic and isolated islands in the Pacific (Mamelund 2011). By the end of the pandemic that had come in three waves between 1918 and 1919 (Taubenberger and Morens 2006), it had reportedly killed 50–100 million people (Patterson and Pyle 1991), mostly affecting the demographic age group of 20–40 years (Simonsen et al. 1998). During the peak of the pandemic in USA, more than 100,000 deaths were occurring in a week (Reid et al. 1999). It is interesting to note that post first infection wave, herd immunity within the population should have immunologically protected a second wave of infection in the same year. But on the contrary, the second wave drastically increased mortality rates. A key enigma, genetically, is the occurrence of the three infection waves in rapid succession within a period of 6–12 months: the H1N1 virus had to accumulate immense mutational changes or acquire key genetic alterations for it to become so fatal so rapidly which normally requires years of circulation within a population. However, in this case the virulence had increased over a matter of weeks (Reid et al. 1999).

Pandemics generally spread through trade routes when infected people travel or migrate from one country to another. The unprecedented proportions at which the Spanish influenza spread were only possible due to incidental military movements from one continent to the other for World War I (Erkoreka 2009).

This was followed by two more influenza epidemics in 1957 and 1968. The two viruses arose by reassortment of human and avian strains. The avian strain was from wild waterfowl found in Eurasia, and the human strain originated from the H1N1 strain that was already present in the population. It is speculated that the surface proteins of hemagglutinin and neuraminidase that had adapted to humans were replaced by those found in the avian strains (Reid et al. 2004). The 1957 epidemic, 'Asian influenza', originated in China and resulted in two million deaths caused by the type A H2N2 virus (Pyle 1986). The Asian flu virus next underwent antigenic shift to give rise to the H3N2 virus that brought about the 1968 'Hong Kong influenza' epidemic; in a report one million people died from the viral infection (Knobler et al. 2005). Towards the end of the century, in 1996, H5N1 virus was identified in China in geese which subsequently jumped the species barrier to infect poultry; some humans were incidentally infected in the 1997 wave and caused serious concern of another pandemic (De Jong et al. 1997). In the next few years, the H5N1 avian flu spread across various countries around the world and millions of wild as well as poultry deaths were seen. Human deaths arising from this avian influenza through physical contact were recorded globally till early 2014, but till date no human-to-human transmission has been observed, fortunately.

The twenty-first century saw the most recent influenza pandemic in 2009. The H1N1(2009)pdm strain, known as the swine flu, started from Mexico and spread rapidly round the globe, leading to a death toll of 106,000–400,000 (Dawood et al. 2012), before it died down. In 2014–2015, the poultry industry in the USA was faced by an epidemic of H5N2 bird flu that is believed to have been a reassortment with Asian H5 and North American N2 proteins, which could mutate to a human-infecting pandemic although no human infection has been noticed to date (Nandy and Basak 2015). In 2017–2018, flu season in the USA, an H3N2 type A influenza epidemic has affected thousands of people and led to the death of around 140 children, once again underlining the high pathogenicity of the influenza virus.

9.3.2 Genetics of Influenza Virus

Influenza virus belongs to the Orthomyxoviridae family. They are of four types, A, B, C and D, referred to as IAV, IBV, ICV and IDV, respectively (CDC 2017a). Types A, B and C infect humans and a variety of animals, among which type A is the most virulent. Influenza D virus was identified as recently as 2012. Although, IDV has been found to infect cattle (Hause et al. 2014; WHO 2018a, b), its detection in swine has raised the possibility of its spread in other mammals, including humans. Wild aquatic birds are the natural reservoir of the influenza A virus. The bird flu scare of around 1997–2005 was noteworthy because of the rapidity with which the virus spread throughout the globe causing huge fatalities among birds and animals; it still continues to precipitate frequent outbreaks. Because of the worldwide spread, virulence and pandemic history, IAV is of primary importance among the four types. In this chapter, we emphasise only on influenza type A virus except where otherwise stated.

Influenza is a negative-sense strand RNA virus. It is characterised by a segmented genome (Bouvier and Palese 2008) where the IAV has 8 segmented genes coding for 11 proteins, as described in Table 9.20. The influenza A virion is roughly spherical in shape, typically around 100 nanometres wide. It consists of an envelope containing the two surface proteins, hemagglutinin (HA) and neuraminidase (NA), in the proportion of 4:1 (Bouvier and Palese 2008). The central hollow core contains the RNA genome and the other proteins essential for packaging and survival of the virion.

IAV is categorised into several subtypes based on their surface-exposed proteins HA and NA, such as H1N1, H3N2 and H5N1, among others. So far, 18 subtypes of HA and 11 subtypes of NA have been observed on the basis of their antigenicity (Tong et al. 2013), out of which only a few subtypes are found to infect humans (Nandy et al. 2014).

Table 9.20 Gene segments of influenza A virus (A/duck/Vietnam/HU5-1571/2016(H5N1)) and their encoded proteins. While most proteins are coded by the gene segments, PB2-F1, M2 and NEP are expressed from spliced RNAs or alternate reading frames. (Based on the table in Bouvier and Palese 2008)

Segment	Gene	Code	Function of the protein	Gene length (nt)	Protein length (aa)
1	Polymerase basic subunit	PB1	mRNA cap recognition	2280	760
2	Polymerase basic subunit	PB2	RNA elongation, endonuclease activity	2274	758
		PB1-F2	Pro-apoptotic activity	273	91
3	Polymerase acidic subunit	PA	Protease activity	2151	712
4	Hemagglutinin	HA	Receptor binding, fusion activities	1704	568
5	Nucleoprotein	NP	RNA-binding protein, nuclear import regulation	1497	499
6	Neuraminidase	NA	Sialidase activity, virus release	1350	450
7	Matrix protein	M1	vRNP interaction, RNA nuclear export regulation, viral budding	759	252
		M2	Virus uncoating and assembly	294	97
8	Nonstructural protein	NS1	Regulation of host gene expression	678	225
		NEP/NS2	Nuclear export of RNA	366	121

HA mediates entry of the virus in the host cell by binding with terminal sialic acid residues of the host cell glycoprotein receptors. The HA protein has two subdomains, HA1 and HA2. HA1 subdomain contains the receptor-binding region and the antigenic regions where many mutations occur. HA2 subdomain is comparatively more stable and functions to anchor the protein on the viral envelope. NA mediates exit of the virus from the host cell by cleaving the terminal sialic acid residues of the glycoprotein or glycolipid receptors (Chen and Li 2013).

The host cells primarily include epithelial cells of the nose, throat and lungs of mammals and intestines of birds. Following binding, the viral lipid membrane fuses with the host cell membrane. The terminal sialic acid residues of the human cell receptors, where the HA binds, are of two types, α-*2-3* linked and α-*2-6* linked, having differential distribution in the lower and upper respiratory tract of humans, respectively (Kumlin et al. 2008). Distribution of the α-*2-3*-linked and α-*2-6*-linked sialic acid receptors in swine respiratory tract is much more uniform. In contrast, birds have only α-*2-3*-linked sialic acid receptor in their respiratory tract. This atypical distribution of sialic-linked glycoprotein receptors in the animal kingdom plays a significant role in evolution and dissemination of the virus. In China and the Far East, where poultry and swine herds are raised in numerous villages and farms, avian influenza with α-*2-3*-linked sialic acid receptors can infect swine and develop the ability to bind *α-2-6*-linked sialic acid receptors. This enables the modified virus to infect humans in the upper respiratory tract and make transmission between humans easy through sneezing and coughing. The scare with the bird flu, H5N1, was for this particular possibility – the virus has a high mortality ratio but to date can infect humans only through the *α-2-3*-linked sialic acid receptors. Were it to develop the ability to bind the *α-2-6* linked sialic acid receptors, humans would be susceptible to a pandemic catastrophe.

In another instance, the Spanish flu of 1918, known to be of avian origin (De 2018), is believed to have mutated to two forms (Reid et al. 2003). During the pandemic, two strains of H1N1 virus was co-circulating in nature. One strain was capable of binding *α-2-6*-linked sialic acid receptors present in the upper respiratory tract, having mutations at amino acid positions 190 (E190D) and 225 (D225G) in the HA1 segment (Reid et al. 2003). The other had the potential to bind both *α-2-6*- and *α-2-3*-linked sialic acid receptors present in the upper respiratory tract and lower respiratory epithelial cells, with mutation only at amino acid position 190 (E190D) in the HA1 segment (Reid et al. 2003).

RNA viruses are prone to very high rates of mutation (Barr and Fearns 2010; Drake 1993). The enzyme required for influenza gene replication in the host cell, RNA-dependent RNA polymerase, is devoid of any proofreading activity, which allows mutations during replication to be retained. Mutational changes in IAV are known to proceed by two well-documented mechanisms, antigenic drift and antigenic shift (CDC 2017b). Antigenic drifts arise out of random point mutations. If they occur in the viral antigenic site, the host cell may lose immunity against the virus, and the infection-immunity response cycle starts all over again. These drifts can cause seasonal outbreaks or epidemics, but they are not potent enough to give rise to pandemics. Antigenic shifts are less recurrent, occurring only when two or more virus subtypes infect a host cell. As the virus has segmented genome, the progeny virus may pack genes from different subtypes in a process known as reassortment. This may give rise to new and unique subtypes, sometimes introducing novel surface-exposed HA or NA that may predispose the viral host to high degree of pathogenesis. Antigenic shifts are often the causes of pandemics.

Interestingly, the Spanish flu virus is believed to have had greater than expected number of silent nucleotide or synonymous mutations (mutations that do not change the corresponding amino acids) in its genome in comparison with its avian and mammalian counterparts (Morens et al. 2010) and caused other pandemics through reassortments: the H2N2 pandemic of 1957, the H3N2 pandemic of 1968 and the H1N1 pandemic of 2009. The H5N1 bird flu is believed to have been a reassortment of genes from a prior H5N1 and a H9N2. Similarly, the H7N9 avian flu that struck China in 2013 (Gao et al. 2013) was a reassorted virus, picking up genes from multiple hosts like HA genes from H7N3 virus of domestic duck, NA from H7N9 virus of wild migratory birds and six other genes from the H9N2 virus of domestic poultry (CDC 2013; Liu et al. 2013). In 2015, there was an avian epidemic in the poultry market in the Midwest USA caused by H5N2 virus (Spackman et al. 2016), another product of reassortment which held possibilities of further reassortants.

There is also a third mechanism by which a virus may mutate to form a new strain. If a host cell is coinfected by two different strains of the same subtypes, the RNA polymerase may jump during replication from one RNA segment to its counterpart on the other strain at a consensus sequence of the gene referred to as a break point, continue replication and then revert to the original strain at the next break point. This phenomenon is termed as copy-choice recombination. This can give rise to progeny virus having a novel RNA segment not present in either of the infecting strains. We have also considered the possibility of the natural division of distinct segments of a gene, such as HA1 and HA2 of HA as break points, a topic we discuss in some detail later. Recombination in viruses is a controversial subject where the incidence of recombination through copy-choice method is expected to be below 2% (Hao 2011; Boni et al. 2012). However, a mechanism where whole segments of a gene are exchanged between two strains of the same subtype by polymerase jumps has occurred in about 5% of the cases examined (De and Nandy 2015).

9.3.3 Characterising the Flu Using Graphical Representation Methods

Quantitative sequence descriptors, like the g_R of the 2D rectangular representation, provide a ready means to follow the progress of typical viral strains across geographical and temporal domains. To this end we have investigated different aspects of the influenza virus. Some of these are briefly recounted below to indicate to the interested reader how such studies can be actually done.

Considering the high morbidity and mortality among the avian population arising from infections with the H5N1 bird flu virus around the turn of the century and the serious concern over possible mutagenesis of the strain to a human-to-human-infecting one, we undertook detailed comparisons of the H5N1 strains' neuraminidase sequences over a 10-year span to determine what changes were taking place (Nandy et al. 2007). Assuming that mutational changes in the sequences will be quantified by estimates of the graph radii, g_R, in a 2D graphical representation model (Nandy 1994), we assembled a database of 173 neuraminidase sequences for the years 1996–2005 and computed the average g_R values separately over periods when the flu was highly pathogenic to humans (1996–1997 and 2003–2005) and when less pathogenic (1998–2002). We found from computations of the g_R that, e.g. an H5N1 neuraminidase A/Chicken/Hong Kong/220/97(H5N1) had a very close similarity to another NA sequence from the same year, A/HongKong/156/97(H5N1), a result that matched with a conventional phylogenetic study by Suarez et al. (1998), whereas comparison with A/chicken/China/1/02-(H5N1)NA 5 years later showed that the two sequences were dissimilar almost seven times as much as the dissimilarity between the previous two sequences as estimated through Δg_R. This provides a descriptor-based insight into how rapidly the virus mutates over the years. A more detailed analysis with the g_R of the individual segments showed that the g_R expanded in the years of high pathogenicity to humans, more so for the period 2003–2005, whereas the period of low pathogenicity showed reduced values of g_R. This indicates a relative increase in the A, T component of the neuraminidase gene sequences in the pathogenic years, which could be a contributor to the high pathogenicity observed (Nandy et al. 2007).

A study of mutational changes in influenza strains such as of H5N1, H1N1, etc. is important for surveillance against the influenza subtypes building up resistance against the therapeutics and vaccines. Studying a group of 682 strains of the H5N1 virus over the years 1997–2008 (Ghosh et al. 2009), we came across an interesting phenomenon: How far a strain had travelled geographically and temporally. Considering the property that the same g_R of two sequences implies very close similarity of sequences (Nandy and Nandy 2003), we found that identical sequences have appeared over significant distances in space and time, raising questions about a virus' longevity. Based on the statistics, we found it compelling to hypothesise that virions could probably survive ex vivo (Ghosh et al. 2009, but also see Bean et al. 1982) in dried mud or dirt carried over long distances by wild birds to infect other birds and poultry in wetlands in distant areas.

The same investigations revealed an interesting facet of mutations through recombination. The role of recombination in viral sequence changes is hotly debated (Hao 2011; Boni et al. 2012), the general opinion being that if recombination does take place in RNA viruses at all through the copy-choice method, it would be below 2%. We enquired into the possibility of another style of recombination, complete segments of a gene being exchanged through polymerase jumps at the segment boundaries: the neuraminidase gene with three segments – transmembrane, stalk and body – is a good example. Taking A/chicken/Afghanistan/1573–65/2006(H5N1)NA as a test case (Ghosh et al. 2009) and computing g_R values for the three segments individually, we compared each segment with corresponding segments of a selected group of H5N1 neuraminidase genes in terms of the g_R values. We found that the g_R value for the transmembrane segment of this strain was identical, for example, with the g_R value of the transmembrane segment of A/turkey/Islamabad/NARC7873/2007(H5N1)NA, i.e. they have the exact same sequence, implying a segment exchange had taken place at some time. Similarly, the stalk segment was found identical in the test sample and in A/greatcrestedgrebe/Denmark/ 7498/06(H5N1) NA, and the body was found to have been duplicated in A/turkey/Islamabad-Pakistan/NARC-7871/02/2007(H5N1)NA, which implies that recombination-like events through exchanges of individual segments do take place (Ghosh et al. 2009).

To understand the extent of this phenomenon, we undertook a detailed survey of the hemagglutinin protein that is comparatively simpler in that it has two segments designated HA1 and HA2. Our study (De et al. 2016) involved a total of 1274 HA sequences comprised of H1N1, H3N2, H5N1 and H7N9 subtypes from Asia over the period 2010–2014. In this database we searched for sets of three strains where a HA1 from one strain and a HA2 from a different strain would combine to form a third, daughter strain (see schematic, Fig. 9.11). This was easily accomplished by computing g_R values for each segment for all the strains and comparing to find duplicates, with the proviso that the two parental strains were from the same time and place. We found a total of 73 daughter strains, but, interestingly, there were no

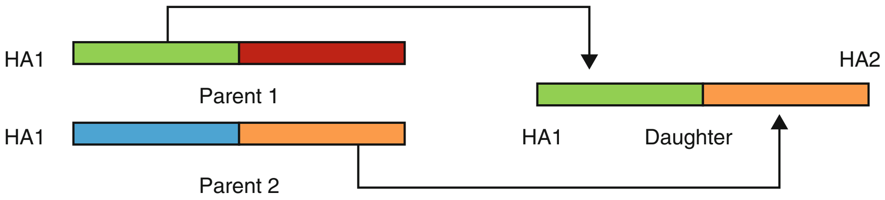

Fig. 9.11 Schematic of segment-wise recombination hypothesised for hemagglutinin. (Reproduced from Fig. 2 of De et al. (2016))

exchanges between different subtypes of the hemagglutinin. The daughter strains comprised 16 out of 427 H1 subtype, 30 of 408 H3 subtype, 20 of 350 H5 subtype and 7 of 89 H7 subtype in our database totalling 5.73% of all the strains in the database. Extending the analysis to 102 new strains from 2015 yielded another 5 instances of segment-wise recombination-like events. Thus, while the evidence for copy-choice recombination event mosaic in any gene sequence may not be readily evident, segment-wise recombination-like events do appear to occur in around 5% instances.

The two surface proteins, hemagglutinin and neuraminidase, have more in common: It appears that there is some kind of correlation between the two proteins such that suppression of the role of one influences changes in the sequence of the other and that coadaptation between the two was required for initiation of the H1N1 pandemic of 2009 (Wei 2010; Wagner et al. 2002; Xu et al. 2012; Hooper and Bloom 2013). There was also an observation that the influenza over the years had shown only a few subtypes instead of large numbers possible from the existence of 18 HA and 11 NA subtypes. Clearly, there was some interdependence between the two proteins, but there was no idea of what and how strong the interdependence could be. We undertook a study through graphical representation method (Nandy et al. 2014) and hypothesised a coupling between the HA and NA nucleotide sequences which could be measured through the g_R values. We assumed that subtypes that exist would have strong coupling and subtypes that are not found in nature would be weakly coupled, i.e. their interdependence would be weak, and hence those strains would not survive in nature. Our results (Nandy et al. 2014) amply bore this out and in fact showed that the couplings were stronger for avian flus compared to human-infecting flus, which one would have expected since wild birds seem to be the reservoir of influenza viruses of different kinds. We carried this analysis further and forecast the effects of possible reassortments of the H5N2 virus affecting poultry in the USA for therapeutic development and surveillance purposes (Nandy and Basak 2015).

As is evident from the above applications, g_R is a convenient way to scan hundreds and thousands of sequences in a short time interval to determine segments of interest. One particularly important query is identifying conserved segments in any gene sequence. By scanning the relevant sequences through a window of the appropriate size and determining the number of instances where the g_R values remain identical, the best conserved segments of a gene sequence can be identified.

Correspondingly, taking the protein sequence and computing the p_R values in the 20D model (see Sect. 9.2.3.2), the conserved domains in the protein can be identified. These techniques are an important tool in rational design of peptide vaccines where we may target those peptides that are conserved and have the capability of appropriate immune response to be the basis for vaccines that can outlive several generations of mutational changes. We had followed this procedure to identify best target sites in the NA of H5N1 (Ghosh et al. 2010) and the HA of H7N9 (Sarkar et al. 2015) influenza subtypes. Analyses for these targets are described briefly in the next section.

9.3.4 Drugs and Vaccines Against the Flu and Vaccine Design by Graphical Method

As in the case of most RNA viruses, drugs and vaccines against the influenza virus also face rapid obsolescence due to the high level of mutational changes. There have been several drugs that have been developed against the influenza, but only one remains effective at this time, while vaccines have to be changed every year.

Drugs used in influenza treatment globally belong to two groups having distinctive pharmacological attributes: M2 ion channel blockers and NA (neuraminidase) inhibitors (Stiver 2003). M2 protein forms a small proton channel across the viral envelope and mediates pH levels across viral membrane during cell entry and progeny maturation (Pielak and Chou 2011). It is vital for viral replication, and drugs amantadine and rimantadine were developed to block its function but has since been withdrawn because mutational changes in the prevailing influenzas have rendered them resistant to these drugs (CDC 2011). Present drugs available in the market are neuraminidase inhibitors are peramivir (trade name Rapivab), zanamivir (trade name Relenza) and oseltamivir (trade name Tamiflu). The neuraminidase inhibitors act both against IAV and IBV. However, according to the World Health Organisation (WHO), there is widespread resistance against oseltamivir globally (WHO 2009).

Vaccines that prepare the body's immune system against foreign pathogens form an alternative to drugs in the fight against viral diseases. Vaccines may be live-attenuated, inactivated or recombinant virus-like particle (VLP) type. With influenza there are so many different strains that finding a traditional vaccine against all types poses a perennial problem. Based on global surveillance data generated by WHO's in-house surveillance system, viz. the Global Influenza Surveillance and Response System (GISRS), vaccine strains are chosen for every year on the basis of preponderance of viral strains in each hemisphere. At present, for 2017–2018, WHO recommended four A and B strains, viz. A/Michigan/45/2015 (H1N1) pdm09-like virus, A/Hong Kong/4801/2014(H3N2)-like virus, B/Brisbane/60/2008-like virus (Victoria lineage) and B/Phuket/3073-like virus (Yamagata lineage), to be used as quadrivalent vaccine in the northern hemisphere (WHO 2017). The CDC also recommends vaccines for a year based on their assessment of prevailing virus strains in the previous year. While these do reduce the scope of the infection, in some years the vaccine is only partly effective as in 2014–2015 season and again in the current year, i.e. 2017–2018. These are partly due to the continuing mutations in the viral strains and partly errors in the vaccine as is reported to have happened this year.

It has therefore been a lingering quest to develop a universal flu vaccine that could act against all strains of the virus all the time. There have been claims off and on of a universal flu vaccine, but to date no such vaccine has been marketed. We undertook a graphical and numerical analyses of 514 strains of the H5N1 and 425 strains of the H1N1 influenza virus neuraminidase gene and determined six regions which remained conserved (Ghosh et al. 2010). This was facilitated by our examination of the graphical representation of the neuraminidase gene sequence which showed a strongly conserved 50/51 base (17 aa) region near the $3'$ end of the gene. When examined in the 3D crystallographic structure of the protein, this region was seen to be responsible for the bonding between the adjacent proteins in the neuraminidase quaternary structure and therefore extremely important for the structure's stability. The other regions we identified as being conserved also turned out to be surface exposed. These observations indicated a possibility, albeit subject to experimental verification, of designing inhibitors for broad-spectrum pandemic control of flu viruses with similar NA structure that could remain active for many generations of mutations of the underlying neuraminidase gene sequences (Ghosh et al. 2010). In another exercise, we explored the HA protein of H7N9 influenza virus using an improved protocol since there were concerns that the H7N9 could mutate to a human-infecting virus. We determined several targets in both the HA1 and HA2 regions of the hemagglutinin that could be used for rational design of peptide vaccines (Sarkar et al. 2015). It is to be noted that the HA1 of the hemagglutinin is highly variable, but HA2 is comparatively much more stable and is one of the targeted regions for a universal flu vaccine.

Another approach has recently been advocated for vaccine development. One of the disadvantages of current vaccines is their instability and weakness towards protease activity which mandates additional operational and transport costs; it also results in reduced biological activity and bioavailability to the immune system. Mile et al. (2018) constructed a T-cell-restricted IAV synthetic peptide containing D-amino subunits. D-amino acids are rarely found in nature and are inherently protected from protease degradation. Peptide was able to elicit an immune response in mice after oral administration. Such synthetic 'mimics' can be a cost-effective alternative to traditional vaccines.

9.4 Surveillance of the Flu

9.4.1 Background

With the high rate of mutations in influenza genome along with the high frequency and intensity of influenza epidemics and pandemics recurring every few years, surveillance of influenza genomic variations is an important issue for monitoring its changes and designing drugs and vaccines. Using influenza, hepatitis, poliomyelitis and malaria as examples, Alexander Langmuir expounded disease surveillance as the constant vigilance over the patterns and spread of disease occurrence via accumulation, integration, assessment and interpretation of data on morbidity and mortality along with other parameters and relevant information (Langmuir 1963). The Centers for Disease Control and Prevention (CDC) outlines surveillance to encompass identification of health problems for surveillance; collection, analysis and interpretation of data; dissemination of the data and its interpretations; and evaluation and improvement of surveillance (CDC 2006).

Surveillance of infectious viral diseases necessitates the analysis of antigenic constitutions of human-infecting viruses, as well as of viruses that belong to the same group which have not yet escaped their respective sylvatic cycles, in order to ascertain whether the molecular modifications have significantly altered known antigenic characteristics (Ghendon 1991). Some of the methods employed in viral discovery and analysis include tissue culture studies, immunohistochemical assays, singleplex and multiplex assays, serology and high-throughput sequencing (Lipkin and Firth 2013). One of the primary applications of disease surveillance and analysis is in the development of drugs and vaccines. The antigenic differences in the existing vaccine and the viral strains in circulation call for the synthesis of appropriate vaccine compositions to establish renewed immunity in the host population against epidemic- and pandemic-causing viruses (Smith et al. 2004). For surveillance of influenza, the Influenza Division at CDC collects specimens from over one million patients USA-wide which then follows a protocol for testing of the patient samples in various laboratories, followed by gene sequencing on approximately 6000 influenza viruses annually. Out of the pool of 6000 viruses, antigenic compositions are analysed in 2000 of them. The hemagglutination inhibition (HI) assay is used to identify the subtype of the hemagglutinin (HA) gene of a new influenza isolate on the basis of inhibition of hemagglutination by subtype-specific antibodies (Pedersen 2014). About 50 virus variants are filtered annually for potential vaccine production.

9.4.2 Application of Graphical Representation-Based Descriptors in Surveillance of Emerging Pathogens

Mathematical model analyses can extend the HI assays by quantitative estimation of antigenic differences; the results can also be reproduced visually to compare many strains at a time and check out significant differences. Examination of a number of plots of the influenza H5N1 neuraminidase gene in a 2D graphical representation system (Nandy 1994) led to the discovery of a 50-base segment at the 3′ end that was very strongly conserved (Nandy et al. 2007) and led eventually to a proposal for design of a peptide vaccine (Ghosh et al. 2010). This was found to be applicable across multiple subtypes of influenza, including H5N1, H1N1, H7N1, H9N1, H10N1 and others.

That same work (Nandy et al. 2007) showed another facet of viral surveillance. As described in some detail in Sect. 9.3.3, for this study 173 strains of the H5N1 of the period 1996–2005 were grouped on the basis of pathogenic years in the human population, two denoting periods of human infections and another representing the absence of human cases. Significant differences in the g_R were observed between the two groups of strains. The observed differences implied that genetic drifts had occurred among the strains of the two periods with the mutations having led to a quantitative decrease in the hydrophobicity of the proteins of strains belonging to the infective period. This paper illustrated that a relatively simple mathematical metric, viz. g_R, can be related to the location and effects of mutations in RNA, sort through conserved regions in the RNA, compute similarity between multiple sequences and thus aid in the surveillance of changes in the viral genomes.

Graphical representation can also help us understand the collected data which might contain incomplete information and annotation. Partial coding sequences can be graphically analysed to determine which part of the whole gene it belongs to through an alignment-free model. The approximate location of the sequence fragment can be determined in the gene by inferences from the nucleotide distribution pattern and the quantification of the pattern in form of g_R values. In their study of the Zika envelope gene, Dey et al. (2017a) evaluated partial CDS fragments and pinpointed their nucleotide start and end positions within the whole envelope gene sequence. Moreover, by cross-checking with existing annotated complete gene sequences, they were also able to identify additional peptide fragments in their sample sequences which contributed to the genetic differences between sequences collected from two different locations, Uganda and Brazil.

Extending the feature of determining sequence similarity from g_R values, which directly correlate with the nucleotide distribution in gene sequences, the graphical method can also be specifically employed to deduce the mixing between viruses originating from different regions due to homologous recombination. This is significant in understanding the geographical dispersal and trend of viral strain and subtype circulation, a key point in surveillance study of disease spread. A study (De et al. 2016) of the hemagglutinin (HA) gene of H1N1, H5N1, H3N2 and H7N9 subtypes from Asia, spanning the years from 2010 to 2014, showed that there were homologous recombinations of whole segments between the same subtypes. Some instances indicated that the parental and progeny strains were from same geographic regions, while others showed unexpected regional disparity, for example, parental H1N1 strains were found to be from Kowloon and Guangdong in China, and the daughter strain was isolated in Singapore. These genomic trails can be used to glean movement patterns of pathogenic viruses. In another study (Ghosh et al. 2009), it was found that H5N1 strains in geese from Qinghai, China, were also observed 5000 km away in strains in swans from Southern Russia in 2005; H5N1 strains were detected in poultry in Turkey that were also observed in chicken in Israel, via the trade route; and the same H5N1 strain were found in Egypt and Ghana, while H5N1 samples collected in 2006 from ducks in Hunan had the same sequence as found in human samples isolated from Indonesia.

Another important utility of the graphical representations is inferring the lineage of the strains. The g_R can be used to generate non-alignment-based phylogenetic trees that can indicate the evolutionary relationships between the viruses (Liao et al. 2005, 2006). Its application has also established new insights about virus families, exemplified by dengue type 2 virus (DENV2) which belongs to the flavivirus family along with Zika virus, yellow fever virus, West Nile virus and Japanese encephalitis virus, among others. The envelope gene sequences of DENV2 greatly deviate from the other aforementioned flaviviruses in their nucleotide distribution and composition. This was a new find compared to available literature which has classified dengue to be similar to the other flavivirus members. The change in their g_R values is significant (Dey et al. 2017b). This type of analytical observation will aid in surveillance of pathogens for antiviral and vaccine development.

These observations show the role of descriptors g_R in the surveillance of viruses. Earlier in the chapter, we described methods for calculation of other sequence descriptors. Different descriptors may encode different structural information on the sequences. So, a collection of sequence descriptors or orthogonal factors like principal components (PCs) derived from them may be more powerful tools for the comparison and surveillance of emerging viral pathogenesis. Such a critical analysis of different methodologies was done by us some time ago (Sen et al. 2016) and will be further developed and implemented in other areas.

9.4.3 Big Data and Social Networks in Surveillance Programmes

Dr. Tarun Weeramanthri introduced the term precision public health (Severi et al. 2014) and is now defined as the usage of computational and technological progress and big data to improve disease surveillance (Dolley 2018). The primary goal of precision public health is the accuracy involved in gathering data including emerging pathogens, reactions based on susceptibility and geographical distribution of the diseases. It makes use of real-time data generated, among other computational methods, to gather quick observable data. Traditional methods of forecasting influenza trends by CDC usually lag behind real time by 1–2 weeks, whereas information contained in cloud-based electronic health records and search queries in the Internet are typically available near real time. Yang et al. (2017) combine these cloud-based records and Internet searches with historical flu data and use dynamically selected set of variables to give the best fit in their model for the 2013–2016 flu season. Their results correctly estimated the peak timing and magnitude of the studies of flu season. These kinds of predictive models of influenza activity help public health officials prepare and allocate resources for possible disease outbreaks.

However, as with any individual's information, the use of public data calls for discretion: consent, privacy and security must be upheld and protected. Furthermore, an application that has come into prominence is the use of millions of data available from the social media platforms on the World Wide Web. However, the risk of inaccuracy in predictive analysis based on such data is high. A wide margin of error is required to compensate for over- or underestimation. Google Flu Trends, which is the recent surveillance tracking arm of the search engine giant, came into the spotlight for over-calculating the doctor visits for influenza-like illness than the reports generated by CDC which relies on records procured from laboratories (Lazer et al. 2014). The idea was to correlate search criteria with flu incidence and establish a trend. The initial trials matched 50 million searches with 1152 data points (Ginsberg et al. 2009) and came out with erroneous results due to certain ad hoc procedures. Even after corrections to the programmes, GFT persistently overestimated flu prevalence and has been relegated to the background for now in favour of more traditional procedures as adopted, for example, by CDC. Ideally, perhaps a combination of the two approaches could auger for better predictive trend analyses.

9.4.4 Ethics in Surveillance

With any scientific research, one must take into account the ethical implications of their work and inferences. A bone of contention in the prevention of spread of avian influenza and possible epidemics in human population is the recent trend in culling of poultry birds suspected to be infected by the H5N1 and H7N9 avian flu. The ethical dilemma arises between the prices of avian life against human population safety. Furthermore, WHO sheds light on the risks which the healthcare professionals place themselves in and also illuminates on the extent of travel restrictions that can be imposed in order to mitigate the spread of disease and the individual rights to freedom of movement. In 2006, WHO published a manual to integrate ethics with the influenza pandemic response structure (WHO 2007). The guide advocates sharing of surveillance information across borders during the pandemic, as well as prior and post pandemic. Resnik (2013) talks about the dilemma faced with the morality of dissemination of scientific knowledge, exemplified by the two censored and redacted papers, authored by Kawaoka and Fouchier Enserink 2012). Their papers showed the conclusions of genetically inducing mutations in H5N1 that conferred onto the viruses the ability to transmit through air among ferrets in the form of respiratory water droplets. The authors asserted that similar conditions among human populations would produce similar results. The NSABB, National Science Advisory Board for Biosecurity, allowed revised and heavily edited versions of the papers to be published in fear of dual-use research concerns (DURC) where scientific findings can be used for catastrophic implications. The NSABB, after much deliberation, recommended the full publication of their work.

9.5 Summary

In this chapter we have outlined powerful and novel alignment-free graphical representation and numerical characterisation (GRANCH) tools for comparative analyses of biomolecular sequences. We explained in brief some of the proposals put forward for graphical representation and numerical characterisation in various dimensionalities and mentioned some of the many applications done using these novel, alignment-free approaches. In particular, we concentrated upon several applications done through the simple and intuitive 2D graphical representation of Nandy (1994) to give an idea of how these approaches can be utilised to compare and contrast the several variants of the influenza virus, determine homologous recombinations between strains of the same influenza subtype, identify conserved segments in influenza gene sequences and design peptide vaccines. We have seen that these approaches can yield phylogenetic trees to understand the relationships between the various strains and subtypes and assist in focused surveillance to ensure advance knowledge of developments that could lead to epidemic and pandemic varieties. Graphical representation and numerical characterisation thus are very general but quantitative approaches that can be used to unravel myriad aspects of viral characteristics.

References

Aguero-Chapin G, Varona-Santos J, de la Riva GA, Antunes A, Gonzalez-Villa T, Uriarte E (2009) Alignment-free prediction of polygalacturonases with pseudofolding topological indices: experimental isolation from coffea arabica and prediction of a new sequence. J Proteome Res 8:2122–2128

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25 (17):3389–3402

Bai F, Wang T (2006) On graphical and numerical representation of protein sequences. J Biomol Struct Dyn 23(5):537–546

Barr JN, Fearns R (2010) How RNA viruses maintain their genome integrity. J Gen Virol 91 (6):1373–1387. https://doi.org/10.1099/vir.0.020818-0

Bean B, Moore BM, Sterner B, Peterson LR, Gerding DN, Balfour HH Jr (1982) Survival of influenza viruses on environmental surfaces. J Infect Dis 146(1):47–51

Bielinska-Waz D, Clark T, Waz P, Nowak W, Nandy A (2007) 2D-dynamic representation of DNA sequences. Chem Phys Lett 442:140–144

Boni MF, Smith GJ, Holmes EC, Vijaykrishna D (2012) No evidence for intrasegment recombination of 2009 H1N1 influenza virus in swine. Gene 494(2):242–245

Bouvier NM, Palese P (2008) The biology of influenza viruses. Vaccine 26(4):D49–D53. https://doi.org/10.1016/j.vaccine.2008.07.039

CDC (2006) Principles of epidemiology in public health practice third edition. Updated on May 2012. https://www.cdc.gov/ophss/csels/dsepd/ss1978/SS1978.pdf

CDC (2011) Antiviral agents for the treatment and chemoprophylaxis of influenza: recommendations of the advisory committee on immunization practices. https://www.cdc.gov/mmwr/preview/mmwrhtml/rr6001a1.htm

CDC (2013) Genetic evolution of H7N9 virus in China. https://www.cdc.gov/flu/pdf/avianflu/h7n9-reassortment-diagram.pdf

CDC (2017a) Types of influenza viruses. https://www.cdc.gov/flu/about/viruses/types.htm

CDC (2017b) How the flu virus can change: "Drift" and "Shift". https://www.cdc.gov/flu/about/viruses/change.htm

Chen L, Li F (2013) Structural analysis of the evolutionary origins of influenza virus hemagglutinin and other viral lectins. J Virol 87(7):4118–4120. https://doi.org/10.1128/JVI.03476-12

Chou PY, Fasman GD (1974a) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13:211–222

Chou PY, Fasman GD (1974b) Prediction of protein conformation. Biochemistry 13:222–245

Cruz-Monteagudo M, González-Díaz H, Borges F, Dominguez ER, Cordeiro MN (2008) 3D-MEDNEs: an alternative "in silico" technique for chemical research in toxicology. 2. Quantitative proteome-toxicity relationships (QPTR) based on mass Spectrum spiral entropy. Chem Res Toxicol 21:619–632

Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng PY, Bandaranayake D, Breiman RF, Brooks WA, Buchy P, Feikin DR, Fowler KB, Gordon A, Hien NT, Horby P, Huang QS, Katz MA, Krishnan A, Lal R, Montgomery JM, Mølbak K, Pebody R, Presanis AM, Razuri H, Steens A, Tinoco YO, Wallinga J, Yu H, Vong S, Bresee J, Widdowson MA (2012) Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. Lancet Infect Dis 12(9):687–695. https://doi.org/10.1016/S1473-3099(12)70121-4

De A, Nandy A (2015) An insight to segment based genetic exchange in influenza a virus: an in silico study. In: Proceedings of the MOL2NET, 5–15 December 2015; Sciforum Electronic Conference Series 1(015). https://doi.org/10.3390/MOL2NET-1-b015

De A (2018) Molecular evolution of hemagglutinin gene of influenza a virus. Front Biosci (Schol Ed) 1(10):101–118

De Jong J, Claas E, Osterhaus A, Webster R, Lim W (1997) A pandemic warning? Nature 389:554–554. https://doi.org/10.1038/39218

De A, Sarkar T, Nandy A (2016) Bioinformatics studies of influenza a hemagglutinin sequence data indicate recombination-like events leading to segment exchanges. BMC Res Notes 9:222. https://doi.org/10.1186/s13104-016-2017-3

Dey S, De A, Nandy A (2016) Rational design of peptide vaccines against multiple types of human papillomavirus. Cancer Informat 15(S1):1–16. https://doi.org/10.4137/CIN.S39071

Dey S, Nandy A, Basak SC, Nandy P, Das S (2017a) A bioinformatics approach to designing a Zika virus vaccine. Comput Biol Chem 68:143–152. https://doi.org/10.1016/j.compbiolchem.2017.03.002

Dey S, Roy P, Nandy A, Basak S, Das S (2017b) Comparison of base distributions in Dengue, Zika and Other Flavivirus envelope and NS5 genes. Paper presented at In: Proceedings of the MOL2NET, International conference on multidisciplinary *Sciences*, Sciforum electronic conference series 3. https://doi.org/10.3390/mol2net-03-04966

Dodin G, Vandergheynst P, Levoir P, Cordier C, Marcourt L (2000) Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. J Theor Biol 206:323–326. https://doi.org/10.1006/jtbi.2000.2127

Dolley S (2018) Big Data's role in precision public health. Front Public Health 6:68. https://doi.org/10.3389/fpubh.2018.00068

Drake JW (1993) Rates of spontaneous mutation among RNA viruses. Proc Natl Acad Sci USA 90:4171–4175

Enserink M (2012) Free to speak, Kawaoka reveals flu details while Fouchier stays Mum. Science. http://www.sciencemag.org/news/2012/04/free-speak-kawaoka-reveals-flu-details-while-fouchier-stays-mum

Erkoreka A (2009) Origins of the Spanish influenza pandemic (1918–1920) and its relation to the First World War. J Mol Genet Med: Int J Biomed Res 3(2):190–194

Estrada E, Uriarte E (2001) Recent advances on the role of topological indices in drug discovery research. Curr Med Chem 8:1573–1588

Gao R, Cao B, Hu Y, Feng Z, Wang D, Hu W et al (2013) Human infection with a novel avian-origin influenza A (H7N9) virus. N Engl J Med 368(20):1888–1897

Gates MA (1986) A simple way to look at DNA. J Theor Biol 11:319–328

Ghendon Y (1991) Influenza surveillance. Bull World Health Organ 69(5):509–515

Ghosh A, Nandy A, Nandy P, Gute BD, Basak SC (2009) 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 49(11):2627–2638. https://doi.org/10.1021/ci9001662

Ghosh A, Nandy A, Nandy P (2010) Computational analysis and determination of a highly conserved surface exposed segment in H5N1 avian flu and H1N1 swine flu neuraminidase. BMC Struct Biol 10(6)

Ghosh A, Chattopadhyay S, Chawla-Sarkar M, Nandy P, Nandy A (2012) In silico study of rotavirus VP7 surface accessible conserved regions for antiviral drug/vaccine design. PLoS One 7(7):e40749. https://doi.org/10.1371/journal.pone.0040749

Gibbs AJ, McIntyre GA (1970) The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. Eur J Biochem 16:1–11

Ginsberg J et al (2009) Detecting influenza epidemics using search engine query data. Nature 457:1012–1014. https://doi.org/10.1038/nature07634

González-Díaz H, Vilar S, Santana L, Uriarte E (2007) Medicinal chemistry and bioinformatics - current trends in drugs discovery with networks topological indices. Curr Top Med Chem 7:1025–1039

Gonzalez-Diaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008a) Proteomics, networks and connectivity indices. Proteomics 8:750–778

Gonzalez-Diaz H, Prado-Prado F, Ubeira FM (2008b) Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. Curr Top Med Chem 8:1676–1690

Hanson, RW (2003) Fast fourier transform analysis of DNA sequences. BA thesis, Reed College, Portland

Hao W (2011) Evidence of intra-segmental homologous recombination in influenza A virus. Gene 481(2):57–64

Harary F (1969) Graph theory. Addison-Wesley, Boston

Hause BM, Collin EA, Liu R, Huang B, Sheng Z, Lu W, Wang D, Nelson EA, Li F (2014) Characterization of a novel influenza virus in cattle and swine: proposal for a new genus in the Orthomyxoviridae family. MBio 5(2):e00031–e00014. https://doi.org/10.1128/mBio.00031-14

Hooper KA, Bloom JD (2013) A mutant influenza virus that uses an N1 neuraminidase as the receptor-binding protein. J Virol 87(23):12531–12540

Janežič D, Miličević A, Nikolić S, Trinajstić N (2007) Graph theoretical matrices in chemistry. CRC Press, Boca Raton

Ji M, Li C (2006) TB curve, a new 2D graphical representation of DNA sequence. J Math Chem 40 (2). https://doi.org/10.1007/s10910-006-9063-3

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30(14):3059–3066. https://doi.org/10.1093/nar/gkf436

Knobler SL, Mack A, Mahmoud A et al (eds) (2005) The story of influenza. The threat of pandemic influenza: are we ready? Institute of Medicine (US) Forum on Microbial Threats Workshop Summary. National Academies Press (US), Washington (DC)

Kobori Y, Mizuta S (2015) Similarity estimation between DNA sequences based on local pattern histograms of binary images. Biorxiv 1–17. https://doi.org/10.1101/016089

Kumlin U, Olofson S, Dimock K, Arnberg N (2008) Sialic acid tissue distribution and influenza virus tropism. Influenza Other Respir Viruses 2(5):147–154

Langmuir AD (1963) The surveillance of communicable diseases of national importance. N Engl J Med 268:182–192

Larionov S, Loskutov A, Ryadchenko E (2008) Chromosome evolution with naked eye: palindromic context of the life origin. Chaos 18:013105

Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google flu: traps in big data analysis. Science 343(6176):1203–1205. https://doi.org/10.1126/SCIENCE.1248506

Leong PM, Morgenthaler S (1995) Random walk and gap plots of DNA sequences. Comput Appl Biosci 11:503–507

Li C, Fei W, Zhao Y, Yu X (2016) Novel graphical representation and numerical characterization of DNA sequences. Appl Sci 6:63. https://doi.org/10.3390/app6030063

Liao B, Wang T (2004) Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases. J Chem Inf Comput Sci 44:1666–1670

Liao B, Ding K (2006) A 3D graphical representation of DNA sequences and its application. Theo Comput Sc 358:56–64

Liao B, Tan M, Ding K (2005) Application of 2-D graphical representation of DNA sequence. Chem Phys Lett 414:296–300

Liao B, Xiang X, Zhu W (2006) Coronavirus phylogeny based on 2D graphical representation of DNA sequence. J Comput Chem 27:1196–1202

Lipkin WI, Firth C (2013) Viral surveillance and discovery. Curr Opin Virol 3(2):199–204. https://doi.org/10.1016/j.coviro.2013.03.010

Liu D, Shi W, Shi Y, Wang D, Xiao H, Li W, Bi Y et al (2013) Origin and diversity of novel avian influenza a H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. Lancet 381(9881):1926–1932. https://doi.org/10.1016/S0140-6736(13)60938-1

Martin PM, Martin-Granel E (2006) 2,500-year evolution of the term epidemic. Emerg Infect Dis 12 (6):976–980

Mamelund SE (2011) Geography may explain adult mortality from the 1918-20 influenza pandemic. Epidemics 3(1):46–60. https://doi.org/10.1016/j.epidem.2011.02.001

Mile JJ, Tan MP, Dolton G, Edwards ESJ, Sae G, Laugel B et al (2018) Peptide mimic for influenza vaccination using nonnatural combinatorial chemistry. J Clin Invest 128:1–12. https://doi.org/10.1172/JCI91512

Morens DM, Taubenberger JK, Harvey HA, Memoli MJ (2010) The 1918 influenza pandemic: lessons for 2009 and the future. Crit Care Med 38(4):e10–e20

Nandy A (1994) A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes. Curr Sci 66(4):309–314

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

Nandy A (1996b) Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. CABIOS 12(1):55–62

Nandy A (2009) Empirical relationship between intra-purine and intra-pyrimidine differences in conserved gene sequences. PLoS One 4(8):e6829. https://doi.org/10.1371/journal.pone.0006829

Nandy A, Basak SC (2015) Prognosis of possible Reassortments in recent H5N2 epidemic influenza in USA: implications for computer-assisted surveillance as well as drug/vaccine design. Curr Comp-Aided Drug Des 11:110–116

Nandy A, Nandy P (2003) On the uniqueness of quantitative DNA difference descriptors in 2D graphical representation models. Chem Phys Lett 368:102–107

Nandy A, Harle M, Basak SC (2006) Mathematical descriptors of DNA sequences: development 1276 and applications. ARKIVOC 9:211–238

Nandy A, Basak SC, Gute BD (2007) Graphical representation and numerical characterization of H5N1 avian flu neuraminidase gene sequence. J Chem Inf Model 47(3):945–951

Nandy A, Ghosh A, Nandy P (2009) Numerical characterization of protein sequences and application to voltage-gated sodium channel a subunit phylogeny. In Silico Biol 9:77–87

Nandy A, Sarkar T, Basak SC, Nandy P, Das S (2014) Characteristics of influenza HA-NA interdependence determined through a graphical technique. Curr Comp-Aided Drug Des 10:285–302

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Patterson KD, Pyle GF (1991) The geography and mortality of the 1918 influenza pandemic. Bull Hist Med 65:4–21

Pedersen JC (2014) Hemagglutination-inhibition assay for influenza virus subtype identification and the detection and quantitation of serum antibodies to influenza virus. Methods Mol Biol 1161:11–25. https://doi.org/10.1007/978-1-4939-0758-8_2

Pielak RM, Chou JJ (2011) Influenza M2 proton channels. Biochim Biophys Acta 1808 (2):522–529

Pyle GF (1986) The diffusion of influenza: patterns and paradigms. Rowan & Littlefield, New Jersey

Qi X, Wu Q, Zhang Y, Fuller E, Zhang C-Q (2011) A novel model for DNA sequence similarity analysis based on graph theory. Evol Bioinforma 7:149–158. https://doi.org/10.4137/EBO.S7364

Randic M, Vracko M, Nandy A, Basak SC (2000a) On 3-D graphical representation of DNA primary sequences and their numerical characterization. J Chem Inf Comput Sci 40:1235–1244

Randic M, Vracko M, Nandy A, Basak SC (2000b) On 3-D representation of DNA primary sequences. J Chem Inf Comput Sci 40:1235–1244

Randic M, Vracko M, Lers N, Plavsic D (2003) Novel 2-D graphical representation of DNA sequences and their numerical characterization. Chem Phys Lett 368:1–6

Randic M, Zupan J, Balaban AT (2004) Unique graphical representation of protein sequences based on nucleotide triplet codons. Chem Phys Lett 397(1–3):247–252 https://doi.org/10.1016/j.cplett.2004.08.118

Randic M, Lers N, Plavsic D, Basak SC, Balaban AT (2005) Four-color map representation of DNA or RNA sequences and their numerical characterization. Chem Phys Lett 407:205–208

Randic M, Zupan J, Balaban AT, Drazen V-T, Plavsic D (2011) Graphical representation of proteins. Chem Rev 111(2):790–862

Raychaudhury C, Nandy A (1999) Indexing scheme and similarity measures for macromolecular sequences. J Chem Inf Comput Sci 39:243–247

Reid AH, Fanning TG, Hultin JV, Taubenberger JK (1999) Origin and evolution of the 1918 "Spanish" influenza virus hemagglutinin gene. Proc Natl Acad Sci 96(4):1651–1656. https://doi.org/10.1073/pnas.96.4.1651

Reid AH, Janczewski TA, Elliot AJ, Daniels RS, Berry CL, Oxford JS, Taubenberge JK (2003) 1918 influenza pandemic caused by highly conserved viruses with two receptor-binding variants. Emerg Infect Dis 10:1249–1253

Reid AH, Taubenberger JK, Fanning TG (2004) Evidence of an absence: the genetic origins of the 1918 pandemic influenza virus. Nat Rev Microbiol 2(11):909–914

Resnik DB (2013) H5N1 avian flu RESEARCH and the ethics of KNOWLEDGE. Hast Cent Rep 43(2):22–33. https://doi.org/10.1002/hast.143

Sarkar T, Das S, De A, Nandy P, Chattopadhyay S, Chawla-Sarkar M, Nandy A (2015) H7N9 influenza outbreak in China 2013: in silico analyses of conserved segments of the hemagglutinin as a basis for the selection of peptide vaccine targets. Comput Biol Chem 59:8–15

Sen D, Dasgupta S, Pal I, Manna S, Basak SC, Nandy A, Grunwald G (2016) Intercorrelation of major DNA/RNA sequence descriptors–a preliminary study. Curr Comput Aided Drug Des 12 (3):216–228. https://doi.org/10.2174/1573409912666160525111918

Severi G, Southey MC, English DR, Jung CH, Lonie A, McLean C et al (2014) Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer. Breast Cancer Res Treat 48(3):665–673. https://doi.org/10.1007/s10549-014-3209-y

Simonsen L, Clarke MJ, Schonberger LB, Arden NH, Cox NJ, Fukuda K (1998) Pandemic versus epidemic influenza mortality: a pattern of changing age distribution. J Infect Dis 178:53–60

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Smith TF, Waterman MS, Burks C (1985) The statistical distribution of nucleic acid similarities. Nucleic Acids Res 13:645–656

Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA (2004) Mapping the antigenic and genetic evolution of influenza virus. Science 305 (5682):371–376

Song J, Tang H (2005) A new 2-D graphical representation of DNA sequences and their numerical characterization. J Biochem Biophys Methods 63:228–239

Spackman E, Pantin-Jackwood MJ, Kapczynski DR, Swayne DE, Suarez DL (2016) H5N2 highly pathogenic avian influenza viruses from the US 2014–2015 outbreak have an unusually long pre-clinical period in turkeys. BMC Vet Res 12:260

Stiver G (2003) The treatment of influenza with antiviral drugs. CMAJ 168(1):49–57

Suarez DL, Perdue ML, Cox N, Rowe T, Bender C, Huang J, Swayne DE (1998) Comparisons of highly virulent H5N1 influenza A virus isolated from humans and chickens from Hong Kong. J Virol 72:6678–6688

Tang XC, Zhou PP, Qiu WY (2010) On the similarity/dissimilarity of DNA sequences based on 4D graphical representation. Chin Sci Bull 55:701–704. https://doi.org/10.1007/s11434-010-0045-2

Taubenberger JK, Morens DM (2006) 1918 influenza: the mother of all pandemics. Emerg Infect Dis 12(1):15–22. https://doi.org/10.3201/eid1201.050979

Tong S, Zhu X, Li Y, Shi M, Zhang J, Bourgeois M et al (2013) New world bats harbor diverse influenza A viruses. PLoS Pathog 9(10):e1003657. https://doi.org/10.1371/journal.ppat.1003657

Wagner R, Matrosovich M, Klenk HD (2002) Functional balance between haemagglutinin and neuraminidase in influenza virus infections. Rev Med Virol 12:159–166

Wang J, Zhang Y (2006) Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation. Chem Phys Lett 423:50–53

Wei H (2010) The interaction between the 2009 H1N1 influenza a hemagglutinin and neuraminidase: mutations, co-mutations and the NA stalk motif. J Biomed Sc Engg 3:1–12

WHO (2009) Influenza A(H1N1) virus resistance to oseltamivir - 2008/2009 influenza season, northern hemisphere. http://www.paho.org/hq/images/stories/ad/hsd/cd/influenza/h1n120081230.pdf

WHO (2017). Recommended composition of influenza virus vaccines for use in the 2017–2018 northern hemisphere influenza season. http://www.who.int/influenza/vaccines/virus/recommendations/2017_18_north/en/

WHO (2018a) Influenza (Seasonal) Fact sheet. who.int. Updated to January 2018. http://www.who.int/mediacentre/factsheets/fs211/en/. Retrieved 12 Mar 2018

WHO (2018b) Recommended composition of influenza virus vaccines for use in the 2018–2019 northern hemisphere influenza season. http://www.who.int/influenza/vaccines/virus/recommendations/2018_19_north/en/

Wiesner I, Wiesnerova D (2010) 2D random walk representation of Begonia x_tuberhybrida multiallelic loci used for germplasm identification. Biol Plant 54:353–356

World Health Organization (2007) Ethical considerations in developing a public health response to pandemic influenza. WHO Press. WHO/CDS/EPR/GIP/20072

Xu R, Zhu X, McBride R, Nycholat CM, Yu W, Paulson JC, Wilson IA (2012) Functional balance of the hemagglutinin and neuraminidase activities accompanies the emergence of the 2009 H1N1 influenza pandemic. J Virol 86:9221–9232

Yang S, Santillana M, Brownstein JS, Gray J, Richardson S, Kou SC (2017) Using electronic health records and internet search information for accurate influenza forecasting. BMC Infect Dis 17 (1):332. https://doi.org/10.1186/s12879-017-2424-7

Yao Y-h, Nan X-y, Wang T-m (2006) A new 2D graphical representation—classification curve and the analysis of similarity/dissimilarity of DNA sequences. J Mol Struct THEOCHEM 764:101–108

Yau SS-T, Wang J, Niknejad A, Lu C, Jin N, Ho Y-K (2003) DNA sequence representation without degeneracy. Nucleic Acids Res 31:3078–3080

Yin D, Chen Y, Yau S-T (2014) A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. J Theor Biol 359:18–28

Zhang Z, Liu L, Li J, Zhang Z (2010) Spectral representation of DNA sequences and its application. IEEE 1023–1027. http://search.ror.unisa.edu.au/media/researcharchive/open/9915909466801831/53108451280001831. Accessed 22 Feb 2018

Zhao Y-B, Qi Z-H, Yang A-P (2015) Characterization and similarity analysis of DNA sequences considering codon degeneracy. Int J Hybrid Inf Technol 8(1):73–84