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

GRAPHICAL REPRESENTATION AND MATHEMATICAL CHARACTERIZATION OF PROTEIN SEQUENCES AND APPLICATIONS TO VIRAL PROTEINS

By AMBARNIL GHOSH* AND ASHESH NANDY†

*Physics Department, Jadavpur University, Jadavpur, Kolkata, India
†Centre for Interdisciplinary Research and Education, Jodhpur Park, Kolkata, India

Abstract

Graphical representation and numerical characterization (GRANCH) of nucleotide and protein sequences is a new field that is showing a lot of promise in analysis of such sequences. While formulation and applications of GRANCH techniques for DNA/RNA sequences started just over a decade ago, analyses of protein sequences by these techniques are of more recent origin. The emphasis is still on developing the underlying technique, but significant results have been achieved in using these methods for protein phylogeny, mass spectral data of proteins and protein serum profiles in parasites, toxicoproteomics, determination of different indices for use in QSAR studies, among others. We briefly mention these in this chapter, with some details on protein phylogeny and viral diseases. In particular, we cover a systematic method developed in GRANCH to determine conserved surface exposed peptide segments in selected viral proteins that can be used for drug and vaccine targeting. The new

ADVANCES IN PROTEIN CHEMISTRY AND STRUCTURAL BIOLOGY, Vol. 83
DOI: 10.1016/B978-0-12-381262-9.00001-X

GRANCH techniques and applications for DNAs and proteins are covered briefly to provide an overview to this nascent field.

I. Introduction

A. *Protein Basics*

Proteins, the most versatile macromolecules in the living system, primarily constitute complex folded chain of amino acids which are encoded by genes. The information content of the folded complex constitutes a functional unit that plays a crucial role in biological processes. The origin of the word ''Protein'' is from the Greek ''prota'' which means ''of primary importance.'' This name was coined by Jöns Jakob Berzelius in 1838 for large organic compounds with a very close similarity in their empirical formulae and of primary importance in animal nutrition, though the evidences were not so prominent at that time. A landmark in protein chemistry came through Frederick Sanger and his colleagues, at the University of Cambridge in 1954 when, after 10 years of hard work, they succeeded in solving the complete primary structure of insulin (Sanger and Tuppy, 1951; Sanger, 1952; Sanger and Thompson, 1953). The very next milestone in protein chemistry was Max Perutz (Perutz and Weisz, 1947; Perutz, 1960; Perutz et al., 1960) and Sir John Cowdery Kendrew (Kendrew et al., 1958, 1960; Kendrew, 1959) solving the 3D structure of hemoglobin and myoglobin. These findings are the basis of modern age of advanced structural protein chemistry research.

Proteins form the building blocks of the structure and function of biological entities. A typical mammalian cell contains as many as 10,000 different proteins having a diverse array of functions (Karp, 2008). The set of proteins expressed in a cell or cell type is called a proteome. Proteins are generally a few hundred amino acids in chain length but can vary in size from a few tens of amino acids to over 34,000 amino acids, for example, the human titins, also known as the largest in protein world (MW = 3–3.7 MDa; Opitz et al., 2003). While a single protein chain can theoretically fold in an unlimited number of ways (Chou and Fasman, 1974b; Fasman, 1989; Feldman and Hogue, 2002), typically a specific amino acid chain folds to a particular structure through a process that is not yet clearly understood (Dill et al., 2007, 2008; Ghosh et al., 2007), but which is the basis for all protein interactions; recent research shows that the folded structure might have conformational changes depending on

the environment too (Makowski et al., 2008). Protein structure is often referred to in terms of four aspects: The primary structure consisting of the amino acid chain, the secondary structure which contains regularly repeating structures like alpha helices and beta sheets stabilized by hydrogen bonds, the tertiary structure which is the final folded structure incorporating the various secondary structures, and a quaternary structure where several proteins are bound together to form one protein complex such as are found in the neuraminidase body of an influenza virion (Russell et al., 2006) or the VP7 of a rotavirus particle (Li et al., 2009b). The tertiary and quaternary structures of a large number of proteins have become available through X-ray crystallography and NMR spectroscopy studies and the data are available in Protein Data Bases (PDB) such as World Wide Protein Data Bank (WWPDB; Berman et al., 2007), RCSB Protein Data Bank (RCSB-PDB; Deshpande et al., 2005; Dutta et al., 2007), Protein Data Bank Europe (PDBe; Velankar et al., 2010), Protein Data Bank Japan (PDBj; Nakamura et al., 2002; Kinjo et al., 2010), and Biological Magnetic Resonance Databank (BMRB; Markley et al., 2008). The difficulty of crystallizing proteins has restricted the number of proteins whose structures are sufficiently well known (Chayen, 2004, 2009; Chayen and Saridakis, 2008). However, taking the protein primary structure as the source material for all subsequent structures, structural genomics and protein structure prediction methods theoretically predict protein secondary and tertiary structures based on known structures (Baker and Sali, 2001).

The importance of proteins in biological function have led to wide ranging studies to understand how proteins fold (Dobson, 2004; Dill et al., 2007; Ghosh et al., 2007), interact with other proteins to regulate enzyme activity (Frieden, 1971), oligomerize to form fibrils (Powers and Powers, 2008), aggregate to protein complexes that lead to conformational changes, and enable signaling networks. These interactions are mediated by the chief characteristic of a protein: the ability to bind other molecules specifically and tightly to it. The specificity arises from unique shapes in the tertiary structure of the protein surface (Roach et al., 2005, 2006) where, for example, a depression acts as a binding site or pocket and by the chemical natures of the side chains of the neighboring amino acids. This also results in total inability to bind in cases where changes in the amino acid composition render conformational changes to the binding site (Moscona, 2005). Such changes arising out of mutations in the amino acid chains are among the main factors responsible for development of drug resistance in bacterial and viral diseases (Moscona, 2004). Enzymatic

role of proteins helps catalyze metabolic reactions but only a small region of the protein consisting of a few amino acids are active in the catalysis; a noncatalytic example of protein includes the antibodies that are part of the adaptive immune systems and act as a binder to antigens for destruction (MacCallum et al., 1996). Ligand-binding proteins such as hemoglobin bind specific small molecules to transport them to other locations in the body of a multicellular organism (Baldwin and Chothia, 1979). Structural proteins such as actin and tubulin confer stiffness and rigidity to the cytoskeleton (Doherty and McMahon, 2008); other structural proteins such as myosin and kinesin generate mechanical forces and are responsible for the motility of many single cell organisms (Rayment, 1996).

Thus, there are numerous processes, and there are numerous proteins that take part in them. These processes and the functions of the proteins are studied through *in vivo* and *in vitro* analysis. *In vitro* analysis helps understand how a protein functions, *in vivo* analysis often helps in understanding its functional location and related parameters in the living system; however, the specifics of how a protein targets particular organelles or cellular structures are often unclear (Bejarano and Gonzalez, 1999). Site-directed mutagenesis techniques (Ruvkun and Ausubel, 1981) that alter the protein sequence and hence its structure and cellular location/function that help to identify susceptibility to regulation provide guidelines to rational drug design or development of new proteins with novel properties.

Among the simplest of biological entities, and of particular interest for this chapter, is the virus. A virus particle like the influenza or rotavirus contains about 8–11 protein-coding genes in a multiprotein coat that protects the RNA or DNA of the virus and also enables the proteins and genetic materials to enter and leave cells. A great range of variability in amino acid composition is observed for these viral proteins (Reid et al., 2000; Ghosh et al., 2009), specifically the surface situated ones like NA (neuraminidase; Ghosh et al., 2010), HA (hemagglutinin), VP4 (variable protein), VP7 (Gunn et al., 1985), and gp120 (of the HIV) but the functional impact remains the same. Often, a single change in the side chain of a single amino acid is enough for producing a new mutant (Lopez et al., 2005). Viruses use this highly mutable property for escaping the host defense mechanism and they are also frequently found to generate escape mutants against a naturally occurring immunity or artificially designed drug or vaccine (Air and Laver, 1989).

B. Drugs and Proteins

Proteins are involved by function or malfunction, in diseases of organism. Bacterial, viral, and other pathogens disrupt the normal protein functions and thereby destabilize the infected host organism (Goldsby et al., 2000). While immunological defenses are called into action by the infection, often these are inadequate by themselves and have to be supported by drugs, vaccines, and other therapeutical regimes. Design of drugs and study of their actions have therefore been an important area of research. Drugs can act through formation of drug–DNA complexes (Chaires, 1997, 1998) or protein–drug complexes (Chicault et al., 1981). Major trends of research into drug–DNA relationships have been recently reviewed (Nandy and Basak, 2010). Stated simply, DNA drugs and vaccines are made of plasmids designed to carry a selected gene into cells where it is translated into a protein. In the case of antiviral DNA vaccine, for example, plasmids are created for producing the selected viral protein in the cell and immune systems are expected to act to prevent future infections from the virus (Ulmer et al., 1996a,b; Gurunathan et al., 2000). Advanced techniques such as codon optimization (Deml et al., 2001) are enhancing the protein production from the plasmids and others such as adjuvant incorporation are enhancing the immune response leading to more effective vaccines and therapies, several of which are already available for treatment of specified animals afflicted with the West Nile virus (Kramer et al., 2007), melanoma and fetal loss, while applications for humans for treating HIV, influenza, hepatitis C, and other diseases are under trial (Morrow and Weiner, 2010).

Pharmaceutical proteins effective against a wide range of bacterial infections can be traced to penicillin, and developed into new class of drugs referred to as antibiotics. Conventional production processes for antibiotics are expensive and face many regulatory issues. Vaccines that enhance the body's immune system consist of attenuated viruses but can, in rare cases, harm the host with a full-blown viral occupancy (Ball et al., 1998; Colgrove and Bayer, 2005). Since viruses use the host's cells to replicate, designing safe and effective antiviral drugs is difficult and also makes it difficult to find targets for the drugs that would interfere with the virus without also harming the host organism's cells. But almost all antimicrobials, including antivirals, are subject to drug resistance as the pathogens mutate over time (Gold and Moellering, 1996), becoming

less susceptible to the treatment. Small molecules are often used as drugs, but the new technology of recombinant proteins (Geigert, 1989; Dingermann, 2008), commonly produced using bacteria or yeast in a bioreactor, potentially provide greater efficacy and fewer side effects because their action can be more precisely targeted toward the cause of a disease rather than treatment of symptoms, is yet to gain wide acceptance.

Peptide-based drugs operate by stimulating the immune response to the peptide and thereby to the invading pathogen. Peptides play an important role in modulating many physiological processes in our body. Use of peptides as drugs have the benefit that they are small, easily optimized, and can be quickly investigated for therapeutic potential. However, peptide drug screening process (Otvos, 2008), although a well-established approach, is long and arduous resulting in high manufacturing costs, and the fact that they have short half-life, and limited *in vivo* bioavailability hampers their effectiveness; new approaches have been proposed to overcome the difficulty of generating sufficient amount of the required tRNAs (Owens, 2004). The peptides can be naturally derived or chemically synthesized, with the latter method being more prevalent. Novel peptide analogs (Lee et al., 2002) are also being synthesized to create more potent drugs.

In practice, protein and peptide drugs are finding increasing acceptance in therapeutics. A drug's efficiency is related to the degree of its binding with the proteins in blood serum (Meyer and Guttman, 1968; Koch-Weser and Sellers, 1976): The less bound a drug is, the more efficiently it can diffuse through cell membrane. Common drug-binding proteins in plasma are human serum albumin, lipoprotein, glycoprotein, etc. It is the unbound fraction of the drug–protein complex that exhibits therapeutic effect and excessive binding may mitigate against rapid action of the drug. However, the same effect can be used for long-lasting dosage by designing drugs that bind to the protein and act as a reservoir so that the unbound fraction is released slowly.

But degradation of the proteins during storage and drug administration routes remains a challenging problem (Frokjaer and Otzen, 2005). These issues of stability of therapeutic proteins toward aggregation and misfolding in long-term storage as well as means of efficacious delivery that avoid adverse immunogenic side effects are engaging the attention of the pharmaceutical industry (Frokjaer and Otzen, 2005). While invasive routes

such as subcutaneous injections are often used, oral delivery faces difficulties in poor permeability across biological membranes due to the hydrophobic nature and large molecular size, susceptibility to enzymatic attack, among others. Formulation strategies for protein therapeutics thus continue to remain a challenging problem.

C. Bioinformatics in Protein Studies

The complexities of protein function and structure have necessitated the development of computational techniques to analyze available data and help in formulating novel ways to predict structure, function, and interaction of proteins. Especially, in view of the requirements of new approaches to drug development through recombinant proteins, synthesizing new peptides, and investigating drugs–DNA complexes, use of computational methods is now of vital importance.

The increased availability and accessibility of genomic and protein sequence data have opened up new possibilities for the search for target proteins, and the success of protein and peptide therapeutics is revolutionizing the biotech and pharmaceutical market, spurring the creation of next-generation products with reduced immunogenicity (Schellekens, 2002; Tangri et al., 2005), improved safety, and greater effectiveness. The protein engineering market is expected to cross $100 billion in sales in 2010 from about $36 billion 4 years ago. The top-selling therapeutic protein is reported to be Amgen's Aranesp (Locatelli and Vecchio, 2001), a reengineered variant of the company's first-generation product Epogen (recombinant human erythropoietin). A number of such products have been launched by Genetech and others, and nonparenteral delivery systems, alongside parenteral protein and peptide drug delivery systems have also been approved (Packhaeuser et al., 2004). Progress in bioinformatics and computational biology as well as new techniques in protein engineering (recombinant proteins through site-directed mutagenesis and posttranslational modifications) are aiding the development of reengineered, improved, whole antibody, and antibody fragment-based products, reducing immunogenicity by using fully human recombinant antibodies or human antibodies derived from transgenic mice and allowing biosimilar products to be differentiated on the basis of superior characteristics. Screening experiments for appropriate molecules rely critically on bioinformatics support for design of experiments and for

interpreting the generated data, for example, to identify interesting differentially expressed genes and to predict the function and structure of putative target proteins (Lengauer and Zimmer, 2000).

Protein characterization and *in silico* protein design and structure analyses form an integral part of these developments. Phylogenetic analyses based on primary sequences have been used to group related proteins and understand their evolutionary history, algorithms have been developed to predict protein secondary structures, and web accessible systems are available to suggest possible folding patterns (Shen and Chou, 2009). A number of epitope prediction tools have been devised with varying degrees of success to aid in drug design (Yang and Yu, 2009); one area of nascent research is concerned with understanding of allosteric conformations that may help or hinder protein interactions (Teague, 2003). In a broader area, computational biology has already proved itself as one of the powerful tools for handling the large genomic databases. The basic applications involve killer tools like sequence alignment, phylogenetic tree drawing, sequence comparison, etc. *In silico* motif search algorithms on primary protein structure can be applied for finding structural information like signal sequence prediction (Menne et al., 2000), cleavage site prediction (Chou, 2001), glycosylation prediction (Blom et al., 2004), posttranslational modifications prediction, etc. Large datasets are frequently found to be utilized in predictions of protein structural levels from primary structure. Software like Modeller (Eswar et al., 2007, 2008), Discovery Studio, etc., can predict 3D structure of proteins from a database of known crystallized proteins. Many theories have been developed in this prediction research but they are often ineffective in case of a completely new protein for comparison with the preexisting database (comparative protein modeling) or a protein without appropriate template. Another very important application of data mining is the use of computational power in handling the proteomics data. In proteomics, proteins are detected by matching a part of it with the whole existing protein database in mass spectrophotometer software (Perkins et al., 1999). The basis of all these data mining and related computational techniques is mathematics and statistics. Different theories like dot-matrix algorithm (Gibbs and McIntyre, 1970), Needleman Wunsch algorithm (Needleman and Wunsch, 1970), Smith Waterman algorithm (Smith and Waterman, 1981; Smith et al., 1985), Hidden Marakov model (Eddy, 1996), Chou-Fasman algorithm (Chou and Fasman, 1974a,b), etc., are widely used.

Some models or algorithms work on interpretation from statistics and probability and others depend on the visual interpretation of genomic data by different techniques (Nandy, 1994; Randic, 2004, 2006; Randic et al., 2005, 2006; Nandy et al., 2007; Basak and Gute, 2008; Gonzalez-Diaz et al., 2008a, 2009).

To aid in protein characterization, ideas of graphical representation and numerical characterization (GRANCH) have been taken up from their success in DNA sequence analysis, but complicated here by the fact that protein sequences are composed of 20 amino acids whereas a DNA sequence is concerned with only the four building blocks of nucleotides. However, while some standard procedures such as dot-matrix plot have been used for a long time, several ingenious schemes have been developed recently that have marked significant success in this nascent field as we show in the next few sections. Coronavirus phylogeny, studies of H5N1 neuraminidase protein mutations and identification of highly conserved peptide stretches on influenza virus and rotavirus proteins that could potentially aid in the development of new drugs and vaccines are some of the significant results of application of these novel techniques. We provide a brief review of these studies in Section III.

II. Graphical Methods

Graphical methods to display sequences have the advantage of visual indications of trends and inherent features. The familiar dot-matrix type of graphs have been widely used to determine systematics in nucleotide and amino acid sequences. The dots plotted on a 2D grid with the sequence running along the positive x- and negative y-axes produce a pattern (Gibbs and McIntyre, 1970) that is useful in determining sequence similarity, direct repeats, inverted repeats, etc., and such plots have also been used in RNA secondary structure predictions theories, for example, complementary sequences in a RNA structure in the dot-matrix analysis of nucleotide sequences of potato tuber spindle viroid (Fig. 1). In the case of proteins, one of the more widely used molecular graphs is the hydrophobicity–polarity lattice graph to model structure–activity relationships and folding dynamics in 2D/3D spaces (Jiang and Zhu, 2005; Chikenji et al., 2006). In continuing developments in the field, a new pseudofolding molecular graph or network-type representation has been proposed recently (Fernandez et al., 2008).

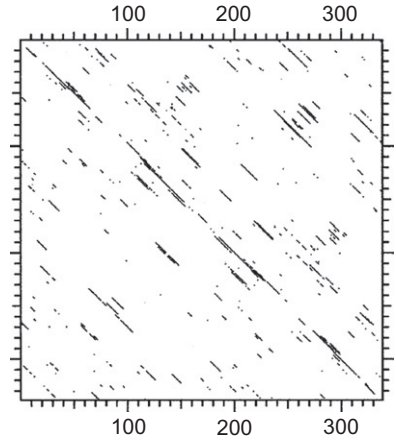


FIG. 1. Dot-matrix graph for RNA secondary structure prediction for Potato Tuber Spindle Viroid RNA sequence. Reproduced with permission from Cold Spring Harbor Laboratory Press, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA. Web: http://www.bioinformaticsonline.org/. Source: Mount, 2004.

A. Graphical Methods for DNA Sequences

1. Graphical Representation of DNA Sequences

Much of the recent interest in graphical methods arose from their applications in analysis of DNA and RNA sequences. Representation of the sequence of bases in a DNA or RNA strand using graphical methods was initiated several years ago with a 3D model proposed by Hamori and Ruskin (1983), followed up subsequently by Gates (1986), Nandy (1994), and Leong and Morgenthaler (1995) with 2D representations, while Peng et al. (1992) and Jeffrey (1990) represented sequence data graphically in more abstract forms. The plot of purine–pyrimidines against base numbers devised by Peng et al. (1992) demonstrated the presence of long-range correlations in DNA sequences while Jeffrey's Chaos Game Representation (CGR) method showed visually for the first time the fractal nature embedded in these sequences (Fig. 2), as also the different patterns for mammalian, bacterial, and phage sequences reflecting the inherent differences in their base organization. The utility of the graphical approach have led to many new techniques of GRANCH of DNA and RNA sequences (see review Nandy et al., 2006).

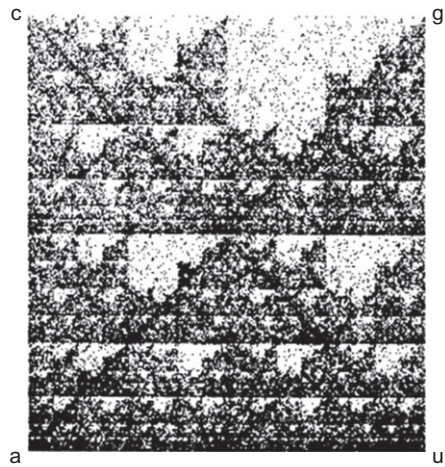
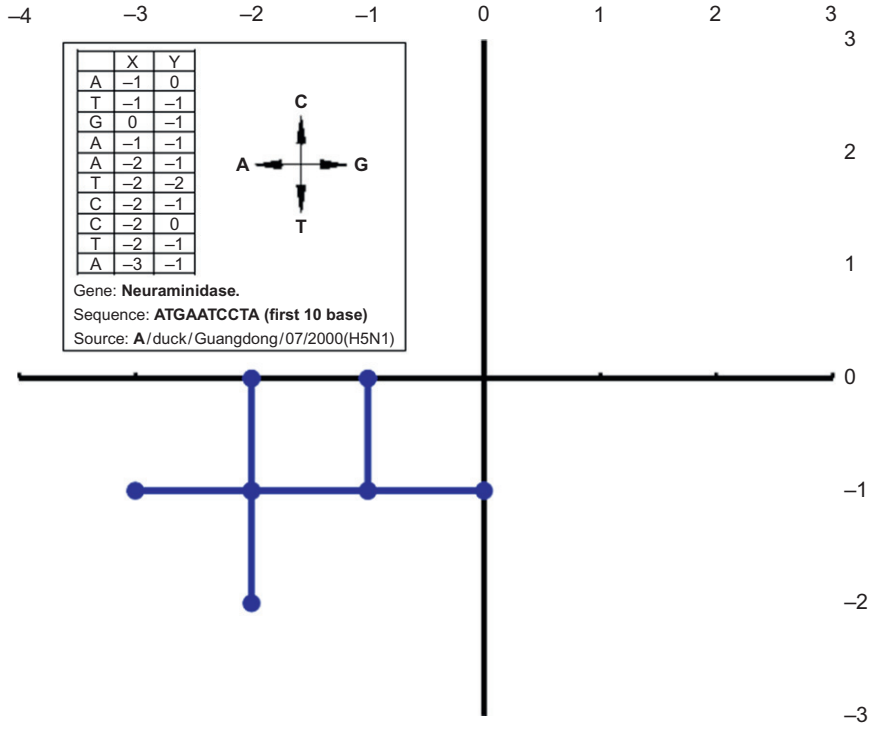


FIG. 2. CGR of human beta-globin (HUMHBB) region of Chromosome 11. Reproduced with permission from Oxford University Press, UK and Copyright Clearance Center (CCC) Web: http://nar.oxfordjournals.org/. Source: Jeffrey, 1990.

The basic approach can be most simply described in the 2D representation where the four cardinal directions are associated with the four bases. Nandy (1994) associated adenine with the negative *x*-axis, cytosine with the $+y$-axis, guanine with $+x$-axis, and the thymine with the $-y$-axis and plotted a sequence starting from the origin and moving, for each base in the sequence, one step at a time in the designated direction depending on the specific base until the entire sequence is plotted. Figure 3 follows the above mentioned direction of graphical representation technique for first 10 nucleotides of neuraminidase RNA (c-DNA) and generates a series of points like a Markov chain that reflects the sequence and distribution of bases in the sequence in the chosen representation. However, this simple approach has the disadvantage of allowing reentry in the random walk path, for example, a sequence like AGAGAG traces only one unit path in the Nandy representation, and several other schemes have been formulated that minimize or eliminate this problem, but with reduced visual appeal (Nandy et al., 2006). Randic and his coworkers, for example, proposed various representations such as ''worm'' curve (Randic et al., 2003c; Randic, 2004), ''four horizontal line'' curve (Randic et al., 2003a,b), four-color maps (Randic et al., 2005), ''spectrum-like'' figures



	X	Y
A	–1	0
T	–1	–1
G	0	–1
A	–1	–1
A	–2	–1
T	–2	–2
C	–2	–1
C	–2	0
T	–2	–1
A	–3	–1

FIG. 3. Graphical representation (according to Nandy, 1994) of first 10 nucleotides of H5N1 neuraminidase RNA (or c-DNA).

(Randic, 2006) among others to reduce or eliminate the degeneracy inherent in the 2D approach. Yau et al. (2003) proposed a 2D graphical representation, where the purines (A, G) and pyrimidines (T, C) are plotted on two quadrants of the Cartesian coordinate system at fixed angles to the *x*-axis; such a system has no degenracy. A sequence is plotted as a progression of points counted along the *x*-axis but rising or falling with the nature of the base thus tracing a pattern that is unique for the particular sequence. Among other proposals, mention may be made of the recent works of Todeschini et al. (2006, 2008) who use partial ordering ideas to compare the first exons of eight beta-globin sequences, and Liu and Wang (2010) who used an 8D representation of DNA sequences for comparison of similarities/dissimilarities of over 40 viral, lipase, phage,

and other genes, both of which methods dispense with visual rendering in favor of more rigorous mathematical approaches.

2. *Numerical Characterization of DNA Sequences*

To obtain a quantitative measure of the graphical representations, different techniques have been devised to convert these representations into numbers or vectors that are expected to be characteristic of each sequence. A simple geometrical technique for the 2D graph of a DNA sequence determines the weighted center of mass of a plot (μ_x, μ_y) and a graph radius (g_R), and therefrom the distance (Δg_R) of two sequences, using Euclidean measures (Raychaudhury and Nandy, 1999):

$$\mu_x = \frac{\sum x_i}{N}, \quad \mu_y = \frac{\sum y_i}{N} \quad \text{and} \quad g_R = \sqrt{\mu_x^2 + \mu_y^2} \tag{1}$$

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

where the (x_i, y_i) represent the coordinates of each point on the plot, N is the total number of the bases in the segment and the μ_1 and μ_2 refer to two different DNA sequences. The g_R here represents a base distribution index that is critically dependent upon the position of each base in the sequence and together with the μ_x, μ_y form a set of biodescriptors for the sequence. The g_R and the Δg_R have been found to be very sensitive measures of the sequence composition and distribution (Nandy and Nandy, 2003). The difference index, Δg_R, provides a quantitative comparison between the sequences: the smaller the Δg_R, the more similar are the underlying DNA sequences and the higher the Δg_R, the more dissimilar are the sequences.

A matrix method of determining numerical indexes for DNA sequences was proposed by Randic et al. (2000) in a 3D graphical representation in which the position of every base of a sequence was related to all other bases through a Euclidean and graph–theoretic distance. The ratios of these distances, D_E/D_G, formed the elements of a DD matrix. Since matrices are well-known objects with well-defined properties, the leading eigenvalues of a DD matrix are considered to be characteristic, or invariants, of the matrix and, by association, to be descriptors of the DNA sequence itself. The authors calculated the leading eigenvalues of the first

exon sequence of the *beta-globin* gene of eight species and determined the similarity/dissimilarity between the various sequences. This was followed by successive proposals for different graphical representations that similarly used matrix methods to determine invariants to characterize each sequence and form vectors of such invariants to estimate the degrees of similarities and dissimilarities between members of a family of DNA sequences (Nandy et al., 2006). The works of Randic, Todischini, and Wang referred to earlier use these techniques of DD matrices or Hasse matrices to compute the distances between species from their gene sequences.

3. Applications

The new GRANCH techniques gave a rich view of the complexities of DNA sequences. Among the first applications of these techniques to human diseases, Liao et al. (2006) showed that mathematical techniques can be used to analyze the underlying DNA/RNA sequences by studying the severe acute respiratory syndrome (SARS) coronavirus, and, separately, that GRANCH techniques could do away with multiple alignment requirements to study gene families. Larionov et al. (2008) broadened the usage by showing that plots of human and mouse chromosomal sequences in a graphical representation were able to reveal long-range palindromes. The 3D and 2D graphical representations visually highlighted the base preferences along a DNA sequence (Hamori and Ruskin, 1983; Nandy, 1994), while 2D representations showed long runs of duplications of a motif as simple runs on the graphs (Nandy and Nandy, 1995). Gates had remarked on large-scale complex repeats that were revealed by 2D graphs (Gates, 1986); Nandy showed that conserved genes have shapes on the 2D maps that are similar across species (Nandy, 1994), a visual rendition no doubt of homology. Viewing a number of maps of the H5N1 neuraminidase gene revealed a conserved region (Nandy et al., 2007), and numerical characterization of the maps, in the whole RNA sequence and in segments, has allowed reconstruction of the wide dissemination and possible recombination of segments of the gene not reported heretofore (Ghosh et al., 2009). In a novel application, using a variation of the 2D graphical representation, Wiesner and Wiesnerova (2010) studied multiallelic marker loci from *Begonia* × *tuberhybrida*. They found significant correlation of graph invariants to genetic diversity of the

marker loci and suggested that DNA walk representation may predict allele-rich loci solely from their primary sequences, which improves current design of new DNA germplasm identificators. Recently, Nandy has shown from inspection of conserved gene representations on 2D maps (Nandy, 2009) that effects of point mutations in gene sequences over evolutionary time scales indicate a polynomial relationship between the intrapurine intrapyrimidine differences on each strand of a DNA sequence.

B. Graphical Methods for Protein Sequences

1. Graphical Representation of Protein Sequences

The experience with GRANCH techniques for DNA and RNA sequences led to many proposals for GRANCH methods for protein sequences, although complicated by the necessity of accommodating 20 residues for proteins compared with four bases for the nucleotide sequences. One of the earliest attempts is the dot-matrix plot for protein sequences, but other techniques were also developed. Among the pioneer works for representing the chemical information in a protein graphically is the representation of protein bonds through the Ramachandran plot (Ramachandran et al., 1963; Ramachandran and Sasisekharan, 1968) and, for protein primary sequences, the Hydropathy plot (Engelman et al., 1986). The former can extract the secondary structural information from protein's bond angles, while the latter draws the graph from thermodynamical and chemical properties of amino acids. The DNA graphical representation methodology led Randic to propose a Magic Circle representation (Randic et al., 2006), where the total protein sequence is represented in a unit circle and the graph starts from the center following the sequence by moving half way toward the corresponding amino acids which are positioned equally spaced on the circumference. The result of the complete execution of the protein sequence within the circle produces a typical graph for a particular protein (Fig. 4), except for large protein sequences which are often found to have lesser visual benefits. Li et al. used a reduction model of abstracting the protein sequence in a five-letter code (Wang and Wang, 1999; Li et al., 2008) each representing a specific group of amino acids and generated a 2D-graph by plotting the reduced sequence on the *x*-axis and all five group representatives horizontally at equal intervals along the *y*-axis resulting in a zig-zag like graphical

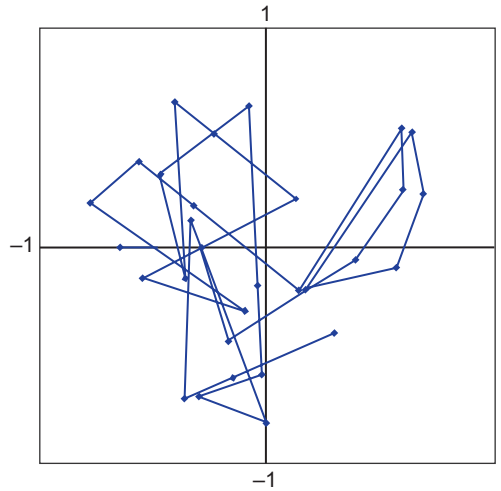


FIG. 4. Magic Circle representation of protein sequence WFFESRNDPANDPIILWLNGGPGCSSFTGL. Reproduced with permission from Elsevier provided by Copyright Clearance Center (CCC). Web: http://www.sciencedirect.com/science/journal/00092614. Source: Randic et al., 2006.

representation of the sequence (Fig. 5). 2D graphical representations based on nucleotide triplet codons (Bai and Wang, 2006) have been proposed for sequence comparison and start–stop sign of a coding region. Liao et al. (2006) used this approach to study 24 coronavirus genomic sequences which have ~29,000 bases each. They classified the 20 amino acids of a protein sequence into four separate groups according to the chemistry of their R groups: amino acids A, V, F, P, M, I, L belong to the hydrophobic chemical group; amino acids D, E, K, R belong to charged chemical group; amino acids S, T, Y, H, C, N, Q, W belong to polar chemical group; the unique G amino acid belongs to glycine chemical group. Starting with the nucleotide sequence, this enabled them to construct three 2D graphs (one for each reading frame) for each gene sequence and compute a distance matrix between the 24 coronaviruses from which they could generate a phylogenetic tree relating all the sequences without the need for any multiple alignments. Gonzalez-Diaz and his coworkers have used 2D lattice graphs for proteins (Aguero-Chapin et al., 2006; Gonzalez-Diaz et al., 2008a), constructed in a similar way to the DNA representations of Nandy and adapted to proteins

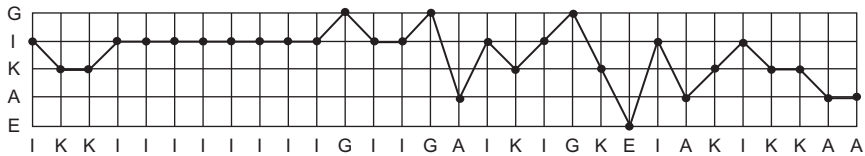


FIG. 5. Zig-zag curve generated from the 2D graphical representation of five-letter coded amino acids IKKIIIIIIIIGIIGAIKIGKEIAKIKKAA. Reproduced with permission from BMB reports, Web: http://www.bmbreports.org. Source: Li et al., 2008.

according to a proposed protocol (Estrada, 2002), and extended to other graph representations such as spiral and star networks (Aguero-Chapin et al., 2008a,b; Dea-Ayuela et al., 2008; Vilar et al., 2008; Munteanu et al., 2009); for example, for a star network, starting from the beginning of the sequence, the amino acids are placed in the corresponding branches transforming the protein sequence into modified branch connectivity graph from which connectivity indices (CIs) can be derived. Other authors have proposed higher dimensional representations such as the 3D model of Bai and Wang (2006) who embedded a dodecahedron in 3D space where each corner represented one of the 20 amino acids and thus generated a walk for the protein sequence, and the 20D representation of Novic and Randic (2008). The present authors have proposed an alternate 20D representation in Euclidean space (Nandy et al., 2009), where each amino acid is assigned to one axis in the 20D space and the sequence plotted using algorithms similar to the random walk model for 2D graphical representation of DNA sequences. This procedure generates a graph in the abstract 20D space from which consequences can be calculated to characterize proteins and quantify similarities and dissimilarities.

2. *Numerical Characterization of Protein Sequences*

Analogously to the GRANCH techniques of DNA sequences, to obtain quantitative measures for protein sequences, Randic, Li, Humberto Gonzales-Diaz, and several other authors have extended the methodologies of numerical descriptors for DNA sequences and of topological indices (TIs) used in QSAR studies to analysis of protein sequences, viral surfaces, and RNA secondary structures (Estrada and Uriarte, 2001; Randic et al., 2004, 2008; Bai et al., 2005; Gonzalez-Diaz et al., 2007b; Li et al., 2009a)

leading to more general biological applications. González-Díaz and collaborators have done extensive work on extension of these representations to the study of protein sequences (Aguero-Chapin et al., 2009) and applied to mass spectral data of proteins and protein serum profiles in parasites (Gonzalez-Diaz et al., 2008b), toxicoproteomics, and diagnosis of cancer patients (Cruz-Monteagudo et al., 2008; Gonzalez-Diaz et al., 2008a). Their group has used mathematical biodescriptors derived from toxicoproteomics maps in conjunction with chemodescriptors of toxic molecules to predict their toxicity (Hawkins et al., 2006). Integrated QSARs (Basak et al., 2006) developed using chemodescriptors for ligands and biodescriptors of a molecular entity, for example, connect structural information of drug molecules, DNA and RNA sequences, or RNA secondary and protein tertiary structures and may be used to predict parameters for new entities (Gonzalez-Diaz et al., 2008a). It has been found that using different type of numerical indices derived from the protein 2D molecular graphics to perform QSAR studies is simpler than having to work with the protein 3D structures (Gonzalez-Diaz et al., 2009; Vilar and González-Díaz, 2010). These indices describe graph/network topology, connectivity, or branching, often referred to as the graph TIs or network CIs used to determine structure–function relationships in cellular biochemistry (Chou and Cai, 2003), and have been applied in theoretical biology and bioinformatics of small-size molecules, macromolecules, proteome mass spectra, and protein interaction networks (Aguero-Chapin et al., 2006; Gonzalez-Diaz et al., 2007a, 2008a). Basak et al. (2011) have in a pathbreaking work using a new differential QSAR approach for study of dihydrofolate reductases (DHFR) from multiple strains of *Plasmodium falciparum* shown that DHFR from the wild strain is substantially different from four mutant strains of their study and remark that the protocols indicated in the paper can be used for the development of drugs to combat drug-resistant pathogens arising continuously in nature due to mutations. Bai and Wang (2005) proposed to numerically characterize protein sequences through the nucleotide triplet codons by using a 2D graphical representation system similar to that of Yau et al. (2003) to generate protein descriptors for the *Homo sapiens* X-linked nuclear protein (ATRX). An intuitively simpler indexation scheme based on the 20D graphical representation of protein sequences proposed by the present authors (Nandy et al., 2009) and described below has been found useful in generating phylogenetic relationships between sequences without necessity of multiple alignments and for

determining conserved surface exposed stretches on viral proteins that could be useful in drug and vaccine designs (Ghosh et al., 2010).

3. *Protein Similarities/Dissimilarities and Phylogeny by Graphical Methods*

Numerical characterization of sequences have also been targeted at the challenging problem of determining evolutionary relationships in protein families; for example, the multiplicity of voltage-gated sodium channel proteins from one for the bacteria (e.g., *Bacillus halodurans*) to 10 in humans, the development of the globin genes, the growth of differences in the highly conserved histones. Popular software like PHYLIP (Retief, 2000), MEGA (Tamura et al., 2007; Kumar et al., 2008), etc., are available for phylogenetic analysis, based generally on complex multiple sequence alignment (MSA) algorithms. Graphical methods like k-tuple, dot-matrix method (Gibbs and McIntyre, 1970), etc., are found as an integral part of MSA algorithm, and other graphical methods assess the extent of similarity/dissimilarity between protein sequence and serve as inputs to the software packages to generate the phylogenetic trees.

Bai and Wang (2006) derived the phylogenetic relationships for selected proteins using their 3D graphical representation where the amino acids are plotted on the corners of a dodecahedron. From the curve of the protein sequence obtained as a walk within this 3D space (Fig. 6), they derive a quotient matrix similar to the DD matrix discussed earlier for the DNA plots (Li and Wang, 2005), from which they can calculate the distance matrix between a set of protein sequences. The application of a similar procedure to a set of nine nerve genes from various organisms led to the generation of phylogenetic trees (Fig. 7). While usual methods of generating such trees are difficult due to the varying lengths of the sequences, the matrix method with leading eigenvalues do not have such problems and generates fairly acceptable relationships, although some of the details show, as the authors point out, that the method requires further refinement. The method is also useful in that it allows visible inspection of protein sequence characteristics and thus is good for comparative study of proteins too.

Li et al. (2009a) proposed a 3D graphical representation of protein sequences where the amino acids were classified into five separate groups based on their interactions. Thus, in terms of the one-letter code of amino acids, Group 1 consisted of the amino acids C, M, F, I, L, V, W, Y; Group 2

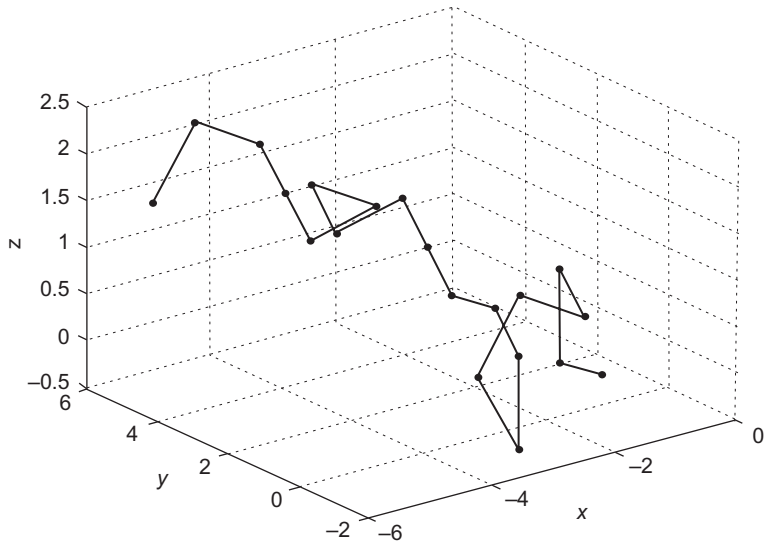


FIG. 6. 3D curve of protein sequence MGAPFVWALGLLMLQMLLFV, proposed by Bai and Wang, 2006. Reproduced (or published after acceptance) with permission from Adenine Press, 2066 Central Ave, New York, USA (Web: http://www.jbsdonline.com). Source: Bai and Wang, 2006.

of A, T, H; Group 3 of G, P; Group 4 of D, E; and the last Group 5 of S, N, Q, R, K. Representing each group by one amino acid, a protein sequence can be reduced to a sequence of five letters only, which can then be used to generate a random walk in a 3D Euclidean space where the steps are in designated cardinal directions (Fig. 8). Taking a cue from the work by Gonzalez-Diaz et al. (2005), they use the charge information of the amino acids and the number of amino acids at each node of the walk to define four charge coupling numbers for each sequence from which, after some combinatorics, they generate a 60-component vector for each sequence. Applying this technique to beta-globin protein sequences from 15 species, they were able to quantitatively assess the similarities and dissimilarities between the proteins from comparison of the sequence vectors. This also led to generation of a distance matrix which, though not explicitly shown by the authors, can be used to draw the phylogenetic tree for this protein family. The results obtained by the authors' prescription are analogous to established data.

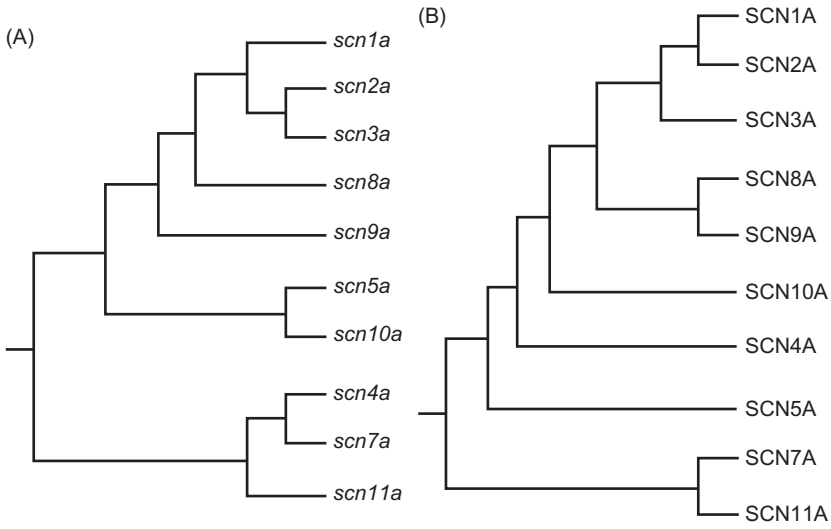


FIG. 7. Phylogenetic trees generated from Δp_R-matrix of nine-rat (*scn1a* to *scn11a*) voltage-gated sodium channel isoforms (A) and nine-human (SCN1A to SCN11A) voltage-gated sodium channel isoforms (B). Here the Δp_R-matrix is generated from the 20D algorithm proposed by Nandy et al. Reproduced with permission from IOS Press BV, Nieuwe Hemweg 6B, 1013 BG Amsterdam, The Netherlands. Web: www.bioinfo.de/isb/. Source: Nandy et al., 2009.

Thus, GRANCH methods are seen to be useful techniques to represent protein characteristics that can be easily computed while avoiding complications arising out of the need for multiple alignments (Altschul, 1989; Gotoh, 1993) and other modeling assumptions. The usefulness of these approaches and the reasonable agreement that we observe with standard results provide a good basis to investigate new phenomena such as viral issues which are the subjects of the next section.

III. Application to Viral Proteins

A. *Unique Features of Viral Proteins*

Viruses, the smallest biological entities, possess distinct groups of proteins holding a number of unique properties like high adaptability, high mutation rate, high structural flexibility, loose packing of the core, high

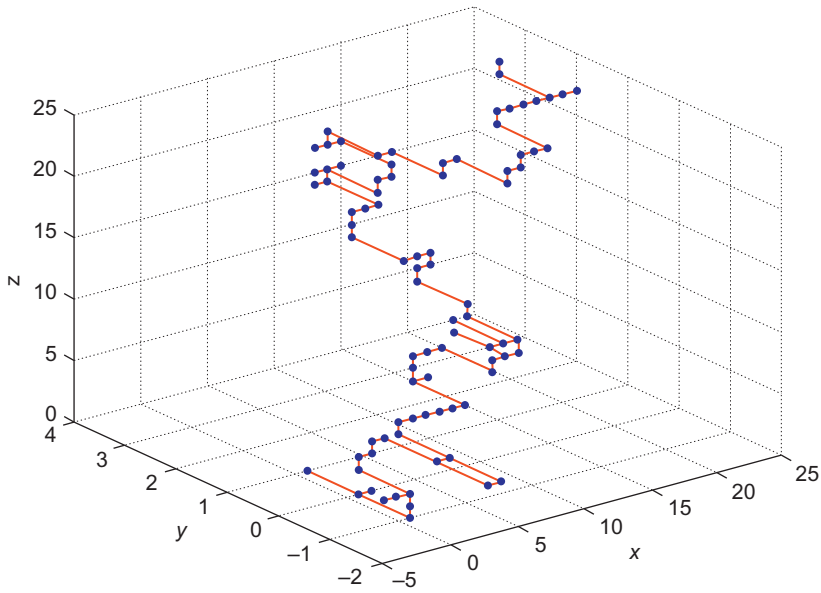


FIG. 8. 3D graph of the five-letter sequence of first 31 residues of Gorilla beta-protein IIAIAGEEKKAIAAIIGKIKIEEIGGEAIGK; each node may contain more than one amino acid. Reproduced with permission from Elsevier provided by Copyright Clearance Center (CCC). Web: http://www.sciencedirect.com/science/journal/03784371. Source: Li et al., 2009a.

proportion of disordered segments, among others (Koonin et al., 2009; Tokuriki et al., 2009; Kristensen et al., 2010). At the genome level, a very specific example of viral uniqueness resides in the existence of virus hallmark genes (Koonin et al., 2006), which play a central role in viral replication and structure, and are shared by a broad variety of viruses. In contrast to thermostable proteins like the heat-shock protein *Thermotoga maritime* (Tokuriki et al., 2009) which have specialized characteristics like high contact density, highly stable sequence composition, and highly compact structural scaffold, viral proteins necessarily have to be more complex to retain their functional characteristics in spite of the high variability. Another remarkable feature of viruses is the diversity in their genetic cycle. Altogether the variety of genetic strategies, genomic complexity, and global ecology of viral evolution lead to the formation of an infective long existing noncellular life form.

Currently prevention and treatment of viral diseases such as influenza rely on inactivated vaccines and antiviral drugs. Impact of mutational changes in amino acid residues on the stability, activity, and sensitivity of the target protein is a widely studied topic in antiviral drug design and for adequate remedy. The general causes like high mutability, altered specificity, environmental adaptability, etc., that are involved in generation of antiviral-resistant variety of the strains have been the main target of the major researches. Several investigations have focused upon phylogenetic relationships in viral evolution and transmission (Vijaykrishna et al., 2010) and reassortment (Lam et al., 2008; Owoade et al., 2008) and some other researchers are trying to correlate the evolution of genomic influenza varieties that affect humans and those that infect other life forms, for example, avian populations with a view that characterization of the causative proteins and determination or isolation of the conserved parts are useful approaches to combating viral diseases. Application of GRANCH techniques to viral proteins indicates one path to achieve this goal.

B. Two Viral Examples: The Avian and Swine Flu Viruses and the SARS Coronavirus

The smallest unit that makes a particular protein identifiable is an eight to nine amino acid long peptide segment. This fundamental unit is frequently used in wet lab and dry lab researches involving protein mass spectrometry data analysis, sequence alignment and phylogenetic algorithms, protein database handling software (Perkins et al., 1999), structure-based drug design, etc. Comparison of a large group of protein sequences often involves comparing the basic units of the proteins (single amino acids to complex structural levels like peptides, secondary structures, domains, etc.) and their organization. GRANCH provides novel ways for identifying sequences or peptides by generating an identifier (Nandy et al., 2009) with the aim to uniquely prescribe a protein and its compositional information. GRANCH techniques for protein sequences have emerged recently with promising applications to studies of coronavirus and the avian and swine flu viruses. We briefly describe the characteristics of the viruses, cover the GRANCH methods used in these studies, and state the significant results.

1. *Avian and Swine Flus*

The H5N1 avian flu erupted in Hong Kong in 1997 (Hatta et al., 2001) and got carried by migratory birds from its place of origin in South Central China to the rest of Asia and to Europe and Africa. The existence of the virus gene pool in China and continuous mutations among the virus strains have led to continued rapid spread worldwide by different carriers with sudden conflagrations erupting at different locations, among aquatic birds, poultry, and farm animals, and also infecting humans resulting in over 300 deaths out of 505 confirmed cases (from World Health Organization; updated August 31, 2010). The H5N1 virus, like all other influenza viruses, is an enveloped virus with an eight-segment single-stranded RNA in the core and two surface proteins on the envelope, the hemagglutinin and the neuraminidase, which are responsible for the glycosilation necessary for cell entry and exit. Although the number of fatalities in humans from this virus appears small, the rapid mutations that can occur in the RNA genome, and the possibility of whole gene or gene fragments shuffling between avian and mammalian hosts (Wu et al., 2008), are considered to carry the potential to cause a pandemic challenge. Since the inhibitors of this influenza virus, principally oseltamivir and zanamivir, act on the neuraminidase component of the H5N1 protein, continuous monitoring of the mutational changes in this gene assumes significance.

The H1N1 swine flu outbreak of 2009, often referred to as Mexican flu or just swine flu, though less severe pathogenically than the H5N1 avian flu, infected humans and spread worldwide rapidly enough to lead the WHO to declare it as a pandemic. The genomic structure closely parallels the H5N1 genome except for the important difference in the hemagglutinin subtype and the virus has responded well to the osletamivir therapy, implying again the importance of the neuraminidase in the control and remedy of these forms of influenza (Moscona, 2005). However, an escape route (Moscona, 2004, 2009) from this standard treatment through genetic mutations remains highly probable and provides ample impetus for continued research into development of alternate therapeutic strategies.

2. *SARS Coronavirus*

The SARS erupted on the world stage in 2003 (Gorbalenya et al., 2004) from its origins in South East Asia and was established to have been caused by a novel form of the coronavirus. Coronaviruses also are enveloped

viruses with a single-stranded multisegment RNA genome, but ranging in size from 16 to 31 kb (Lai, 1990). The virus primarily infects the upper respiratory and gastrointestinal tracts of mammals and birds, but the human SARS coronavirus also affects the lower respiratory tract. Experimental studies are complicated by the fact that the human coronaviruses are difficult to grow in the laboratory. While earlier only two coronaviruses were known, the HcoV-229E and HcoV-OC43 (Gorbalenya et al., 2004), after the SARS epidemic three more coronaviruses were identified by 2005, the SARS-CoV, the NL63 (van der Hoek et al., 2004), and HKU1 leading to interest in the evolutionary history of this virus.

C. *Graphical Representation Methods in Viral Studies*

1. *The 2D Method of Liao et al. (2006)*

As mentioned earlier (Section II), Liao et al. (2006) constructed 2D graphs with the four R-groups of the amino acids at predetermined angles on either side of the x-axis. For each nucleotide sequence, they constructed three separate graphs for the three reading frames of the gene sequence. For each graph, they defined a geometric center of mass x_0, y_0 and a covariance matrix CM as

$$x_0 = \sum \frac{x_i}{N}, \quad y_0 = \sum \frac{y_i}{N} \tag{3}$$

$$\mathrm{CM}_{xx} = \sum \frac{(x_i - x_0)(x_i - x_0)}{N} \tag{4}$$

$$\mathrm{CM}_{xy} = \mathrm{CM}_{yx} = \sum \frac{(x_i - x_0)(y_i - y_0)}{N} \tag{5}$$

$$\mathrm{CM}_{yy} = \sum \frac{(y_i - y_0)(y_i - y_0)}{N} \tag{6}$$

where the summations are over the subscript i which runs from 1 to N, the length of the sequence. The covariance matrix CM is a 2×2 square matrix with a leading eigenvalue λ. Thus, for the three graphs of each sequence there will be a set of three geometric centers of masses and three leading eigenvalues. From these eigenvalues, they defined a distance measure for two sequences i and j as

$$d_{ij} = \sqrt{\left\{ \left(\lambda_{i1} - \lambda_{j1}\right)^2 + \left(\lambda_{i2} - \lambda_{j2}\right)^2 + \left(\lambda_{i3} - \lambda_{j3}\right)^2 \right\}} \tag{7}$$

which can be used for studies of evolutionary relationships between species without having to make any evolutionary model assumptions or multiple alignments of the sequence.

2. *The 2D Method of Li et al. (2008)*

Another 2D graphical method has been described by Li et al. (2008) where they ascribe a 60-component vector to each of the proteins and construct a distance matrix

$$d_{ij} = \sqrt{\sum \left(x_{ir} - x_{jr}\right)^2} \tag{8}$$

where i and j refer to two different sequences and r=1,2,3,..., 60. This structure allows them to generate a phylogenetic tree in similar fashion to Liao et al.

3. *The 20D Method of Nandy et al.*

Taking a cue from the graphical representations of DNA sequences, Nandy et al. (2009) proposed an abstract 20D Cartesian coordinate system to generate a protein sequence walk by plotting one point for each amino acid in the sequence along a designated axis for that acid as shown in Table I; the choice of association is equivalent for all residues and can be arbitrarily assigned but once assigned will be fixed for the duration of the computation.

The walk as per the sequence will result in a series of points in the abstract 20D space generating a curve, each point on the walk being specified by 20 coordinate values. For example, for a protein sequence like MVHLTPEEKS the coordinate of the end point will then be (0,0,0,2,0,0,1,0,1,1,1,0,1,0,0,1,1,1,0,0) and the exercise can likewise be performed for any protein sequence.

Unlike some of the 2D graphical representation of DNA and protein sequences, there are no degeneracies or path retracements (Nandy et al., 2009) in this representation and all amino acids are represented on equal footing. While the disadvantage of this method is clear that the graph cannot be visualized, numerical characterization of the sequences

TABLE I
Assignment of Axis to Individual Amino Acids

Axis No.	Amino acid	Three-letter code	Single-letter code
1	Alanine	Ala	A
2	Cysteine	Cys	C
3	Aspartic acid	Asp	D
4	Glutamic acid	Glu	E
5	Phenylalanine	Phe	F
6	Glycine	Gly	G
7	Histidine	His	H
8	Isoleucine	Ile	I
9	Lysine	Lys	K
10	Leucine	Leu	L
11	Methionine	Met	M
12	Asparagine	Asn	N
13	Proline	Pro	P
14	Glutamine	Gln	Q
15	Arginine	Arg	R
16	Serine	Ser	S
17	Threonine	Thr	T
18	Valine	Val	V
19	Tryptophan	Trp	W
20	Tyrosine	Tyr	Y

can be easily computed as described below and used for comparison between sequences irrespective of sequence lengths (Nandy et al., 2009).

The quantification procedure in this representation characterizes a sequence by a weighted center of mass approach first used for DNA sequences with the CM coordinates given by

$$\mu_1 = \sum \frac{x_1}{N}, \mu_2 = \sum \frac{x_2}{N}, \ \mu_3 = \sum \frac{x_3}{N}, \ldots, \mu_{20} = \sum \frac{x_{20}}{N} \tag{9}$$

here the x_i's are the coordinate values of each point on the abstract curve and N, a normalization factor for the μ_i's, is the number of amino acids in the protein chain. Using these weighted averages, the procedure defines a protein graph vector p_R (μ_1, μ_2,..., μ_{20}) and a protein graph radius

$$p_R = \sqrt{(\mu_1^2 + \mu_2^2 + \cdots + \mu_n^2)} \tag{10}$$

Again, the distance between two sequences i and j can be defined as

$$\Delta p_{\mathrm{R}} = \sqrt{\left\{\left(\mu_{i_1} - \mu_{j_1}\right)^2 + \left(\mu_{i_2} - \mu_{j_2}\right)^2 + \cdots + \left(\mu_{i_{20}} - \mu_{j_{20}}\right)^2\right\}} \qquad (11)$$

where the sum is taken over all 20 coordinates. Obtaining a distance matrix from comparison of a family of sequences can enable generating a phylogenetic tree to study evolutionary relationships, again, as in all GRANCH methods, without having to introduce multiple alignments or any other model dependencies. Nandy et al. (2009) has successfully applied this algorithm in tree construction for human globin variants and between voltage-gated sodium channel isoforms.

It is to be noted that this numerical characterization method refers strictly to the identities of the amino acids and is transparent to their chemical properties, that is, no distinctions are made between residues that are mutationally conservative or nonconservative, between polar and nonpolar residues, between basic and acidic residues, etc., and all residues are treated at par. As in the case of the g_{R} for the DNA sequences, the p_{R} values also are found to be sensitive to changes in the amino acid sequences (Ghosh et al., 2009, 2010), and equal values of the p_{R} imply exact duplication of the amino acid composition and distribution along the sequences.

D. Results for the Coronavirus and the Flu Viruses

1. Phylogenetic Studies

For the 24 coronavirus genomes selected for their study, Liao et al. constructed a 24×24 distance matrix (Liao et al., 2006) from which they were able to generate a phylogenetic tree of the whole genome of the virus for different species using their 2D GRANCH technique. MSA, the popular phylogenetic tree generation algorithm, does not work properly for the whole genome, and the evolutionary model used may produce a wrong interpretation (Liao et al., 2006). Here, the phylogenetic tree (Fig. 9) obtained by their method clearly defined the evolutionary relationships between the whole genomes of 24 different species of the coronaviruses.

Liu and Wang (2010) using an L-tuple-based DNA representation constructed a set of $L \times L$ matrices whose mathematical characterization led to

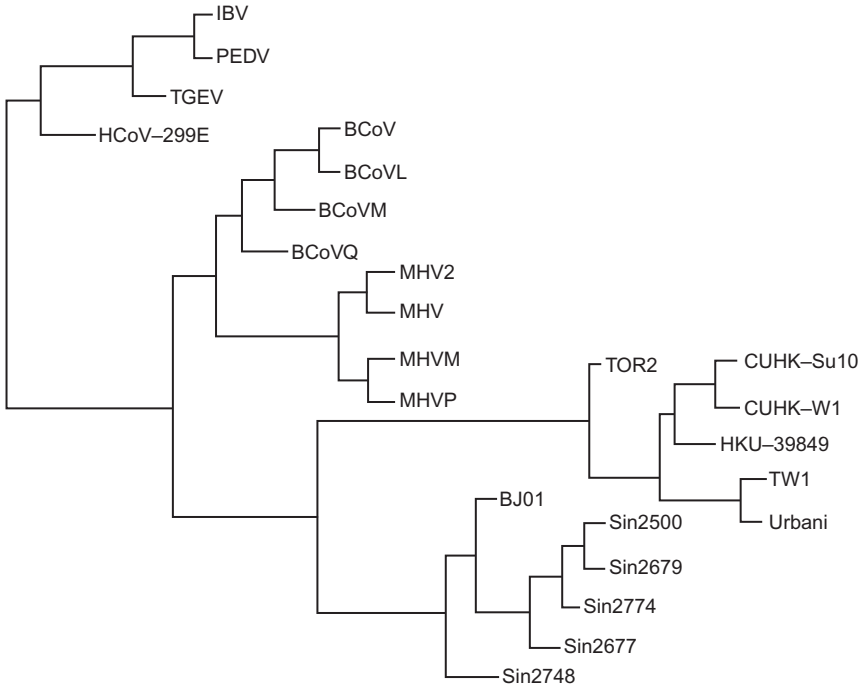


FIG. 9. The phylogenetic tree of the whole genome of 24 Coronavirus species. Reproduced with permission from Elsevier provided by Copyright Clearance Center (CCC). Web: http://www.sciencedirect.com/science/journal/00092614. Source: Liao et al., 2006.

characterization of the DNA sequences. Obtaining the distance matrices between a set of eight H5N1 avian flu genomes, they were able to generate a phylogenetic tree where the evolutionary relationships between the various strains of the virus were clearly identified.

2. *Similarity/Dissimilarity Analyses*

Nandy et al. (2007) and Ghosh et al. (2009) used the 2D GRANCH techniques for DNA sequences and the 20D methods for the protein sequences for analyses of global characteristics of over 680 H5N1 neuraminidase sequences to determine any systematic and exceptional behavior that may have arisen from mutational changes.

They found from detailed comparison of the g_R and p_R values that, at the protein level, only about 62% unique strains are observed, whereas for the nucleotide sequences the percentage of unique strains is considerably high at 80%, implying that about 22% (percentage of synonymous sequences to uniques) of the purportedly new strains of the neuraminidase gene have synonymous mutations (Ghosh et al., 2009). Considering the neuraminidase's segmented structure of transmembrane, stalk, and body regions, it was found that the body region appears proportionately less stable than the transmembrane or the stalk regions (Ghosh et al., 2009). In contrast, a 50-base segment at the 5′-end of the gene is found highly stable, mutations there being observed in less than 4.5% of the sequences at the RNA level and about 1.9% at the protein level raising the possibility of investigating this region as potentially useful for designing novel neuraminidase inhibitors.

The duplicate sequences identified by the p_R analysis showed sequence duplication across species and distributed over substantial distances in space and time (Ghosh et al., 2009). While localized or cosynchronous distributions can be expected to occur due to rapid dissemination of specific strains through viral shedding as one mechanism, the appearance of identical strains in geographically widely separated locations several thousand kilometers apart, or after a lapse of 2 years or more, is puzzling since viral genes are known to mutate rapidly in replication. The authors hypothesized that this may arise out of viral shedding in aquatic and nonaquatic habitats that are subsequently spread across wide regions by the migratory or local birds who themselves might not be infected but act merely as carrier agents. The p_R analysis from this technique also showed for the first time that recombinations between segments in the neuraminidase gene may have been taking place. Thus, sequence similarities and dissimilarities analysis done comparatively easily through the numerical descriptors can reveal many interesting features of viral spread and mutational changes.

3. Conserved and surface exposed peptide stretch identification

In nature, viruses are found to carry a great quantity of sequence variation in both the RNA and the protein level. These variations in viral sequences (Phillips et al., 1991; Chen and Deng, 2009) generally come from spontaneous mutation, adaptive forces, various mutagenic

effects, sequence recombination, etc. Such mutations are observed more in the parts in contact with the environment and therefore readily develop resistance to drugs. Conserved region in such parts of the protein, when determined, can be used for many purposes like structure-based drug design, viral proteins activity determination, vaccine design, etc. Ghosh et al. have applied the methods of GRANCH to determine just such regions in the H5N1 avian flu neuraminidase protein (Ghosh et al., 2010).

Using the 20D similarity/dissimilarity technique (Ghosh et al., 2009) through comparisons of p_R values, all the proteins in the dataset were scanned by a window size of 6–14 amino acids and the p_R values compared to find regions of least variability. This variability profile is then compared with a solvent accessibility profile to determine regions of low variability and high solvent accessibility implying that these identified regions would be accessible to drugs and vaccines and also offer target sites over many cycles of mutations. A view of the 3D structure also ensures that high surface exposed regions are actually selected. The authors determined six such regions on the neuraminidase protein (Fig. 10), of which the most promising appears to be the 50-base (16 amino acid) stretch at the 5′-end of the gene mentioned earlier. A special feature of this 16 amino acid long

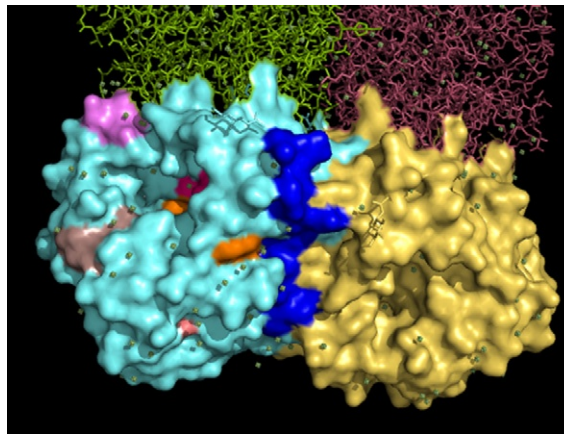


FIG. 10. Conserved surface exposed regions are shown in different colors in the cyan colored monomer of neuraminidase (other monomers are colored in magenta, green, and yellow). Here the six conserved regions are shown in six different colors. The conserved C-terminal portion is shown in blue.

peptide is its location on the dimeric interface (indicated by blue color in Fig. 10) of the quaternary structure of the neuraminidase protein, implying that any disruption of this stretch could interfere with the stability of the protein itself.

Nandy (2010) has reported on a similar work done on the rotavirus in association with several others. There, seven such distinct conserved surface exposed regions have been identified with this procedure. The most promising four regions have tested positive by epitope prediction servers (Peters et al., 2005; Vita et al., 2010) and reportedly hold promise for peptide drug and vaccine development.

IV. Conclusion

Thus, the GRANCH techniques for protein sequences are turning out to be quite useful novel method for analysis of proteins. Extension of these techniques to applications of different measurements as espoused by Gonzalez-Diaz et al. and others are opening up new methods to visualize and analyze experimental data and provide new insights. Applications by various authors to viral issues have generated new model independent ways to establish evolutionary relationships. In particular, the GRANCH techniques have provided for the first time a systematic method to determine conserved surface exposed peptide stretches on viral proteins that could be potentially very useful for drug and vaccine development.

References

Aguero-Chapin, G., Antunes, A., Ubeira, F. M., Chou, K. C., Gonzalez-Diaz, H. (2008). Comparative study of topological indices of macro/supramolecular RNA complex networks. *J. Chem. Inf. Model.* **48**, 2265–2277.

Aguero-Chapin, G., Gonzalez-Diaz, H., de la Riva, G., Rodriguez, E., Sanchez-Rodriguez, A., Podda, G., et al. (2008). MMM-QSAR recognition of ribonucleases without alignment: comparison with an HMM model and isolation from Schizosaccharomyces pombe, prediction, and experimental assay of a new sequence. *J. Chem. Inf. Model.* **48**, 434–448.

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

Aguero-Chapin, G., Varona-Santos, J., de la Riva, G. A., Antunes, A., Gonzalez-Vlla, T., Uriarte, E., et al. (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.

Air, G. M., Laver, W. G. (1989). The neuraminidase of influenza virus. *Proteins* **6**, 341–356.

Altschul, S. F. (1989). Gap costs for multiple sequence alignment. *J. Theor. Biol.* **138**, 297–309.

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

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

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

Baker, D., Sali, A. (2001). Protein structure prediction and structural genomics. *Science* **294**, 93–96.

Baldwin, J., Chothia, C. (1979). Haemoglobin: the structural changes related to ligand binding and its allosteric mechanism. *J. Mol. Biol.* **129**, 175–220.

Ball, L. K., Evans, G., Bostrom, A. (1998). Risky business: challenges in vaccine risk communication. *Pediatrics* **101**, 453–458.

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

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

Basak, S. C., Mills, D., Hawkins, D. M. (2011). Characterization of dihydrofolate reductases from multiple strains of *Plasmodium falciparum* using mathematical descriptors of their inhibitors. *Chem. Biodivers* **8**, 440–453.

Bejarano, L. A., Gonzalez, C. (1999). Motif trap: a rapid method to clone motifs that can target proteins to defined subcellular localisations. *J. Cell Sci.* **112**(Pt 23), 4207–4211.

Berman, H., Henrick, K., Nakamura, H., Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **35**, D301–D303.

Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S., Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**, 1633–1649.

Chaires, J. B. (1997). Energetics of drug-DNA interactions. *Biopolymers* **44**, 201–215.

Chaires, J. B. (1998). Drug–DNA interactions. *Curr. Opin. Struct. Biol.* **8**, 314–320.

Chayen, N. E. (2004). Turning protein crystallisation from an art into a science. *Curr. Opin. Struct. Biol.* **14**, 577–583.

Chayen, N. E. (2009). High-throughput protein crystallization. *Adv. Protein. Chem. Struct. Biol.* **77**, 1–22.

Chayen, N. E., Saridakis, E. (2008). Protein crystallization: from purified protein to diffraction-quality crystal. *Nat Methods* **5**, 147–153.

Chen, J., Deng, Y. M. (2009). Influenza virus antigenic variation, host antibody production and new approach to control epidemics. *Virol. J.* **6**, 30.

Chicault, M., Luu Duc, C., Boucherle, A. (1981). Drug protein interactions. *Arzneimittelforschung* **31**, 1015–1020.

Chikenji, G., Fujitsuka, Y., Takada, S. (2006). Shaping up the protein folding funnel by local interaction: lesson from a structure prediction study. *Proc. Natl. Acad. Sci. USA* **103**, 3141–3146.

Chou, K. C. (2001). Prediction of protein signal sequences and their cleavage sites. *Proteins* **42**, 136–139.

Chou, K. C., Cai, Y. D. (2003). Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J. Cell. Biochem.* **90**, 1250–1260.

Chou, P. Y., Fasman, G. D. (1974a). Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry* **13**, 211–222.

Chou, P. Y., Fasman, G. D. (1974b). Prediction of protein conformation. *Biochemistry* **13**, 222–245.

Colgrove, J., Bayer, R. (2005). Could it happen here? Vaccine risk controversies and the specter of derailment. *Health Aff. (Millwood)* **24**, 729–739.

Cruz-Monteagudo, M., Gonzalez-Diaz, H., Borges, F., Dominguez, E. R., Cordeiro, M. N. (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.

Dea-Ayuela, M. A., Perez-Castillo, Y., Meneses-Marcel, A., Ubeira, F. M., Bolas-Fernandez, F., Chou, K. C., et al. (2008). HP-Lattice QSAR for dynein proteins: experimental proteomics (2D-electrophoresis, mass spectrometry) and theoretic study of a *Leishmania infantum* sequence. *Bioorg. Med. Chem.* **16**, 7770–7776.

Deml, L., Bojak, A., Steck, S., Graf, M., Wild, J., Schirmbeck, R., et al. (2001). Multiple effects of codon usage optimization on expression and immunogenicity of DNA candidate vaccines encoding the human immunodeficiency virus type 1 Gag protein. *J. Virol.* **75**, 10991–11001.

Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., et al. (2005). The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.* **33**, D233–D237.

Dill, K. A., Ozkan, S. B., Shell, M. S., Weikl, T. R. (2008). The protein folding problem. *Annu. Rev. Biophys.* **37**, 289–316.

Dill, K. A., Ozkan, S. B., Weikl, T. R., Chodera, J. D., Voelz, V. A. (2007). The protein folding problem: when will it be solved? *Curr. Opin. Struct. Biol.* **17**, 342–346.

Dingermann, T. (2008). Recombinant therapeutic proteins: production platforms and challenges. *Biotechnol. J.* **3**, 90–97.

Dobson, C. M. (2004). Principles of protein folding, misfolding and aggregation. *Semin. Cell Dev. Biol.* **15**, 3–16.

Doherty, G. J., McMahon, H. T. (2008). Mediation, modulation, and consequences of membrane-cytoskeleton interactions. *Annu. Rev. Biophys.* **37**, 65–95.

Dutta, S., Berman., H. M., Bluhm, W. F. (2007). Using the tools and resources of the RCSB protein data bank. *Curr. Protoc. Bioinformatics* **20**, 1.9.1–1.9.24.

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361–365.

Engelman, D. M., Steitz, T. A., Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* **15**, 321–353.

Estrada, E. (2002). Characterization of the folding degree of proteins. *Bioinformatics* **18**, 697–704.

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

Eswar, N., Eramian, D., Webb, B., Shen, M. Y., Sali, A. (2008). Protein structure modeling with MODELLER. *Methods Mol. Biol.* **426**, 145–159.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-y., Pieper, U., Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* **50**, 2.9.1–2.9.31.

Fasman, G. D. (1989). Protein conformational prediction. *Trends Biochem. Sci.* **14**, 295–299.

Feldman, H. J., Hogue, C. W. (2002). Probabilistic sampling of protein conformations: new hope for brute force? *Proteins* **46**, 8–23.

Fernandez, M., Caballero, J., Fernandez, L., Abreu, J. I., Acosta, G. (2008). Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines. *Proteins* **70**, 167–175.

Frieden, C. (1971). Protein-protein interaction and enzymatic activity. *Annu. Rev. Biochem.* **40**, 653–696.

Frokjaer, S., Otzen, D. E. (2005). Protein drug stability: a formulation challenge. *Nat. Rev. Drug Discov.* **4**, 298–306.

Gates, M. A. (1986). A simple way to look at DNA. *J. Theor. Biol.* **119**, 319–328.

Geigert, J. (1989). Overview of the stability and handling of recombinant protein drugs. *J. Parenter. Sci. Technol.* **43**, 220–224.

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., Nandy, A., Nandy, P., Gute, B. D., Basak, S. C. (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**, 2627–2638.

Ghosh, K., Ozkan, S. B., Dill, K. A. (2007). The ultimate speed limit to protein folding is conformational searching. *J. Am. Chem. Soc.* **129**, 11920–11927.

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

Gold, H. S., Moellering, R. C., Jr. (1996). Antimicrobial-drug resistance. *N. Engl. J. Med.* **335**, 1445–1453.

Goldsby, R. A., Kindt, T. J., Osborne, B. A. (2000). Overview of Immune System. Kuby Immunology. W.H. Freeman and Company, United State of America, pp. 3.

Gonzalez-Diaz, H., Gonzalez-Diaz, Y., Santana, L., Ubeira, F. M., Uriarte, E. (2008). Proteomics, networks and connectivity indices. *Proteomics* **8**, 750–778.

Gonzalez-Diaz, H., Molina, R., Uriarte, E. (2005). Recognition of stable protein mutants with 3D stochastic average electrostatic potentials. *FEBS Lett.* **579**, 4297–4301.

Gonzalez-Diaz, H., Perez-Montoto, L. G., Duardo-Sanchez, A., Paniagua, E., Vazquez-Prieto, S., Vilas, R., et al. (2009). Generalized lattice graphs for 2D-visualization of biological information. *J. Theor. Biol.* **261**, 136–147.

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

Gonzalez-Diaz, H., Saiz-Urra, L., Molina, R., Gonzalez-Diaz, Y., Sanchez-Gonzalez, A. (2007). Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments. *J. Comput. Chem.* **28**, 1042–1048.

Gonzalez-Diaz, 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**, 1015–1029.

Gorbalenya, A. E., Snijder, E. J., Spaan, W. J. (2004). Severe acute respiratory syndrome coronavirus phylogeny: toward consensus. *J. Virol.* **78**, 7863–7866.

Gotoh, O. (1993). Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Comput. Appl. Biosci.* **9**, 361–370.

Gunn, P. R., Sato, F., Powell, K. F., Bellamy, A. R., Napier, J. R., Harding, D. R., et al. (1985). Rotavirus neutralizing protein VP7: antigenic determinants investigated by sequence analysis and peptide synthesis. *J. Virol.* **54**, 791–797.

Gurunathan, S., Klinman, D. M., Seder, R. A. (2000). DNA vaccines: immunology, application, and optimization. *Annu. Rev. Immunol.* **18**, 927–974.

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

Hatta, M., Gao, P., Halfmann, P., Kawaoka, Y. (2001). Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* **293**, 1840–1842.

Hawkins, D. M., Basak, S. C., Kraker, J., Geiss, K. T., Witzmann, F. A. (2006). Combining chemodescriptors and biodescriptors in quantitative structure-activity relationship modeling. *J. Chem. Inf. Model.* **46**, 9–16.

Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic Acids Res.* **18**, 2163–2170.

Jiang, M., Zhu, B. (2005). Protein folding on the hexagonal lattice in the HP model. *J. Bioinform. Comput. Biol.* **3**, 19–34.

Karp, G. (2008). Biological molecules. Cell and Molecular Biology. John Wiley & Sons (Asia) Pte Ltd, Asia, p. 31.

Kendrew, J. C. (1959). Structure and function in myoglobin and other proteins. *Fed. Proc.* **18**, 740–751.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662–666.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., et al. (1960). Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution. *Nature* **185**, 422–427.

Kinjo, A. R., Yamashita, R., Nakamura, H. (2010). PDBj Mine: design and implementation of relational database interface for Protein Data Bank Japan. *Database (Oxford) Database published online August 25, 2010)* **2010**, baq021. http://database.oxfordjournals.org/cgi/crossref-forward-links/2010/0/baq021.

Koch-Weser, J., Sellers, E. M. (1976). Binding of drugs to serum albumin (first of two parts). *N. Engl. J. Med.* **294**, 311–316.

Koonin, E. V., Senkevich, T. G., Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biol. Direct* **1**, 29.

Koonin, E. V., Wolf, Y. I., Nagasaki, K., Dolja, V. V. (2009). The complexity of the virus world. *Nat. Rev. Microbiol.* **7**, 250.

Kramer, L. D., Li, J., Shi, P. Y. (2007). West Nile virus. *Lancet Neurol.* **6**, 171–181.

Kristensen, D. M., Mushegian, A. R., Dolja, V. V., Koonin, E. V. (2010). New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.* **18**, 11–19.

Kumar, S., Nei, M., Dudley, J., Tamura, K. (2008). MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform.* **9**, 299–306.

Lai, M. M. (1990). Coronavirus: organization, replication and expression of genome. *Annu. Rev. Microbiol.* **44**, 303–333.

Lam, T. T., Hon, C. C., Pybus, O. G., Kosakovsky Pond, S. L., Wong, R. T., Yip, C. W., et al. (2008). Evolutionary and transmission dynamics of reassortant H5N1 influenza virus in Indonesia. *PLoS Pathog.* **4**, e1000130.

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

Lee, D. G., Kim, P. I., Park, Y., Woo, E. R., Choi, J. S., Choi, C. H., et al. (2002). Design of novel peptide analogs with potent fungicidal activity, based on PMAP-23 antimicrobial peptide isolated from porcine myeloid. *Biochem. Biophys. Res. Commun.* **293**, 231–238.

Lengauer, T., Zimmer, R. (2000). Protein structure prediction methods for drug design. *Brief. Bioinform.* **1**, 275–288.

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

Li, Z., Baker, M. L., Jiang, W., Estes, M. K., Prasad, B. V. (2009). Rotavirus architecture at subnanometer resolution. *J. Virol.* **83**, 1754–1766.

Li, C., Wang, J. (2005). New invariant of DNA sequences. *J. Chem. Inf. Model.* **45**, 115–120.

Li, C., Xing, L., Wang, X. (2008). 2-D graphical representation of protein sequences and its application to coronavirus phylogeny. *BMB Rep.* **41**, 217–222.

Li, C., Yu, X., Yang, L., Zheng, X., Wang, Z. (2009). 3-D maps and coupling numbers for protein sequences. *Physica A* **388**, 1967–1972.

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

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

Locatelli, F., Vecchio, L. D. (2001). Darbepoetin alfa Amgen. *Curr. Opin. Investig. Drugs* **2**, 1097–1104.

Lopez, J. A., Maldonado, A. J., Gerder, M., Abanero, J., Murgich, J., Pujol, F. H., et al. (2005). Characterization of neuraminidase-resistant mutants derived from rotavirus porcine strain OSU. *J. Virol.* **79**, 10369–10375.

MacCallum, R. M., Martin, A. C., Thornton, J. M. (1996). Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.* **262**, 732–745.

Makowski, L., Rodi, D. J., Mandava, S., Minh, D. D., Gore, D. B., Fischetti, R. F. (2008). Molecular crowding inhibits intramolecular breathing motions in proteins. *J. Mol. Biol.* **375**, 529–546.

Markley, J. L., Ulrich, E. L., Berman, H. M., Henrick, K., Nakamura, H., Akutsu, H. (2008). BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J. Biomol. NMR* **40**, 153–155.

Menne, K. M., Hermjakob, H., Apweiler, R. (2000). A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* **16**, 741–742.

Meyer, M. C., Guttman, D. E. (1968). The binding of drugs by plasma proteins. *J. Pharm. Sci.* **57**, 895–918.

Morrow, M. P., Weiner, D. B. (2010). DNA drugs come of age. *Sci. Am.* **303**, 49–53.

Moscona, A. (2004). Oseltamivir-resistant influenza? *Lancet* **364**, 733–734.

Moscona, A. (2005). Oseltamivir resistance–disabling our influenza defenses. *N. Engl. J. Med.* **353**, 2633–2636.

Moscona, A. (2009). Global transmission of oseltamivir-resistant influenza. *N. Engl. J. Med.* **360**, 953–956.

Mount, D. (2004). Bioinformatics, sequence and genome analysis. 2nd edn. Cold Spring Harbor Press, Cold Spring Harbor, NY, pp. 337.

Munteanu, C. R., Magalhaes, A. L., Uriarte, E., Gonzalez-Diaz, H. (2009). Multi-target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. *J. Theor. Biol.* **257**, 303–311.

Nakamura, H., Ito, N., Kusunoki, M. (2002). Development of PDBj: Advanced database for protein structures. *Tanpakushitsu Kakusan Koso* **47**, 1097–1101.

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

Nandy, A. (2009). Empirical relationship between intra-purine and intra-pyrimidine differences in conserved gene sequences. *PLoS ONE* **4**, e6829.

Nandy, A. (2010). Towards stable vaccines: Contributions from DNA and protein numerical characterization studies. 50th Anniversary Celebration with Mathematical Chemistry, Universidad de Pamplona, Pamplona, Colombia.

Nandy, A., Basak, S. C. (2010). New approaches to drug-DNA interactions based on graphical representation and numerical characterization of DNA sequences. *Curr. Comput. Aided Drug Des.* **6**, 283–289.

Nandy, A., Basak, S. C., Gute, B. D. (2007). Graphical representation and numerical characterization of H5N1 avian flu neuraminidase gene sequence. *J. Chem. Inf. Model.* **47**, 945–951.

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

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

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

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

Needleman, S. B., Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.

Novic, M., Randic, M. (2008). Representation of proteins as walks in 20-D space. *SAR QSAR Environ. Res.* **19**, 317–337.

Opitz, C. A., Kulke, M., Leake, M. C., Neagoe, C., Hinssen, H., Hajjar, R. J., et al. (2003). Damped elastic recoil of the titin spring in myofibrils of human myocardium. *Proc. Natl. Acad. Sci. USA* **100**, 12688–12693.

Otvos, L., Jr. (2008). Peptide-based drug design: here and now. *Methods Mol. Biol.* **494**, 1–8.

Owens, J. (2004). Building blocks for peptide drugs. *Nat. Rev. Drug Discov.* **3**, 476.

Owoade, A. A., Gerloff, N. A., Ducatez, M. F., Taiwo, J. O., Kremer, J. R., Muller, C. P. (2008). Replacement of sublineages of avian influenza (H5N1) by reassortments, sub-Saharan Africa. *Emerg. Infect. Dis.* **14**, 1731–1735.

Packhaeuser, C. B., Schnieders, J., Oster, C. G., Kissel, T. (2004). In situ forming parenteral drug delivery systems: an overview. *Eur. J. Pharm. Biopharm.* **58**, 445–455.

Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., et al. (1992). Long-range correlations in nucleotide sequences. *Nature* **356**, 168–170.

Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567.

Perutz, M. F. (1960). Structure of hemoglobin. *Brookhaven Symp. Biol.* **13**, 165–183.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., North, A. C. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature* **185**, 416–422.

Perutz, M. F., Weisz, O. (1947). Crystal structure of human carboxyhaemoglobin. *Nature* **160**, 786.

Peters, B., Sidney, J., Bourne, P., Bui, H. H., Buus, S., Doh, G., et al. (2005). The design and implementation of the immune epitope database and analysis resource. *Immunogenetics* **57**, 326–336.

Phillips, R. E., Rowland-Jones, S., Nixon, D. F., Gotch, F. M., Edwards, J. P., Ogunlesi, A. O., et al. (1991). Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature* **354**, 453–459.

Powers, E. T., Powers, D. L. (2008). Mechanisms of protein fibril formation: nucleated polymerization with competing off-pathway aggregation. *Biophys. J.* **94**, 379–391.

Ramachandran, G. N., Ramakrishnan, C., Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99.

Ramachandran, G. N., Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Protein Chem.* **23**, 283–438.

Randic, M. (2004). Graphical representations of DNA as 2-D map. *Chem. Phys. Lett.* **386**, 468.

Randic, M. (2006). Spectrum-like graphical representation of DNA based on codons. *Acta Chim. Slov.* **53**, 477–485.

Randic, M., Butina, D., Zupan, J. (2006). Novel 2-D graphical representation of proteins. *Chem. Phys. Lett.* **419**, 528–532.

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

Randic, M., Novic, M., Vracko, M. (2008). On novel representation of proteins based on amino acid adjacency matrix. *SAR QSAR Environ. Res.* **19**, 339–349.

Randic, M., Vracko, M., Lers, N., Plavsic, D. (2003a). Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* **371**, 202.

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

Randic, M., Vracko, M., Nandy, A., Basak, S. C. (2000). 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., Zupan, J., Novic, M. (2003c). Compact 2-D graphical representation of DNA. *Chem. Phys. Lett.* **373**, 558.

Randic, M., Zupan, J., Balaban, A. T. (2004). Unique graphical representation of protein sequences based on nucleotide triplet codons. *Chem. Phys. Lett.* **397**, 247–252.

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

Rayment, I. (1996). Kinesin and myosin: molecular motors with similar engines. *Structure* **4**, 501–504.

Reid, A. H., Fanning, T. G., Janczewski, T. A., Taubenberger, J. K. (2000). Characterization of the 1918 ''Spanish'' influenza virus neuraminidase gene. *Proc. Natl. Acad. Sci. USA* **97**, 6785–6790.

Retief, J. D. (2000). Phylogenetic analysis using PHYLIP. *Methods Mol. Biol.* **132**, 243–258.

Roach, P., Farrar, D., Perry, C. C. (2005). Interpretation of protein adsorption: surface-induced conformational changes. *J. Am. Chem. Soc.* **127**, 8168–8173.

Roach, P., Farrar, D., Perry, C. C. (2006). Surface tailoring for controlled protein adsorption: effect of topography at the nanometer scale and chemistry. *J. Am. Chem. Soc.* **128**, 3939–3945.

Russell, R. J., Haire, L. F., Stevens, D. J., Collins, P. J., Lin, Y. P., Blackburn, G. M., et al. (2006). The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **443**, 45–49.

Ruvkun, G. B., Ausubel, F. M. (1981). A general method for site-directed mutagenesis in prokaryotes. *Nature* **289**, 85–88.

Sanger, F. (1952). The arrangement of amino acids in proteins. *Adv. Protein Chem.* **7**, 1–67.

Sanger, F., Thompson, E. O. (1953). The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem. J.* **53**, 353–366.

Sanger, F., Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. *Biochem. J.* **49**, 463–481.

Schellekens, H. (2002). Bioequivalence and the immunogenicity of biopharmaceuticals. *Nat. Rev. Drug Discov.* **1**, 457–462.

Shen, H. B., Chou, K. C. (2009). Predicting protein fold pattern with functional domain and sequential evolution information. *J. Theor. Biol.* **256**, 441–446.

Smith, T. F., Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.

Smith, T. F., Waterman, M. S., Burks, C. (1985). The statistical distribution of nucleic acid similarities. *Nucleic Acids Res.* **13**, 645–656.

Tamura, K., Dudley, J., Nei, M., Kumar, S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599.

Tangri, S., Mothe, B. R., Eisenbraun, J., Sidney, J., Southwood, S., Briggs, K., et al. (2005). Rationally engineered therapeutic proteins with reduced immunogenicity. *J. Immunol.* **174**, 3187–3196.

Teague, S. J. (2003). Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* **2**, 527–541.

Todeschini, R., Ballabio, D., Consonni, V., Mauri, A. (2008). A new similarity/diversity measure for the characterization of DNA sequences. *Croat. Chem. Acta* **81**, 657–664.

Todeschini, R., Consonni, V., Mauri, A., Ballabio, D. (2006). Characterization of DNA primary sequences by a new similarity/diversity measure based on the partial ordering. *J. Chem. Inf. Model.* **46**, 1905–1911.

Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezovsky, I. N., Tawfik, D. S. (2009). Do viral proteins possess unique biophysical features? *Trends Biochem. Sci.* **34**, 53–59.

Ulmer, J. B., Deck, R. R., Yawman, A., Friedman, A., Dewitt, C., Martinez, D., et al. (1996). DNA vaccines for bacteria and viruses. *Adv. Exp. Med. Biol.* **397**, 49–53.

Ulmer, J. B., Sadoff, J. C., Liu, M. A. (1996). DNA vaccines. *Curr. Opin. Immunol.* **8**, 531–536.

van der Hoek, L., Pyrc, K., Jebbink, M. F., Vermeulen-Oost, W., Berkhout, R. J., Wolthers, K. C., et al. (2004). Identification of a new human coronavirus. *Nat. Med.* **10**, 368–373.

Velankar, S., Best, C., Beuth, B., Boutselakis, C. H., Cobley, N., Sousa Da Silva, A. W., et al. (2010). PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.* **38**, D308–D317.

Vijaykrishna, D., Poon, L. L., Zhu, H. C., Ma, S. K., Li, O. T., Cheung, C. L., et al. (2010). Reassortment of pandemic H1N1/2009 influenza A virus in swine. *Science* **328**, 1529.

Vilar, S., González-Díaz, H. (2010). QSPR models for human Rhinovirus surface networks. *In* Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and Social Networks, (González-Díaz, H. and Munteanu, C. R. Eds.), pp. 145–161.

Vilar, S., Gonzalez-Diaz, H., Santana, L., Uriarte, E. (2008). QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks. *J. Comput. Chem.* **29**, 2613–2622.

Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., et al. (2010). The immune epitope database 2.0. *Nucleic Acids Res.* **38**, D854–D862.

Wang, J., Wang, W. (1999). A computational approach to simplifying the protein folding alphabet. *Nat. Struct. Biol.* **6**, 1033–1038.

Wiesner, I., Wiesnerova, D. (2010). 2D random walk representation of Begonia × tuberhybrida multiallelic loci used for germplasm identification. *Biologia Plantarum* **54**, 353–356.

Wu, W. L., Chen, Y., Wang, P., Song, W., Lau, S. Y., Rayner, J. M., et al. (2008). Antigenic profile of avian H5N1 viruses in Asia from 2002 to 2007. *J. Virol.* **82**, 1798–1807.

Yang, X., Yu, X. (2009). An introduction to epitope prediction methods and software. *Rev. Med. Virol.* **19**, 77–96.

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