BIOINFORMATIC ANALYSIS OF VIRAL GENOMIC SEQUENCES AND CONCEPTS OF GENOME-SPECIFIC RATIONAL VACCINE DESIGN

by

Sharmistha P. Chatterjee

A Dissertation Submitted to the Faculty of

The College of Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Florida Atlantic University

Boca Raton, Florida

May 2013

BIOINFORMATIC ANALYSIS OF VIRAL GENOMIC SEQUENCES AND CONCEPTS OF GENOME-SPECIFIC RATIONAL VACCINE DESIGN

by

Sharmistha P. Chatterjee

This dissertation was prepared under the direction of the candidate's dissertation advisor, Dr. Perambur S. Neelakanta, Department of Computer and Electrical Engineering and Computer Science, and has been approved by the members of her supervisory committee. It was submitted to the faculty of the College of Engineering and Computer Science and was accepted in partial fulfillment of the requirements of the degree of Doctor of Philosophy.

SUPERVISORY COMMITTEE:

Perambur S. Neelakanta, Ph.D.
Dissertation Advisor

Mirjana Pavlovic
Mirjana Pavlovic, Ph.D.

Abhijit Pandya
Abhijit Pandya, Ph.D.

Dolores De Groff
Dolores De Groff, Ph.D.

Shihong Huang
Shihong Huang, Ph.D.

Borko Furht
Borko Furht, Ph.D.
Chair, Department of Computer and Electrical Engineering
and Computer Science

Mohammad Illyas, Ph.D.
Interim Dean, College of Engineering and Computer Science

Barry T. Rosson
Barry T. Rosson, Ph. D.
Dean, Graduate College

March 27, 2013
Date

ACKNOWLEDGMENTS

This study would not have been completed without the direction and guidance from Dr. Perambur S. Neelakanta, Committee Chairperson and dissertation advisor. This dissertation has been one of the most significant academic challenges I have ever had to face. Dr. Neelakanta was always available to provide constructive feedback and guidance despite his many other commitments.

I also thank Dr. Mirjana Pavlovic for her advice and providing her educated assessment and valuable knowledge to this study conducted. My sincere thanks to Dr. Pavlovic for providing her insights and help throughout this study.

I take this opportunity to also thank the Committee Members, Dr. Abhijit Pandya, Dr. Dolores DeGroff and Dr. Shihong Huang for their time serving as members of the committee and providing their advice and comments to this dissertation.

Finally, I want to greatly thank my husband for his patience, encouragement and moral support, my son for helping me out in whatever ways he could think of, my parents and parents-in-law for their constant support, encouragement and profound understanding and last, but, not the least, my friends for being there for me and helping me out without even my requesting them for their help.

ABSTRACT

Author:	Sharmistha P. Chatterjee
Title:	Bioinformatic Analysis of Viral Genomic Sequences and Concepts of Genome-specific Rational Vaccine Design
Institution:	Florida Atlantic University
Dissertation Advisor:	Dr. Perambur S. Neelakanta
Degree:	Doctoral of Philosophy
Year:	2013

This research is concerned with analyzing a set of viral genomes to elucidate the underlying characteristics and determine the information-theoretic aspects of the genomic signatures. The goal of this study thereof, is tailored to address the following: (i) Reviewing various methods available to deduce the features and characteristics of genomic sequences of organisms in general, and particularly focusing on the genomes pertinent to viruses; (ii) applying the concepts of information-theoretics (entropy principles) to analyze genomic sequences; (iii) envisaging various aspects of biothermodynamic energetics so as to determine the framework and architecture that decide the stability and patterns of the subsequences in a genome; (iv) evaluating the genomic details using spectral-domain techniques; (v) studying fuzzy considerations to ascertain the overlapping details in genomic sequences; (vi) determining the common subsequences among various strains of a virus by logistically regressing the data obtained *via* entropic, energetics and spectral-domain exercises; (vii) differentiating informational profiles of coding and non-coding regions in a DNA sequence to locate aberrant (cryptic) attributes evolved as a result of mutational changes and (viii) finding the signatures of CDS of genomes of viral strains toward rationally conceiving plausible designs of vaccines.

Commensurate with the topics indicated above, necessary simulations are proposed and computational exercises are performed (with MatLab^{TM} R2009b and other software as needed). Extensive data gathered from open-literature are used thereof and, simulation results are verified. Lastly, results are discussed, inferences are made and open-questions are identified for future research.

BIOINFORMATIC ANALYSIS OF VIRAL GENOMIC SEQUENCES AND CONCEPTS OF GENOME-SPECIFIC RATIONAL VACCINE DESIGN

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 General

In the context of modern biological and medical sciences, there are compelling reasons to elucidate the genetic information of living systems. Specifically, information on human genomics is necessary for the purpose of understanding the underlying physico-chemical attributes as well as the gene details useful for diagnostic and therapeutic efforts. In addition to knowing such genomic aspects of human (or animals, at large), it is necessary to elucidate the gene information of innumerable pathogens that cause adverse effects on living systems. Knowing their genetic features and molecular details is essential to formulate methods of preventive schedules against adverse influences from the universe of germs. In view of the above, knowing the intricacies of human genome and concurrently understanding the molecular biology of pathogens that may invade and embed into humans are imperative in deducing diagnostic and therapeutic options. Relevant efforts can facilitate the control and curative measures across innumerable diseases that prevail in the society.

In the sphere of grossly pervading, but immensely agonizing disease processes, a plethora of exercise in bioinformatics has been directed in evolving methods of comprehensive analyses to extract the underlying informatics. In past, relevant foundation laid by the Human Genome Project (HGP) [1.1] has enabled extensive explorations on the unlimited frontiers of microbiology, wherein skillful melding of biological know-how through tools of computer science, concepts of physics, methods of mathematics, techniques of statistics, avenues of clinical research etc. are diligently adopted. Further, fuelled by the sequencing of human genome *via* innumerable wet-lab studies along with supplementing computational strategies, exists today a host of databases created with brimming of information on nucleic acids, proteins, system-level biology, biochemical and biophysical details supported by relevant annotations [1.2 - 1.6] and descriptions as regard to a vast trove of biological species- small and big.

Notwithstanding the abundance of such data prevailing, more and better understanding of biological details enabled by wet-lab instrumentation and super-computing capabilities of modern times, have led to the quest of knowing the genomic blueprint at various strata of significance. This quest thereof has posed inquisitiveness of knowing something that is unknown yet. Rightly, as said in [1.7] "A remarkable landscape of opportunity lies before next generation of biologist" to explore the morsels of microbiology and envisage the intricacies of bioinformatics beneficial to living systems at large. Supplementing such opportune direction, are pedagogical principles of bioinformatics covering the major areas such as analyzing sequences of whole genomes and/or variations, exploring microarrays, portraying the patterns of proteomics and scything the systems biology as a whole.

With such avenues of research unlimited, it is considered in the present research, a set of select topics viable for some novel analyses that could be eventually be adopted for rational vaccine design (RVD) considerations. Hence, the scope of the work can be specified as follows:

1.2 Scope of the Research

The topics of research advocated in the present study have the general scope of applying information-theoretic algorithms, energetics principle and spectral-domain concepts to analyze genome sequences in general. Further, these are applied exclusively on a set of viral genomic sequences so as to elucidate the underlying genomic characteristics viably useful for vaccine design applications. This conceived scope as above can be expanded as a gist of objectives listed in the following section.

1.3 Objective of the Research

The research performed and presented in this dissertation is centered on the following objectives commensurate with the scope outlined above:

- Applying entropy considerations and formulating information-theoretic methods (in Shannon's sense) so as to analyze genomic sequences. Hence, relevant concepts of statistical divergence and/or distance measures are invoked thereof and applied to the sequence analyses in question

- Using the concept of biothermodynamic energetics [1.8 - 1.10] in the framework of genetic stability, the buried details of subsequences in DNA and/or RNA architectures are determined

- Fuzzy considerations in genomic information-theoretics [1.11] are studied for the purpose of ascertaining overlapping details in genomic sequences

- The efforts of entropy and energetics-based analyses as above are supplemented further with the evaluation of corresponding details using spectral domain techniques [1.12 - 1.14]

- The pursued analyses are aimed at knowing the following: (i) Differentiating the informational profiles of coding and non-coding regions in a DNA sequence [1.15]: (ii) detecting buried signals (such as, splice-junctions [1.16]) within a DNA; (iii) locating aberrant (cryptic) attributes that exists in DNA sequences [1.17] as a result of mutational changes; (iv) observing fuzzy details on overlapping subsequences (v) finding the signatures of CDS (coding DNA sequence) like CpG, TATA regions etc. and (vi) ascertaining the secondary structural details (such as hairpin bends, loops, bulges etc.) of ssDNA and/or RNA sequences *via* nearest-neighbor (NN) energetic profiles [1.8 - 1.10]

- Deducing spectral domain peaks and troughs in the test sequences to determine the underlying spatial patterns [1.12 - 1.14]

- A focused study on viral DNA sequences of pathogens such as B19 [1.18] and dengue viruses

- Exclusive analyses of multiple strains of a given virus (for example dengue virus) [1.19 - 1.22] so as to determine the common subsequences among them by logistically regressing the triple data mined from entropic, energetic and spectral-domain evaluations [1.23]

- Co-relating the details of the sequences from the multiple strains of a given virus (as above) for rational vaccine [1.24] design purposes

Consistent with the objectives enumerated above, exhaustive background details and literature survey are gathered to supplement the research performed. The final objective refers to presenting the results of the research, deducing relevant inferences and listing open-questions for future research.

1.4 Motivation

This work is mainly and objectively motivated by the impetus to develop a multiple set of analyses of genomic sequences for concurrent comparison, logistic compilation (*via* regression) and robust mining of underlying details gathered in terms of different perspectives of analyses pursued.

Traditionally, genomic sequence analysis is performed in three perspectives: (i) By considering the associated entropy features of the test sequence; (ii) by evaluating the nearest-neighborhood interaction of base residues that presents the energetics profile across the sequence in question; and (iii) by elucidating the spectral features of the spatial disposition of sequence content.

Inspired by the aforesaid and established methods, the motivated effort in the present study is to apply cohesively all the above (three) methods on a given sequence and determine the distinguishing details in each case, so that, these details can be collectively compiled to mine the subtle features of the test sequence exhaustively. In other words, it is surmised in the present study that one single version of the analysis may not suffice to portray the variety in gene structures. If one method is effective in showing certain unique details (of the sequence), the other methods may indicate certain other features. It is possible in such multiple analyses that the data acquired may overlap; but, non-overlapping details otherwise will also be obtained.

Hence, the motivated research advocated and described in this dissertation focuses on elucidating subtle feature in the test sequences *via* cohesive data compilation through three methods mentioned earlier. Again, there are different versions of applying the notions of entropy, energetics and Fourier transform methods to genomic sequence analysis. In the present study certain novel concepts are improvised thereof to mine the details effectively and in pragmatic sense. Further, the cohesive data-mining (by the three methods pursued) is applied to a practical, health-related objective of the rational vaccine design. Relevant underlying motive is described below:

Considering the fact that a virus may be present in different guises, it has been a concern in immunological studies to find a common vaccine for the entire set of the serovar. In this regard, notwithstanding the associated wet-lab assertions, bio-informatics can indicate a viable avenue in deducing certain common features between multiple viral strains; and, such common features can be adopted as possible epitopal formats (across the entire set of the strains). Hence, relevant common vaccine design can be attempted based on the epitopal format ascertained.

Motivated by the need to determine the common denominator across genomes of multiple viral strains, this study is extended to identify and apply different bioinformatic algorithms on the test genomic sequences in question. As stated earlier, though sequence analysis is a well-known topic in bioinformatics with feasible approaches and their variants, using appropriate algorithms in the context of genomes of multiple viral strains, as done in this study, is rather sparse.

Further, as discussed earlier, pursuing one method of sequence analysis (applied to all test viral strains) may not possibly and comprehensively identify the prevailing common (and subtle) features between them. For example, perusal of the well-known method of entropy segmentation (information-theoretic based approach) will definitely yield results on any test sequence with significant accuracy (in view of the various related studies that exist in literatures and reviewed in [1.25]). However, whether the deduced information on the test sequence will enable knowing the common features between a set of sequences comprehensively (and accurately) is questionable.

As such, it is a motivated effort exercised in this dissertation to identify more than one independent sequence analyses and apply them on a set of test sequences (such as genomic sequences of the strains of a virus). Accordingly the entropy segmentation method, nearest-neighbor energetics approach and spectral domain analyses are independently applied to the test sequences; and, the results of these three independent studies are logistically combined to determine the existing common subsequence segments in all the genomic structures of multiple viral strains. Therefore, the motivation of knowing a common epitopal structure among the test viral strains is facilitated.

In short, the motivated considerations of the present study can be enumerated as follows:

- To apply cohesively, the three well-known methods of genomic sequence analysis (namely entropic, energetic and spectral-domain algorithms). Absence of such concurrent three-prong approach in practice forms the core motivation in pursuing this research

- Tuning the aforesaid three algorithmic methods and applying them in context of multiple sequences so as to extract the underlying common features: Again existence of an exclusive scheme of combined methods to deduce distinguishable features across a set of sequences is sparse and offers a motivated push for research

- Application of a robust technique envisaged in the context of distinguishing genomic features of multiple strains of a virus is rare: This lacuna has motivated to use the triple approach a above in viral genomic analyses

- Finding a common epitopal attribute among the genes of multiple strains of a given virus is a widely desired vaccine research effort: Relevant considerations offer a push to the study envisaged here

1.5 Contributions: Outcomes of the Research

The salient outcomes of this research can be listed as follows:

- Developing an exclusive framework to apply bioinformatic concepts in the context of viral genomes

- Formulating a methodology using the existing concept of statistical divergence and ascertain the unique genomic structural details of single-stranded DNA of viruses. Hence, obtained are particulars as regard to splice-junctions (canonical and/or cryptic), structural details like hairpin bend, WC matching, bulges, loops etc.

- Invoking the so-called *nearest-neighbor* energetic concepts on nucleotides and apply it to the test viral sequences

- Representing the residues of test sequences *via* numerical chain of associated ionic interaction potential parameter and deduce the signatures of interest using Fourier transform analysis

- Indicating a method to combine the results of the three aforesaid methods and conclude cohesively on highly probable protein details that exist in viral sequences toward vaccine synthesis

- Providing a review on rational vaccine design considerations

- Identifying the scope for future research and enumerate possible open-questions

1.6 Dissertation Organization and Closing Remarks

In order to cohesively address the research efforts and the outcome, this dissertation is written with an organized set of chapters (tentative) as follows:

- **Chapter I: Introduction -** This (present) chapter provides an introduction to the topic of research pursued with the indication of relevant scope and objectives. The dissertation format is outlined

- **Chapter II**: **Genomic Sequence Analysis: A Review-** Chapter II presents a brief review on various methods available to deduce the features and characteristics of genomic sequences of organisms. Starting from the central dogma of microbiology, the genomic structure through the proteomic formation of organisms is briefly discussed and essential sequence features of practical interest are identified

- **Chapter III: Viral Genomic Features –** An outline on the characteristics of the genomes pertinent to viruses is furnished in this chapter. Hence, a review on the family of viral species is presented and the distinctions between them at microbiological levels are detailed. Corresponding single- and double-stranded DNA/RNA structures as well as their unique RNA features in the context of viral family are pointed out. In terms of such unique features, the underlying sequence signatures of importance (such as, finding protein- forming amino acid chains) are identified for the purpose of analytical considerations and computational determinations

- **Chapter IV: Entropy-based Viral Genomic Sequence Analysis-** The objective of this chapter is to analyze viral genomic sequences using information-theoretic methods. Classically, information-theoretic methods (also known as entropy-based techniques) have been adopted in analyzing genomic/proteomic sequences of eukaryotic organisms such as human, yeast, bacteria etc. Relevant studies have also been indicated to ascertain viral genomic details. Notwithstanding the existence of such studies, presented in this work is an effort to determine more aggressive and broader information-theoretic formulations compatible for specific genomes such as those of viruses. The entropy features of a test sequence are reviewed in terms of various information-theoretic considerations and relevant formulations (such as the so-called Kullback-Leibler and other statistical distance/divergence measures). As an example, the CpG motifs are determined using Jensen-Shannon measure. Hence, the concepts of relative entropy, mutual information and information redundancy considerations are revisited.

  Further, genomic sequences have substructures that are not crisply separated. Relevant overlapping features are viewed in terms of fuzzy considerations [1.11]. In this chapter, fuzzy splicing in precursor mRNA sequence is considered and prediction of aberrant splice junction in viral DNA context is indicated.

In addition, the analytical framework and computational details to specify the differential features (not otherwise obviously seen across DNA/ RNA or amino acid sequences) of multiple sequences are described in this chapter. Such differential properties are elucidated *via* Shannon's information redundancy formulation applied to a complex system. Illustrative example and results thereof are furnished with reference to the dengue virus and its serovar.

- **Chapter V: Energetics-based Viral Genomic Sequence Analysis -** The energetics aspect of viral sequences and its implications are discussed in this chapter. The gene structures are known to exhibit exclusive thermodynamic energetics profiles in order to organize themselves into stable structures. That is, the nucleotide alphabets of the DNA, namely {A, T, G, C} arrange themselves in posing a genetic statistics such that, they not only present negentropic details (in Shannon's sense), but also, the associated chemistry (*via* Crick-Watson pairing A↔T and G↔C) renders a minimum global energy profile (at least sub-optimally) across the interacting neighbors. As such, genomic structures assume unique sequence patterns. Deducing the underlying features thereof in viral genomic structures with relevant algorithms (based on energetics consideration) is the topic exercised in this chapter.

  Further, the use of entropy-based segmentation method and energetics method are also addressed side-by-side in this chapter as regard to finding specific structural details of genomes. It refers to characterizing the subregions of genomic sequences such as loops and bulges described earlier in Chapter III. For example, considering the ssDNA of Parvovirus B-19, relevant nucleotide positions wherein the loop, bulge, hairpin etc. are observed are determined *via* statistical measures and nearest-neighbor (NN) based energetics approach. The results due to both methods are compared and discussed.

- **Chapter VI: Fourier Spectral Characteristics of Viral Genomes -** This chapter discusses the spectral characteristics of viral genomic profiles. As indicated in earlier chapters, by virtue of entropy (information-specified) details and energetics-dictated format of genomic structures, the associated residues form characteristic patterns along the stretch of the sequence. In addition, such spatial domain features would also correspondingly reflect another unique set of characteristics in a transformed domain, such as in the Fourier spectrum. Hence, described in this chapter is the avenue to apply Fourier spectral analysis to a set of viral sequences and the results are compiled

- **Chapter VII: A Metalearning Approach to Explore Viral Genomics: A Modular Framework of Data Mining** - In virological context, a particular virus may prevail in different forms of serotypes (as in the case of dengue 1 to 4 viral strains) with common and distinct genomic features. Finding such genomic details of a serogroup is useful in knowing related information for unique vaccine designs compatible for immunity across the viral diversity. For robust comparison of genomes of serovar of a virus in order to decide on their common and differential genomic details, proposed here is a set of sequence analyses exercised side-by-side *via* entropy, energetic and spectral-domain methods. Results obtained thereof with dengue viral serotypes namely, DEN1, DEN2, DEN3 and DEN4 are presented. Hence, inferences on distinct as well as common features extracted are annotated and indicated for possible vaccine design applications

- **Chapter VIII: Viral Genomic Sequences and Vaccine Design Considerations** – This chapter is presented to indicate the major scope of possible uses of bioinformatic analytical frameworks addressed. Specifically, relevant implications are identified *vis-à-vis* vaccine designs. As well known, the gene expression in a virus morphs to different patterns at the molecular (DNA/RNA) level across its different strains. These discernment features offer a viable opportunity to conceive a set of distinct vaccine designs usable to prevent the differentiable pathology likely to be caused by the strains concerned. In this study, it is hypothesized such diverse vaccines can be intelligently synthesized by considering the underlying DNA signature features of the various strains of a given virus. Essentially, the expression seen in each viral DNA/RNA structure as regard to its CDS, CpG, TATA box etc., sites of homology specified by the spatial-spectrum (Fourier domain) details, long-range correlation of coding/noncoding segments and nearest-neighbor energetic-interactions and stability-seeking bends/loop formation (in the case of single-strand DNA or in RNA sequences) are target data that can be profitably utilized in the strain-specific vaccine synthesis. As an example, the computed data on the distinguishable features pertinent to the RNA structures as ascertained by the authors *via* appropriate models are used in proposing a smart vaccine design approach for the Dengue virus having four strains, namely DEN 1, DEN 2, DEN 3 and DEN 4 with distinct RNA features.

- **Chapter IX: Inferential Conclusions and Open-questions for Future Research -** This chapter is written to offer an overview of the dissertation. Also, essential conclusions are enumerated and discussed. Possible research items for future efforts are identified as open- questions

- **Chapter X: Executive Summary**

Thus this introduction chapter is written to outline the overall content of the dissertation and provides details on the scope of the research, underlying objectives and driving motivations. Further, organization of the dissertation is indicated with a format outline on the pursuing chapters.

CHAPTER II

GENOMIC SEQUENCE ANALYSIS: A REVIEW

2.1 General

In living systems, viewed at its most primitive level where meaningful information prevails and deciphered is the DNA molecule. This DNA (or deoxyribonucleic acid) entity is a giant, linear, polymeric molecule existent in all living systems. It consists of two polynucleotide chains wound helically about a long central axis. The bases on opposite chains are joined through Hydrogen bonds with specified constraints.

The linear polymeric bonds of DNA are based on four sub-units, namely, the bases Adenine (A), Guanine (G), Cytosine (C) and Thymine (T), linked in a chain through Phosphate (P) and Deoxyribose (R) bridges. DNA sequences represent very long strings -of the order of 104-109 base-pairs (also known as Watson-Crick (WC) pairs) of (A ⇔ T) and (G ⇔ C).

The occurrence of base in the DNA sequence exhibits statistical ordering. That is, characteristic to every individual living system, random dispositions of bases along the DNA chains are unique and represent a meaningful message borne by statistical uncertainty. That is, a negentropy value can be attributed to the stochastical features of the DNA chain.

This negentropy value denotes specific genetic information possessed by the living system at the DNA level. (For mammals, the genetic information content realised *via* genetic coding is in the order of 1010 bits). This extensively large framework of DNA complex renders the associated entities stochastically viable for representation in the information-theoretic plane.

The process of protein-making (known as translation) follows the central dogma of molecular biology illustrated in Figure 2.1. It involves first the transcribing of information in sections from the DNA strand into an intermediate polymer called messenger RNA (or mRNA), which is similar to DNA except that the sugar residue R is replaced with a slightly different one, namely Ribose R′; and, a base called Uracil (U) replaces the Thymine (T) of the DNA. The genetic information contained originally in the DNA is carried forward when the bases are set in triplet forms (called codons). With the four {A, C, T, G} bases, the possible permuted triplets are 64 (4^3) and they are grouped into 20 amino acids; and, each amino acid being the triplet of bases (or anticodons) bears the mapped genetic code of the DNA for its positioning the peptide chain. The transcription occurs at a specific site on one strand of DNA known as transcription initiation site, marked by a characteristic base sequence. The transcription proceeds through a specific chemical pairing namely, the WC-paring (A ⇔ T/U) and (G ⇔ C) mentioned earlier. The transcription process, in essence, is an information retrieval technique from the original memory units of the DNA. In the next process of translation, the information contained in mRNA constraints the cell in what order amino acids be strung together in making of protein constituents with polypeptide chain.

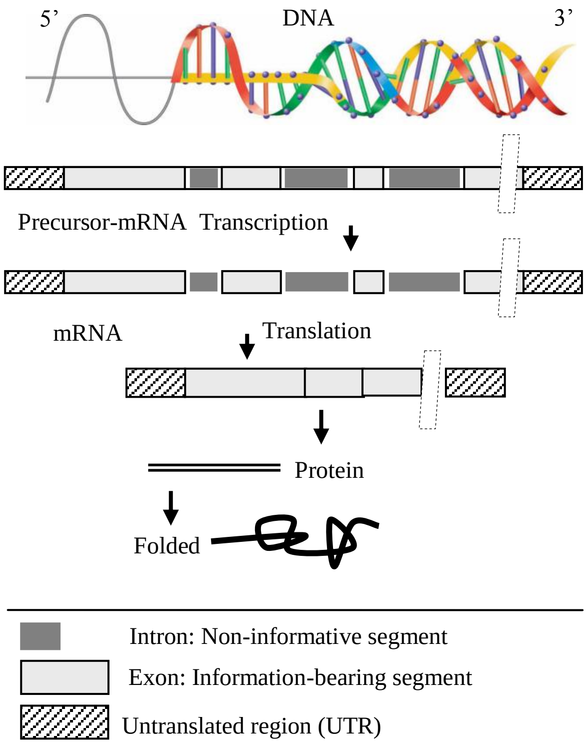

5’
DNA
3’
Precursor-mRNA Transcription
mRNA
Translation
Protein
Folded
Intron: Non-informative segment
Exon: Information-bearing segment
Untranslated region (UTR)

Figure 2.1: Making of a protein: The central dogma of microbiology

The eventual (correct) translation of eukaryotic genomic data into a protein complex is, however, subject to the effects of mutations on the evolutionary conservation tree. Any underlying corruptions may manifest at the so-called splice-junctions that separate/delineate two subsequences in a DNA sequence, namely, the (genetic) information-bearing codon segment (called an exon) and the non-informative “junk” codon, also known as non-codon or intron. (Exons bear necessary information towards protein-making, whereas non-codons are non-informative and their genetic role has not been fully elucidated. Exons and introns appear randomly along the DNA sequence as shown in Figure 2.1. Codons tend to be typically no more than 200 characters long, while noncodons could be tens of thousands of characters in length. Thus in majority, introns prevail mostly in a typical eukaryotic gene.

Towards the process of translation, introns are first scissored out (in the transcription stage) from the sequence and the remaining exons are spliced together constituting the mRNA, which is rendered ready for translation into a protein complex (at the cell interior). Should any errors have occurred (due to mutations), they would give room to the possibility of evolving wrong or cryptic splice-junctions and lead to (imperfect) translations. That is, aberrant splice-junctions may result from mutational spectrum and would hamper the making of correct proteins.

2.2 Biological Sequences Information

Classically, models of biological sequences have been developed based on principles of probability [2.1], [2.2] and the scope of such models has led to various algorithms like the use of Hidden Markov models (HMMs), score-matrices for determining sequence alignment etc. These efforts constitute the basic profile searches of genomic sequences to identify and determine the similarity/dissimilarity between phylogenetically related sequence families. Applying the heuristics of information theory and negentropy considerations (in Shannon’s sense) have been the classical topics of interest in bioinformatics. That is, bioinformatic algorithms in vogue largely rely on information-theoretic notions of Shannon [2.3] in excerpting genomic details and in data-mining exercises pertinent to molecular biology. Most of the models of biological sequence analysis *via* computational techniques are essentially based on the characteristics of the DNAs/RNAs like their structures, chemistry etc. and their associated statistics. In this thesis, a cohesive approach using three perspectives has proposed. They are (i) entropy and information considerations; (ii) energetics-profiles; and (iii) spectral domain characteristics algorithms. These three methods can complement each other in identifying the distinctions among test sequences. Such feature details obtained by a diverse set of analyses could enable distinguishing one DNA sequence from the other robustly in terms of the associated genomic information. It is envisaged that these three algorithms used cohesively will help to determine locally as well as globally the common and contrasting features of phylogenetically related, but, varying strains of a virus.

2.3 Entropy Considerations of Genomic Structures

Relevant to genomic sequence, entropy implies probabilistic (uncertainty) aspects of nucleotides distributed spatially along the sequence length. Inasmuch as the sequence is a mix of codon and non-codon parts, the appearance of the nucleotides {A, T, G, C} will be inherently random, but the Shannon information of the underlying genetic code would remain in the coding sequence segments (CDS). Relevant entropy aspects of coding and non-coding regions in non-viral DNA sequences have been studied, for example, in [2.4 - 2.6]. It is attempted here to apply similar analytical considerations to a viral ssDNA/ssRNA. Finding the delineating border of separation between codon and noncodon regions in a massive stretch of a DNA chain is a bioinformatic problem. Essentially, the focused efforts thereof, refers to entropy characterization of genomic details such as, differentiating informational profile of coding and non-coding regions in a DNA sequence, detecting buried signal (such as splice-junctions) within the DNA, elucidating aberrant (cryptic) attributes introduced in the DNA sequence (as a result of mutational changes), finding fuzzy aspects of overlapping bioinformatic details, knowing the locations of characteristic subsequence (such as CpG islands, TATA box), information redundancy measures (IR) etc. Further, applying the concept of fuzzy logic has shown promising trend in bioinformatics and the door is ajar for more and advanced research.

In bioinformatics, the major interest lies in analyzing genomic and proteomic sequence details. The following algorithmic/computational considerations can be identified thereof is described briefly here (and in details in Chapter IV).

2.3.1 Use of Fisher Discrimination

The Fisher discriminant metric (F-measure) is based on a linear discrimination function with a set of coefficients optimized on the basis of statistical features of a data collection. It can be used as a scoring metric to contrast statistical subsets (possessing relative uncertainty) in a data set. That is, the concept of Fisher discriminant function can be applied to classify (or distinguish) a pair of data sets. Fisher-metric can be applied to the test sequence so as to delineate the codon and noncodon parts. It involves constructing a set of discrimination matrices *vis-à-vis* the subsets of data in the parts to be discriminated. F-metric is useful in identifying the motifs such as TATA-boxes in the DNA sequence.

2.3.2. Use of Complexity Measure

Another approach due to Neelakanta et al. [2.7] uses the concept of information redundancy in complex systems and defines a complexity metric that is adopted to differentiate codon/noncodon segments.

2.3.3. Use of Hamming Distance

In this method the random structure of a DNA base composition is represented by a binary sting and comparison is made between a pair of the sequences *via* modulo-2 (XOR) operation. A test DNA string thereof is compared against a random binary string and the dissimilarity is assessed in terms of the Hamming distance (or the count of 1's in the resulting (XOR)-ed string.)

Suppose **S** depicts a vector representing the binary-coded DNA sequence; **NS** denotes the vector corresponding to the binary-coded complementary DNA sequence; and, **R** is a vector of a random binary sequence. Performing the modulo-2 operations, namely, $\mathbf{S} \oplus \mathbf{R}$ and $\mathbf{NS} \oplus \mathbf{R}$ result in a set of two Hamming distance populations (that is, counts of 1's in each resultant vector). By determining the statistical contrast between these two populations, it leads to an implicit index of comparison of **S** against **NS** (or *vice versa*).

2.3.4. Use of Csiszár Measure

Yet another non-parametric symmetric divergence measure belongs to the class of Csiszár's f-divergences, which are more general than the KL version. Details on such measures are available in [2.8]. Without any loss of generality, all the f-divergences due to Csiszár can be adopted in genomic analyses pursuits. The KL measure is a subset of this Csiszár family of cross-entropy functionals given by: $[p_c \times F(p_c/p_{nc})]$ or $[p_{nc} \times F(p_{nc}/p_c)]$, where F(.) is a doubly differentiable convex function [2.8]. A host of measures thereof can be specified by proper choice of F(.). Popularly, the KL-measure, the Jensen-Shannon (JS) measure etc. have been used in bioinformatics contexts.

2.3.5. Fuzzy Attributes of Genomic Sequences

The large data spaces of bioinformatics can be in many instances overlapping. Such imprecise domains possess unique difficulties when the underlying details are extracted. This is because of the non-specificity of the values and sharplessness of the boundaries of activity variables that can be described mostly in linguistic norms of fuzziness or grey facts. With fuzzy attributes prevailing in bioinformatic contexts, a multi-value logic has to be appropriately built, making the grey truth into complex schemes of formal reasoning.

Exclusive applications of fuzzy logic in bioinformatics are indicated in [2.4] and can be summarised as follows:

- Computing similarity between gene products (annotated via ontology of gene clustering and gene function) can be applied with fuzzy logic to protein secondary structure prediction and the associated structural bioinformatics

- Considering micro-data analyses, the clustering algorithm in fuzzy sense (such as fuzzy co-clustering, fuzzy seaming etc.) can be formulated

- Other application of fuzzy logic in bioinformatics include sequence motif identification, protein sequence alignment, protein sub-cellular localization, prediction, 3D protein structure comparison and computational proteomics

- Building fuzzy thematic clusters and mapping them to higher ranks in taxonomy appears to be a real-world knowledge management in the art of gene ontology.

Arredondo et al. elucidated fuzzy attributes of a DNA complex *via* a fuzzy inference engine designed exclusively to delineate codon and junk codon sub-spaces [2.4]. Recently the authors have developed a technique for fuzzy splicing in pre-cursor mRNA sequence. Hence, aberrant splice junctions in viral DNA contexts are predicted [2.9].

2.4 Energetics Aspects of Genomic Sequences

The chemistry of the bases in a nucleotide chain implicates neighborhood energy-level dependency. As such, the nucleotide bases in their locations exhibiting neighborhood energy profile across the sequence can also be considered to analyze the underlying genomic features. The RNA (transcribed from the double-stranded DNA) as well as in the so-called single-stranded DNA (ssDNA) that exist in viral genes, sequences tend to become more compact by “folding” or “bending” themselves into a stabilized hairpin structure (mostly towards 3’ end) via nucleotide base-matching set by Watson Crick (WC) pairing of A ↔ T and G ↔ C [2.10]. In addition to the favored (neg)entropy enabled by WC-pairing toward stability of hairpin structures formed, relevant (stability) dynamics also relies on free-energy minimization specified by nearest-neighbor (NN) parametric attributes [2.11], [2.12] of base-pairs in the test sequence. That is, the stability in question conforms to the rules stipulated by each base-pair depending only on the most adjacent pairs, with the associated total free-energy being the sum of each contribution of the neighbors. The underlying considerations are as follows:

- Known generally as individual nearest neighbor (INN) model, it implies a preferential stacking of energetically conducive pairs with loop-initiation leading to an eventual hairpin structure

- The free-energy increments of the base-pairs in the sequence can be counted as stacks of adjacent pairs. For example, the consecutive CG base-pairs are worth about (~3.3 kcal/mol). The loop-region formed normally has unfavorable increments called loop initiation energy that largely reflects an entropic cost expended in constraining the nucleotides within the loop. For example, the hair-pin loop made of four nucleotides may have an initiation of energy as high as + 5.6 kcal/mol.

In using the relative free-energy data-set of NN specifications, a sliding-window method can be invoked to get the profile of free-energy variation across the test sequence. Relevant information can be profitably utilized in identifying characteristic sub-spaces in the DNA/RNA such as stem and loop features, CpG island motifs and isolation of CDS parts in a typical viral DNA structure of B19 virus; and, relevant considerations are exclusively useful in conceiving rational vaccine designs [2.13], [2.14]. Relevant algorithmic heuristics are as follows: The disposition of bases adjacent to each other or in *nearest-neighbor* (NN) sense is consistent with a minimum energy profile. This attribute assures a thermodynamic stability decided by chemical bonding consideration. Relevant INN -model has been proposed and discussed in the literature, largely in predicting RNA secondary structures [2.11], [2.10]. Apart from RNA sequences, relevant stability considerations also prevail in the case of single-stranded DNA (ssDNA) that exists in certain pathogens. The underlying structural stability of ssDNA sequences is specified by Watson-Crick (WC) pairing achieved with the sequence morphology of loops and hairpin bends (wherein the energy profile across adjacent neighbor seek a global minimum). Characterizing such unique profiles in such ssDNA contexts has been indicated in [2.13] *via* energetics-based analyses. Detailed procedures and results of the above method has been discussed in Chapter V.

2.5 Spatial-domain Analysis of genomic Details

It is largely concerned with still sparsely known dependence between nearby bases and their occurrence statistics across the genomic sequence (in Markov's sense). Hence, it is argued that the Fourier-transform (FT) may be adopted to overcome the aforesaid obstacle considering the fact that, real and imaginary parts of the Fourier coefficients are all independent random variables and as such, they may yield two distinct sets of fortifying details on the associated statistics [2.15 - 2.17]. Spectrograms are powerful visual tools for biomolecular sequence analysis [2.18]. Defining a spectrogram for use in analyzing 2D-patterns, the display of the magnitude of can be realized *via* short-time Fourier transform (STFT) [2.19]. In this thesis, the FT-based algorithmic and computational efforts pursued in the present study essentially follow the procedure due to [2.20] and has been discussed in details in Chapter VI.

2.6 Conclusion

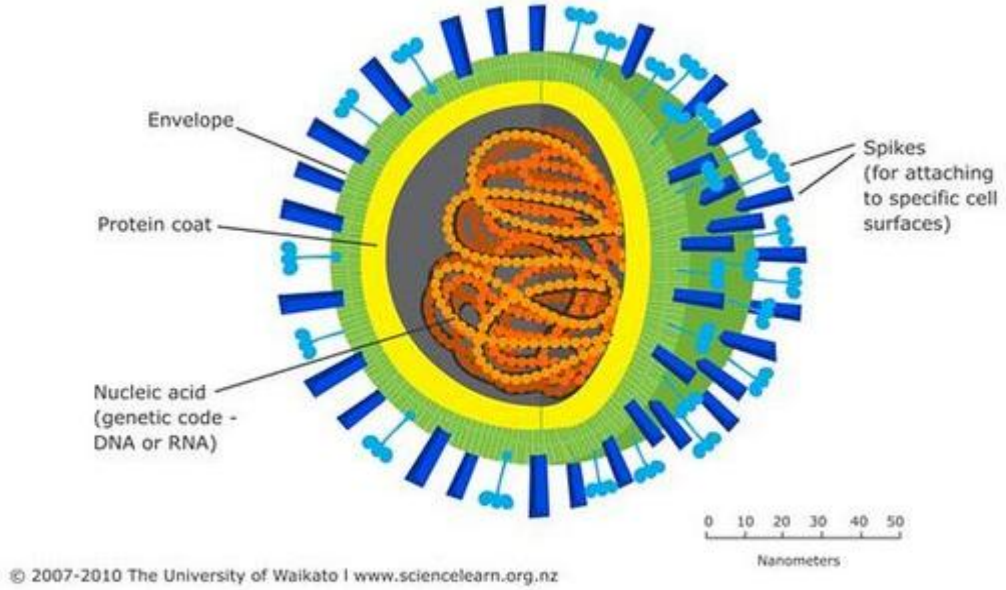
The scope of this chapter is to provide background details of biological sequences and lay the foundation for information-theoretic, energetics and spatial-spectral domain method for analysis of genomic sequences. These methods have been cohesively exercised on genomic sequences to obtain results on genomic features like characteristic loops and bends, delineation of exon-intron boundary, CpG island etc. and the results have been furnished in the ensuing chapters.

CHAPTER III

VIRAL GENOMIC FEATURES: A REVIEW

3.1 General

Viral particles are the most abundant biological entities present on the earth [3.1]. The term virus is derived from the Latin word *virus,* which means poison. Viruses are non-living, microscopic particles consisting of either a RNA or DNA genome surrounded by a protective, virus-coded protein coat. The genomic material of the virus is packaged inside a structural capsid protein. In enveloped viruses, this structure is surrounded by a lipid bilayer with an outer layer of virus envelope glycoproteins [3.2] as shown in Figure 3.1. Viruses are considered non-living as they do not possess the most basic characteristics of living system (metabolism, growth, reproduction and reaction to stimuli) [3.3], [3.4]. For replication, viruses depend on the host cell. In fact, before the viruses enter into the host, they are called as “virions” [3.5]. In his landmark paper, Bândea [3.5] distinguishes virions and viruses, aptly explaining the difference as, virions being the "spores" or reproductive forms of the virus, possessing life only as a potential property. Thus, virions are packaged genetic materials, which can be passed through direct contact or carrier to the host, where it replicates.



© 2007-2010 The University of Waikato I www.sciencelearn.org.nz

Figure 3.1: Structure of virus

Almost all life forms, including all forms of fungi, bacteria, plants and animals can host atleast some type of virus. Also, there exist different types of viruses, which have different types of genetic material, structure, morphology etc. The replication mechanisms of the viruses also differ widely depending on the type of the virus and the host. Roizman [3.6] has described the mechanism of replication of various types of viruses in a range of hosts. Upon entering into a host, the virus may stay in vegetative state (resulting into latent infection, meaning they not be able to replicate at all in the host) or start infesting the cell of the host immediately, depending on host ambient.

Since, the viruses depict the most abundant biological entity; their classification and nomenclature are significant in biological contexts. The first system of classification of viruses was suggested by Holmes in 1942 [3.7]. He suggested Linnaean system of binomial nomenclature to classify viruses into three groups under one order, *Virales*.

Accordingly, viruses are classified depending on the type of living organism (fungi, bacteria, plants or animals) that they infected. In 1962, Lwoff e*t. al.* [3.8] suggested classifying viruses using complete Linnaean hierarchical system for viral nomenclature (unlike binomial classification suggested by Holmes) based on their size, symmetry, nucleic acid, physico-chemical properties and presence or absence of envelope around the virus [3.8]. Inasmuch as a single virus can infect multiple species of organisms, Lwoff e*t. al.* ruled out the classification of virus based on the types of cells they infected. The outline laid down in [3.8] was accepted by *International Committee on Taxonomy of Viruses (ICTV)* [3.9] as the standard method for nomenclature of virus with few changes and additions.

A classification system different from the one suggested by Lwoff *at. al.*, was developed by David Baltimore in 1971 [3.10]. He argued that due to the small size of the viruses, it is difficult to identify their shapes even under electron microscope. As such, he suggested classifying the viruses according to their genome type (namely, type of nucleic acid (DNA or RNA) and its structure (linear, circular or segmented)) and on the method of viral mRNA synthesis. If the viruses are classified into categories as per these characteristics, then, all viruses in a given category will all behave in a similar way. This suggestion was accepted by ICTV and has been included in classifying viruses along with the framework laid down by Lwoff *at. al.* In all, viruses are categorized into seven different types as specified in Table 3.1 below:

Table 3.1: Baltimore classification of viruses

Group	Type of genome	Common examples of family affecting human
DNA virus		
I	dsDNA	Adenovirus, Papillomaviridae (HPV1, HPV11, HPV16, HPV18. HPV 16 and 18 can become cancerous), Herpesviridae (HHV1-8 causes diseases like Herpes, Roseola, Epstein Barr, Chickenpox etc.) Poxviridae (smallpox), Hepadnaviridae (partially ds-causes Hepatitis B), etc.
II	ssDNA	Parvovirus B19V (causing fifth disease in children), Anelloviridae etc.
RNA virus		
III	dsRNA	Rotavirus etc.
IV	(+)ssRNA	Coronaviridae (SARS), Picornaviridae (Polio virus, common cold virus, Hepatitis A virus etc.), Hepevirus (Hepatitis E virus), Togaviridae (Rubella virus, Ross River virus, Sindbis virus, Chikungunya virus etc.), Flaviviridae (Yellow fever virus, West Nile virus, Hepatitis C virus, Dengue fever virus etc.) etc.
V	(−)ssRNA	Filoviridae (Ebola virus, Marburg virus), Paramyxoviridae (Measles virus, Mumps virus), Rhabdoviridae (Rabies virus), Orthomyxoviridae – (Influenza viruses), Bunyaviridae (Hantavirus, Crimean-Congo hemorrhagic fever) etc.
Reverse transcribing virus		
VI	ssRNA-RT	Retroviridae (HIV)
VII	dsDNA-RT	Hepadnaviridae (Hepatitis B)

3.2 Types of Viruses as per Baltimore Classification: Details

3.2.1 DNA Virus

These viruses have DNA as its genomic material and affect almost all domains of life. Double-stranded DNA (dsDNA) has two strands of DNA and replicate inside the host cell using its DNA polymerase [3. 10]. Double-stranded DNA has linear or circular genome and follows the central dogma of molecular biology, except that it uses the host cell's machinery for multiplying. Most of the DNA viruses are dsDNA viruses. Single-stranded DNA (ssDNA) has a single strand of DNA and requires formation of an intermediate double-stranded DNA form for genome replication. This is achieved by assuming structures like hairpin bends or loops (as has been discussed in Section 3.3 with reference to Parvovirus B19V).

3.2.2 RNA Virus

As the name suggests, the main genomic material of RNA virus is ribonucleic acid (RNA). Many of the deadly and widespread diseases in the world are spread by RNA virus, some of which are mentioned in Table 3.1. Most of the RNA viruses are single stranded (ssRNA). The ssRNA virus can be further classified into +ve sense and –ve sense [3.10]. If the RNA base sequence is identical to the viral mRNA sequence, then, it is known as +ve sense RNA virus; whereas, if the RNA base sequence is complementary to the viral mRNA sequence, it is known as –ve sense RNA virus. The double-stranded RNA virus (dsRNA) viruses represent a diverse group of viruses with varying characteristics. It has been discussed exhaustively in [3.11]. The RNA viruses are genetically very unstable with a very high rate of mutation. One of the principal reasons for this is the lack of the corrective mechanisms available inherently in DNA molecules [3.12]. This high rate of mutation greatly limits the design of vaccine and its future effectiveness. In fact, it is suggested by Holland *et. al.* [3.13] that RNA viruses have the most important role in evolution and survival of different species and maintenance of diverse ecology in general. The −ve sense RNA viruses can also be divided into two groups: i) Viruses containing non-segmented genomes for which the first step in replication is transcription from the -ve stranded genome by the viral RNA-dependent RNA polymerase yielding monocistronic mRNAs that code for various viral proteins. A +ve sense genome copy is then produced that serves as template for production of the -ve strand genome. Replication occurs within the cytoplasm. ii) Viruses with segmented genomes for which replication occurs in the nucleus; and correspondingly the viral RNA-dependent RNA polymerase produces monocistronic mRNAs from each genome segment. The largest difference between the two groups of –ve sense RNA viruses is the location of replication site [3.14]. Details on replication of RNA viruses can be seen in [3.15].

3.2.3 Reverse Transcribing Virus

This category consists of some of the most deadly as well as useful viruses. The economic impact of retrovirus (Group VI) is enormous. Infection by HIV, a member of retrovirus (Group VI reverse transcribing virus) is considered pandemic in the modern world. Most untreated people infected with HIV-1 eventually develop AIDS. Also, retroviruses play a role in some forms of cancer. However, by careful design, retroviruses can also be used as the vectors for gene therapy and gene delivery systems [3.16].

In most viruses, DNA is transcribed into RNA. Further, RNA is translated into protein. But, in retrovirus, RNA is reverse-transcribed into DNA, which is integrated into the host cell's genome and then undergoes the usual transcription and translational processes to express the genes carried by the virus. So, the information contained in a retroviral gene is used to generate the corresponding protein *via* the sequence: RNA → DNA → RNA → protein. This extends the fundamental process identified by Crick and

Watson, in which the sequence is: DNA → RNA → protein [3.17]. The term "retro" in retrovirus in fact refers to this reversal (making DNA from RNA) of the central dogma of molecular biology. Like other RNA viruses, retrovirus also mutates quickly and enormously, thus making it difficult to produce effective antiretrovirus drugs that can stay effective for some time or producing a vaccine against these viruses.

Hepadnaviruses (Group VII viruses) are very small genomes consisting of partially double-stranded and partially single-stranded circular DNA. The genome consists of two uneven strands of DNA. One has a negative-sense orientation, and the other (shorter) strand has a positive-sense orientation [3.18]. For multiplication, the reverse transcriptase process described above [3.19] is pursed. These viruses cause diseases of the liver. The mechanism of replication of Hepadnaviruses is discussed in [3.20].

For the present research, virus mainly considered is: (i) Parvovirus B19V and (ii) Dengue virus. The details of which are presented in the following sections.

3.3 Structural Details of Parvovirus B19V

B19V parvovirus is a member of the family parvoviridae responsible in causing a variety of diseases in human. Further, B19V is a member of the genus erythrovirus in the family of parvoviridae. Structurally, B19V contains a small linear single-stranded DNA (ssDNA) genome of 5.6kb length, which harbors two identical *inverted terminal repeats (ITR)* that serve as origin of DNA replication in the host cell. The virus has framework of overall folding of the DNA structure into hairpin formats. Such hairpins formed from a ssDNA consist of a base-paired stem-structure and a loop sequence with unpaired or mismatched nucleotides as shown in Figure 3.2. Relevant conformational studies of DNA hairpins indicate possible tri-dimensional forms with variations, suggesting high profile of complexity of ssDNA structures. Part of this complexity can however be simplified with appropriately rationalized bioinformatic description of loop-bases and the stem part in the frame of backbone structures.

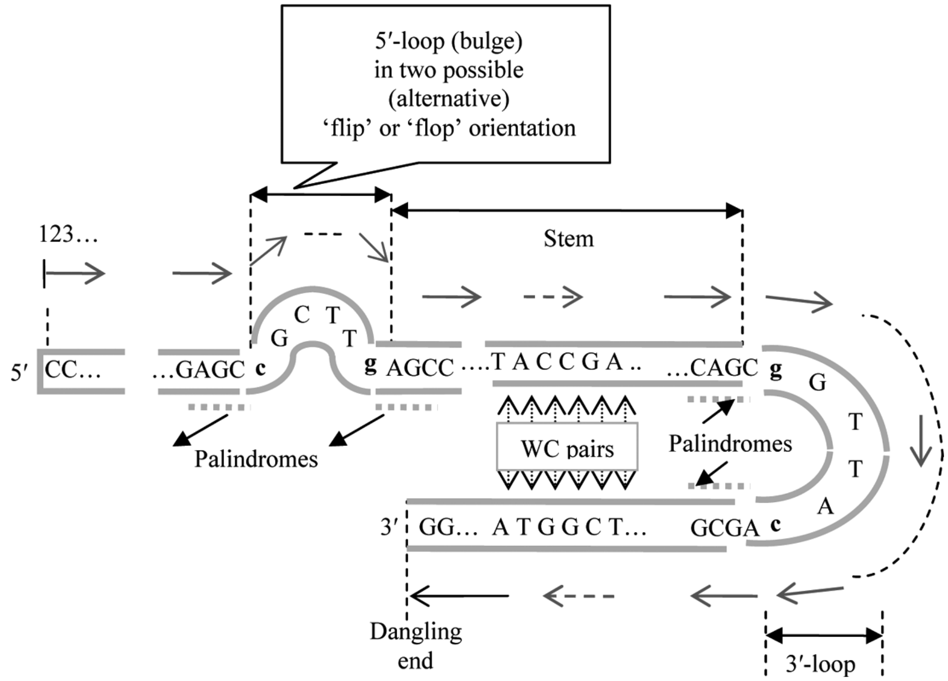
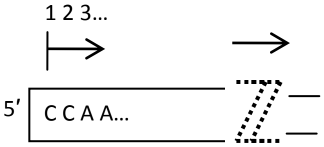


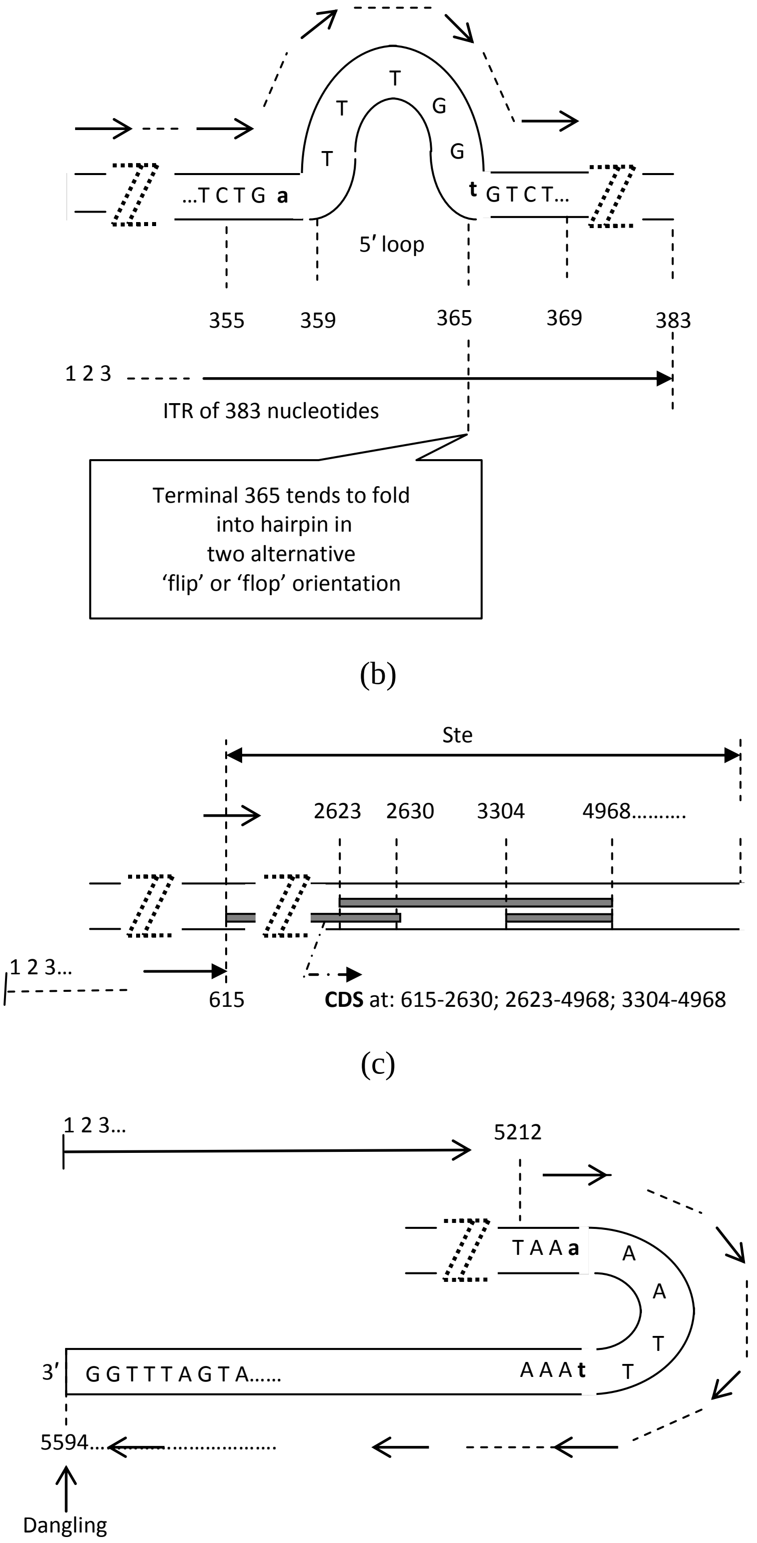
Figure 3.2: A typical hairpin folding of a ssDNA genome.

The B19V virus, considered in this study, according to Baltimore classification has an ssDNA (+sense) genome structure and it is classified as a parvovirus. Normally, the parvovirus has a non-segmented linear ssDNA genome, with an average genome size of 5kbps (5594 nucleotides in length in the case of B-19) with a short double-stranded hairpin formation at the 3'-end. Characteristically, the 3'-hairpin may have inverted repeats across its stem region with site-specific nicks. The microbiological description of such DNA hairpin bends offers a distinct scope of study. Apart from understanding the biological importance of DNA hairpin bends, a parallel pursuit of research is also directed at the physics of thermodynamics on the structural aspects of folded ssDNA as reported in Hilbers [3.21].

Typically, inverted terminal repeats (ITR) can be observed as illustrated in Figure 3.3 (a). For example, the B19V has an ITR of 383 nucleotides and of these nucleotide bases, the terminal 365 tend to fold into hairpins in two alternative ‘flip’ or ‘flop’ orientation. The identical ITRs being present at each end of the genomes correspond to unpaired or mismatched bases in the palindromes represented by the bulges or bubbles. The hairpin folding not only enforces the stability, but, also solves the problem of linear DNA molecules replicating their 5' ends due to the requirement of DNA polymerase for a primer with a free 3'-OH group [3.22]. The hairpin transfer mechanism solves this issue by relying on terminal palindromic sequences to foldback on themselves, forming hairpin structures that prime the DNA replication.



(a)



(b)

(c)

(d)

Figure 3.3 An expanded view of B19V viral structure of bases from 1 through 5594 with

details such as, a bulge at 5′-end, the 3′-end loop (hairpin-bend), the stem parts, ITRs and locations of *coding DNA sequences* (CDS). Relevant details available as GenBank data on human parvovirus B19V (NC_000883.1) (Websites, 2011)

(a) The start of 5′-end of the genomic sequence of B19V.
(b) The stretch of 5′-end of 383 nucleotides tends to fold into a bulge around 365th base. The sets, {TCTG**a**} and {**t**GTCT} on either side of the bulge constitute the *inverted terminal repeats* ITRs) of palindromes. The bases (**a** and **t**) shown in lower case bold fonts depict the *closing pairs* in the loop. (The bases **a**TTTGG**t** in the bulge can flip-flop to a complement set of bases, namely **t**AAACC**a**; that is, the bulge can format itself in two alternative ‘flip’ or ‘flop’ orientations).
(c) The stem-part of nucleotides stretches up to 5212th base and contains CDS at: 615-2630; 2623-4968; 3304-4968
(d) Towards the 3′-end of the genomic sequence, a hairpin structure (3′-loop) is formed with a bending and a reversed stem-part that ends at the last nucleotide, 5594 (called “the dangling end”). The sets {TAA**a**} and {**t**AAA} on either side of this 3′-end loop constitute the palindromes of inverted repeats. (Again, the set **a**AATT**t** in the loop can flip-flop to its complement, **t**TTAA**a**)

In addition to basic aspects of hairpin structure of a viral DNA, mismatched nucleotides in the 5' terminal hairpin can be observed. Further, more complex information may prevail with wild type palindrome in viral DNA sequences [3.23]. The complexity of hairpin structure as above is constrained by the associated molecular dynamics involving characteristic mismatches (such as tandem G-T mismatches at the stem and across non-canonical base pairs.) [3.24]

Learning about the profile of hairpin structures is pertinent in the context of: (i) Understanding the virus replication process [3.23] and (ii) in drug synthesis applications, where, a relevant compound is sought, which in a specific DNA acts as a binding agent and inhibit the replication of certain viruses. [3.24]

Like functional RNAs, which are intensely folded, constituting stable, compact structures, viral ssDNA also assume a hairpin-bending (mostly towards 3’ end). This happens to preserve the stability of the sequence in the single-stand format. In the folded structure, the nucleotide base pairing (also known as Watson Crick (WC) pairing) of A↔T and G↔C takes place around the hairpin region (at the 3’ end), as illustrated in Figure 3.3 (b).

The sequence of bases in the viral ssDNA contains the chemical signature necessary for storing and expressing genetic information. As well known, in the double-stranded helical form of the DNA, the base sequence (5'-3') in one strand matches (A↔T and G↔C) with the complementary base sequence (3'-5') as shown in Figure 3.3. The associated chemical configuration allows the double-stranded helical structure of the DNA to be structurally stable with the underlying features of energetics. The self-complementing feature of the double helix does not however prevail in the single-stranded format of any DNA structure such as in the viral ssDNA. As such single strand versions tend to “fold” or bend themselves so as to get stabilized *via* feasible base-pair matching. Relevant annotated details of the proteins made by the virus are presented in Table 3.2 below:

Table 3.2: CDS range (from NIH website) of Parvovirus B19V [3. 25]

Name of Protein	**Range of bases** **(NIH accession NC_000883.2)**
non-structural protein (NS1)	616…2631
7.5 kDa protein	2084...2308
minor capsid protein	2624...4969

protein X	2874..3119
major capsid protein	3305..4969
11 kDa protein	4890..5174

3.4 Structural Details of Dengue Virus

Dengue virus is a small (approximately 10.7 kb) positive-sense, single-stranded RNA virus (ss-RNA). It is another virus considered in the present research. Dengue belongs to the family Flaviviridae and has inverted complementary sequences at the ends of the molecule that mediate long-range RNA-RNA interaction and genome cyclization. Studies have demonstrated that alternative conformations of the genome are necessary for infectivity. Dengue virus has four different serotypes or strains, namely DEN1, DEN2, DEN3 and DEN4. The complete sequences of all the four strains are available in NCBI databank [3.26 – 3.29]. The genome of dengue virus encodes three structural proteins that form the coat of the virus and deliver the RNA to target cells, and seven nonstructural proteins that are responsible for the production of new viruses once the virus gets access to the host cell. These four different strains have slightly varying genomic characteristics (exhibits almost 60-80% homology between different strains) [3.30], resulting in slightly different proteins. One of the principle obstacles in developing an effective dengue vaccine is due to the need of simultaneously stimulating the immune system against all the four strains and thus generating antibodies against all the four different forms of the dengue virus. This is because when a person is infected with one serotype of the virus, and then infected later by a second serotype, the antibodies and immunity, gained from the first infection appear to assist with the infection by the second subtype, instead of providing a general immunity to all serotypes. This means that an effective vaccine will have to stimulate protective antibodies against all four types at once, a feat that has not yet been achieved. Studies on nucleotide divergence characteristics among different strains of the given virus are limited. Such lack of basic information of viral diversity severely limits vaccine and anti-viral therapy development efforts.

The first dengue viruses were isolated from sick soldiers in Calcutta (India), New Guinea, and Hawaii. The viruses from India, Hawaii, and one strain from New Guinea were antigenically similar, whereas three other strains from New Guinea appeared to be different. They were called dengue 1 (DEN1) and dengue 2 (DEN2) and designated as prototype viruses (DEN1, Hawaii and DEN2, New Guinea-C) [3.31]. Since then, outbreaks of dengue fever (DF) and dengue hemorrhagic fever (DHF) have taken place in hundreds of countries, spread by mosquitoes especially in tropical and subtropical areas, resulting into thousands of deaths.

The genome of dengue virus encodes a single long open reading-frame (ORF), flanked by highly structured 5' and 3' untranslated regions (UTRs). After entering the host cell *via* receptor mediated endocytosis, the virus releases the genomic RNA into the cytoplasm, which serves as mRNA for translation. Like all other Flaviviruses, the mRNA is translated as a single polypeptide and then cleaved into constituent proteins. Dengue virus has stem loop structure at both 5' and 3' untranslated region. These UTRs are needed for the stability and functioning of the viral RNA. Dengue virus 5' UTRs are between 95 to 101 nucleotides long. They contain two RNA domains with distinct functions during viral RNA synthesis. The first domain consists of approximately 70 nucleotides and is predicted to fold into a large stem-loop (SLA), a common feature found in all viruses included in the Flavivirus genus. The second domain of the dengue virus 5' UTR is predicted to form a short stem loop (SLB), which contains essential sequences for long-range RNA-RNA interaction and replication [3.32]. Within these two stem loops are found some loops and bulges.

The first protein to be formed is the capsid protein. It is one of the most important proteins as it contains a hairpin bend between two AUG start codons. This structure is absolutely necessary for efficient viral replication in human and mosquito cells [3.33].

The 3' UTR is approximately 450 nucleotides long and can be divided into three domains. Domain I is located immediately after the stop codon of NS5 and is the most variable region within the viral 3' UTR. It exhibits extensive size variation between serotypes; it can be from more than 120 nucleotides to less than 50 nucleotides. Domain II comprises of many hairpin structures. Particularly interesting is a characteristic dumbbell structure containing conserved sequences with several pseudoknot structures. Domain III of the 3' UTR consists of a conserved sequence which is involved in a long-range RNA-RNA interaction between the ends of the viral genome, followed by a terminal stem-loop structure. A conserved feature of all strains of dengue virus and other flavivirus genomes is the presence of inverted complementary sequences at the ends of the RNA that mediate long-range RNA-RNA interactions.

The annotated details as available in literature [3.26 - 3.29] as regard to the four strains of dengue virus under discussion are summarized in Table 3.3:

Table 3.3: CDS range (from NIH website) [3.26 - 3.29]

Name of Protein	DEN1 (NIH accession NC_001477)	DEN2 (NIH accession NC_001474)	DEN3 (NIH accession NC_001475)	DEN4 (NIH accession NC_002640)
Capsid protein	94….394	97..396	95….394	102…398
Anchored capsid protein	94….436	97…. 438	95….436	102…440
Membrane glycoprotein	710…934	712… 936	710…934	714…938
Membrane glycoprotein precursor	437...934	439… 936	439…934	441…938
Envelope protein	935...2419	937… 2421	935…2413	939…2423
Nonstructural protein 1	2420...3475	2422.. 3477	2414...3469	2424..3479
Nonstructural protein 2a	3476...4129	3478... 4131	3470..4123	3480..4133
Nonstructural protein 2b	4130...4519	4132... 4521	4124..4513	4134..4523
Nonstructural protein 3	4520...6376	4522.. 6375	4514..6370	4524..6377
Nonstructural protein 4a	6377...6757	6376... 6756	6371..6751	6378..6758
2k protein	6758...6826	6757... 6825	6752..6820	6759...6827
Nonstructural protein 4b	6827...7573	6826.. 7569	6821..7564	6828..7562
Nonstructural protein 5	7574..10270	7570… 10262	7565..10264	7563..10262

3.5 Viruses as a Challenge for Design of Vaccine

Most infectious diseases are caused by pathogens that are reasonably genetically stable and host-specific such that a single widely administered vaccine can be used to effectively prevent widespread disease and especially epidemics [3.35]. However, an enormous mutating variety of genomic structures can be seen among viral species, specially ssDNA viruses and RNA viruses. Among RNA viruses, the genome is often divided up into separate parts within the virion (*segmented)*. Each segment often codes for one protein and they are usually found together in one capsid. Every segment is not required to be in the same virion for the overall virus to be infectious. Antigenic diversity among ribonucleic acid (RNA) viruses occurs as a result of rapid mutation during replication, short replication times and recombination/reassortment between genetic material of related strains during co-infections [3.36]. Hence, effective vaccination against such unstable and rapidly mutating viruses requires surveillance programs to monitor circulating serotypes and their evolution to ensure that vaccine strains match field viruses [3.37].

3.6 Conclusion

In this chapter, the general structure of viruses and their classification has been summarized to indicate the diversity in the different families of viruses and the diseases caused by some of the members of the families. Particularly broad outlines on the genomic features, replication, life cycle and the dynamic structures of the genome of the viruses have been discussed. In the present research, details of single-stranded DNA and RNA virus has been studied by bioinformatic methods by considering the genomic sequences of parvovirus B19V and the four serotypes of dengue virus family. Finally, the principal obstacle in designing anti-viral vaccine has been discussed briefly.

CHAPTER IV

ENTROPY-BASED VIRAL GENOMIC SEQUENCE ANALYSIS

4.1 General

The concept of entropy seen in the perspectives of Shannon's information is described in this chapter to present the underlying details statistically strewn across genomic sequences. Relevantly, the motive of this chapter is specified to frame an objective towards analyzing genomic sequences using information-theoretic methods and related entropy concepts.

Classically, information-theoretic methods (also known as entropy-based techniques) have been adopted in analyzing genomic/proteomic sequences of eukaryotic organisms such as human, yeast, bacteria, virus etc. The entropy features of a test sequence can be described in terms of various information-theoretic considerations and relevant formulations (such as the so-called Kullback-Leibler and other statistical distance/divergence measures) [4.1] and [4.2]. Hence, the concepts of relative entropy, mutual information and information redundancy considerations are invoked in this study and the use of relative entropy and mutual information formulations in the context of genomic analysis is elaborated with necessary examples in subsequent chapters.

Exclusively addressed in this chapter are two topics pertinent to genomic entropy information-theoretics: (i) fuzzy aspects of genomic entropic detail transitions; (ii) information redundancy profile of genomic entities. In information-theoretic framework, comparison of two statistical profiles can also be accomplished *via* what is known as *information redundancy* (IR). Relevant analytical framework and computational details to specify differential features (not otherwise obviously seen across DNA/ RNA or amino acid sequences) of multiple sequences using IR concept are described in this chapter. An illustrative example and results thereof are furnished with reference to the genomics of dengue virus and its serovar; and (iii) recognizing the presence and extent of the so-called CpG sequences in the test ssDNA genome.

4.2 Entropy and Information: An Overview

Information theory (IT) is a probability-based concept developed on the basis of principles of entropy. Objectively, it serves to evaluate the contents and the nature of information buried in messages and it refers to rationally perceiving meaningful information from a set of data *via* stochastical considerations.

Classically, entropy is a thermodynamic concept and the "thermodynamic information" decides the number of choices or alternatives posed by the uncertainty (or entropy) involved as a result of order-disorder conflict existing (naturally) in a system. Order in a system presents a negentropy as regard to the certainty of details generated, stored, transcribed, and copied (retrieved). It "informs" the system of details to organise towards an objective function. Hence, the negentropic details constitute "information". In contrast, any associated disorder would try to off-set the system's objective and therefore, constitutes a posentropy. The order-disorder conflict always prevails in a system constituting a “thermodynamic tug-of-war”, which can be studied in the suites of entropy concepts of IT framework [4.1].

A “Siamese twins” relation prevails between entropy and energy, which can be specified *via* entropy *versus* energy equivalence of (the bits of) information. That is, if E is the energy consistent with the thermodynamic indicator namely, the temperature, T, then the Boltzmann’s energy relation indicates that,

$$E = (K_B T)\log_e(N) \text{ erg} \qquad (4.1)$$

where, N is the number of choices in the disordered state and k_B denotes the Boltzmann constant (14×10^{-16} ergs per degree). For example, at $T = 298^o$ K and N = 2 (binary state), $E = 3 \times 10^{-14}$ ergs.

The uncertainty in a disordered system can be estimated by considering the statistical probabilities of the states involved. Correspondingly, the measure of information (H) or negentropy (S) is functionally dependent on those probabilities leading to the well-known Shannon’s law:

$$H = S = \sum p_i \log_2(p_i) \text{ bits} \qquad (4.2)$$

where p_i’s represent the relative numbers or the probability of occurrence of choices/states.

Further, entropy (S) links the transforming (but conserved) energy and information, consistent with the change of state(s) involved. It is specified by the Boltzmann's entropy relation, namely,

$$S = K_B \times [\log_2(\Omega)] \text{ bits} \tag{4.3}$$

where, the Boltzmann constant k_B links the thermoentropy (or temperature) associated with the (thermal) energy with the various states involved as per equation [4.1] and explicitly, $E = k_BT$, known as Boltzmann equation. Also, the function $\log_2(\Omega)$ describes changes encountered in a system having a constant (conserved) energy and mass.

Applications of information theory include communications, computer science, genomics, economics, linguistics and others. The most fundamental problem of any type of communication is reproducing a message exactly or approximately at the receiving end. As indicated in [4.1], if the number of possible received messages is finite, then, the number of possible messages can be thought of as a measure of the information produced when one chosen message is received. For example, natural language provides a system that generates long sequences of symbols that can be considered as realizations of different random processes [4.3]. Then the entropy measure has been used to indicate how much information is produced on average by each alphabet of a language and the redundancy in a language implies how much repetitiveness is imposed on the language by its statistical characteristics.

Relevant entropy aspects and profiles of coding and non-coding regions in non-viral DNA sequences have been studied, for example, in [4.4] and [4.5]. It is attempted here to apply similar analytical considerations to a viral ssDNA and ssRNA.

4.3 Entropy Considerations in Bioinformatics

Biological systems are intrinsically information-rich and the growth process involves consumption of nutrients, emergence of new cells and excretion of waste products. This process is accompanied by heat exchange to and from the reservoir (or stored energy) such that the total entropy change is always positive consistent with the Second Law of Thermodynamics. Shannon's entropy concepts and Boltzmann's perspectives on thermodynamic entropy are cohesively addressed in [4.1] and [4.6].

In living systems, viewed at its most primitive level where meaningful information prevails and deciphered is the DNA/RNA molecule. The occurrence of base in the DNA/RNA sequence exhibits statistical ordering. That is, characteristic to every individual living system, random dispositions of bases along the DNA/RNA chain are unique and represent a meaningful message borne by statistical uncertainty. That is, a negentropy value can be attributed to the stochastical features of the DNA/RNA chain. This negentropy value denotes specific genetic information possessed by the living system at the DNA/RNA level. (For mammals, the genetic information content realised *via* genetic coding is in the order of 1010 bits). This extensively large framework of DNA complex renders the associated entities stochastically viable for representation in the information-theoretic plane.

Characteristically, it is well known that gene information is contained in the so-called exon segment of a nucleotide sequence, wherein the associated genetic information (in Shannon's sense) is decided by the statistics of {A, T, G, C} resulting from the permuted set of bases in the exons. Concurrently existing along the genomic sequence are introns, which supposedly bear no genetic details useful towards protein encoding. Towards the process of translation (a protocol in the central dogma of microbiology), introns are first scissored out (in the transcription stage) from the sequence and the remaining exons are spliced together constituting the mRNA, which is rendered ready for translation into a protein complex (at the cell interior). The contents of exons (or introns) can further be specified by a permuted set of 64 (= 4^3) triplets of {A, T, G, C}, which are grouped into 20 amino acids (AAs). The non-uniform probability of occurrence of such triplets in the exons support the uncertainty (entropy) aspects of the genomic framework furnishing (in Shannon's sense), the genetic information; whereas, in the case of introns, the associated triplets are present on equally-likely basis (that is, with uniform probability distribution); as such, the intron set is regarded as "junk" and non-informative [4.4].

Codons tend to be typically no more than 200 characters long, while noncodons could be tens of thousands of characters in length. Thus in majority, introns prevail mostly in a typical eukaryotic gene. Should any errors have occurred (due to mutations), they would give room to the possibility of evolving wrong or cryptic splice-junctions and lead to (imperfect) translations. That is, aberrant splice-junctions may result from mutational spectrum and would hamper the making of correct proteins

4.4 Information-theoretic (IT) Measures

In bioinformatics, the major interest lies in analyzing genomic and proteomic sequence details. Relevant to genomic sequence, entropy implies probabilistic (uncertainty) aspects of nucleotides distributed spatially along the sequence length. Inasmuch as the sequence is a mix of codon and non-codon parts, the appearance of the nucleotides {A, T, G, C} will be inherently random, but the Shannon information of the underlying genetic code would remain in the coding sequence segments (CDS). The following algorithmic/computational considerations can be identified thereof. For example, finding the delineating border of separation between codon and noncodon regions in a massive stretch of a DNA chain is a bioinformatic problem. Relevant methodology developed for this purposes uses information-theoretics based metrics so as to score the differentiating extents of statistics between codon-noncodon populations at a given site on the DNA sequence.

Several measures of information have been proposed in literature [4.1] each with distinct properties leading to variety in their applications. To classify such measures, they can be first categorized as parametric, non-parametric and entropy-type measures of information.

4.4.1 Parametric Measures of Information

Parametric measures of information estimate the extent of information about an unknown parameter θ contained in a data set and are functions of θ. The best known measure of this type is the so-called Fisher's measure of information. Specified as the Fisher discriminant metric (F-measure), it is based on a linear discrimination function with a set of coefficients optimized on the basis of statistical features of a data collection. It can be used as a scoring metric to contrast statistical subsets (possessing relative uncertainty) in a data set. That is, the concept of Fisher discriminant function enables classifying (or distinguishing) a pair of data sets. It was originally developed by R. A. Fisher in 1935 [4.7] with reference to taxonomic studies. Relevant effort refers to finding out the extent to which two sets of data are statistically similar or dissimilar.

The classical approach due to Fisher involves prescribing a linear function (F) with unknown coefficients $\{\lambda_i\}$ for a set of measurements $\{\theta_i\}$carried out on a population; and, this function in effect, is optimized with a choice of $\{\lambda_i\}$, so as to provide the largest scoring that distinguishes the two subsets of test data. In other words, the Fisher linear discriminant with optimized coefficients applied to a set of measurements would enable the subsets of the population "best discriminated" [4.1], [4.4] and [4.8]. For each set of measurements (pertinent to an "inherent" variable/parameter), the associated *a posteriori* distribution (depicting actual statistics of the parameter (variable) in question) can be compared against an *apriori*, set of uniformly-distributed variable data set. The underlying heuristic of such comparison leads to the equation:

$$I_F(\theta) = f \quad \frac{\det\left|\mathrm{Var}\ (\theta_{\mathrm{Uniform}})\right|}{\det\left|\mathrm{Var}\,(\theta_{\mathrm{Actual}})\right|} \qquad (4.4)$$

where, f is a function that should be continuous and strictly increasing and var denotes the variance of (θ). Fisher-metric, for example can be applied to a test sequence so as to delineate the codon and noncodon parts [4.4]. It involves constructing a set of discrimination matrices *vis-à-vis* the subsets of data in the parts to be discriminated. F-metric is useful in identifying

4.4.2 Non-parametric Measures of Information

Non-parametric measures give the amount of information supplied by the data to discriminate in favor of a probability distribution f_1 against another f_2, or for measuring the distance (or affinity) between f_1 and f_2. For example the so-called, Kullback-Leibler (KL) measure belongs to this class [4.1] and [4.2]. It is an entropy-estimator method that extracts "meaningful signal" to distinguish the exon/intron segments of a test genomic sequence. Such conditional entropy aspects of statistical divergence (SD) are based on relative entropy (mutual information) considerations on coding *versus* non-coding regions. Given that p_c denotes the probability of codon-population statistics and p_{nc} depicts the probability of noncodon population statistics, the KL measure is specified by:

$$\sum_i [p_c \log(p_c / p_{nc})]_i \rightarrow \{A, T, G, C\}$$

or

$$\sum_i [p_{nc} \log(p_{nc} / p_c)]_i \rightarrow \{A, T, G, C\} \qquad (4.5)$$

There also exist a set of other distance measures such as, Mahalanobis measure, Bhattacharyya measure, Ali-Silvey measure etc., which have been adopted for genomic sequence analyses purposes [4.1] and [4.4]. Yet another non-parametric symmetric divergence measure belongs to the class of so-called Csiszár's f-divergences, which are more general than the KL version. Details on such measures are available in [4.1] and [4.2].

In essence, KL measure is a subset of this Csiszár family of cross-entropy functionals given by:

$$[p_c \times F(p_c / p_{nc})] \text{ or } [p_{nc} \times F(p_{nc} / p_c)] \tag{4.6}$$

where, F(.) is a doubly differentiable convex function. A host of measures thereof can be specified by proper choice of F(.). Popularly, the KL-measure, the Jensen-Shannon (JS) measure etc. have been used in bioinformatics contexts [4.1].

The main driver behind the success of the above methods in genomic context is due to distinguishable statistical characteristics of exon and intron segments. That is, a non-uniform codon usage prevails in the exon part meaning that, specific to coding regions not all bases of {A, T, G, C} occur with the same probability; but, there are subtle differences between the statistics of their appearance that exist depending on the position of each base in the codon triplets. In contrast, in non-informative intron segments, the occurrence probabilities of A, T, G and C are the same (equal to ¼). A typical KL-measure based codon/non-codon delineation is indicated in Chapter V. Also, an example of recognizing the presence and extent of the so-called CpG sequences in the test ssDNA genome has been indicated in Section 4.8.

4.4.3 Measures of Entropy: Bioinformatic Context

As indicated earlier, measures of entropy implicitly express the amount of information contained in a distribution, namely, the amount of uncertainty associated with the outcome of an experiment. The classical measures of this type are specified explicitly *via* Shannon's concept and extended as Rényi's measures [4.1]. However, without any loss of generality, the genomic analyses can be done *via* any one of the aforesaid versions.

4.5 Characterization of Genomic Feature *via* IT Measures

In order to quantify the negentropic characteristics of the DNA, the following questions should be answered:

- Are the bases in the DNA chain independent events?
- Does the occurrence of any one base along the chain alter the probability of occurrence of the base next to it implying (Markovian attribute)?
- What is the conditional probability that when a base A occurs, it will be followed by A or any other base designated at T, C or G?

Relevant salient aspects of DNA information can be summarized as follows [4.1]: (i) DNA information is inherently redundant; (ii) it represents at least a first-order Markov source output; (iii) DNA information can be regarded as the output of an ergodic source; (iv) Shannon's redundancy factor (R) corresponds to: 1 – $[H_m/\log_2 (a)]$, where 'a' is the number of states (or epochs) of the statistics involved; and, $[\log_2 (a)]$ corresponds to the maximum entropy of the sequence of equiprobable, independent elementary events I = 1, 2, ..., a; H_m denotes the entropy of the first-order Markov chain and $[H_m/\log_2 (a)]$ is defined as the relative entropy [4.9]. Further, (v) Shannon's Second Theorem on genetic information transmission [4.1]: "It is possible within limits, to increase the fidelity of the genetic message without loss of potential message, provided that the entropy variables change in the proper way, namely, by increasing D_2 at relatively constant D_1". Here, the set $\{D_1\ D_2\}$ refers to the D-indices. D_1 is the divergence from the equiprobable state of independent occurrences; and, corresponding divergence from equiprobable state of non-independent occurrences, which is denoted by D_2, an evolutionary index that seperates higher organisms (like vertebrates) from lower species. Vertebrates can accomplish such a source encoding. That is the reason for them to be “higher” organisms. A review on the application of information theory to DNA sequence analysis is available in [4.10].

4.6 Fuzzy Attributes of Genomic Sequences

The large data spaces of bioinformatics can be in many instances overlapping. Such imprecise domains possess unique difficulties when the underlying details are extracted. This is because of the non-specificity of the values and sharplessness of the boundaries of activity variables that can be described mostly in linguistic norms of fuzziness or grey facts. With fuzzy attributes prevailing in bioinformatic contexts, a multi-value logic has to be appropriately built, making the grey truth into complex schemes of formal reasoning.

As mentioned earlier, exclusive to this chapter, the entropy/IT considerations on genomic analysis is focused on presenting the associated fuzzy attributes. The task indicated thereof refers to elucidating the fuzzy profile that may prevail at the splice-junctions between codons and noncodons. As regard to such fuzzy considerations, developed in [4.4] is a strategy that identifies the splice-junctions between codon and non-codon regions present in a massive stretch of a DNA chain, especially when the delineating boundary in question is submerged in a subspace where codon and non-codon parts exist as overlapping and ambiguous/fuzzy entities. A fuzzy inference engine (FIE) developed thereof uses again information-theoretic based metrics (with relevant algorithms applied to symbolic as well as binary sequence data representing the DNA) so as to score differentiating extents of codon/non-codon populations at a given site in the DNA sequence. The information-theoretic metrics adopted in [4.4] refer to various statistical divergence (such as KL and JS measures) as well as distance and discriminant concepts. Further, the algorithms indicated in [4.4] yield consistent results on the delineation boundary sought on test subspaces that are fuzzy; and simulated studies using human as well as bacteria codon-statistics confirm the efficacy of the approach pursued.

Notwithstanding the existence of pursuits as above in locating the splice-junctions, the statistical divergence (SD) can be extended in getting mapped into a novel membership function that specifies the fuzzy subspace of overlapping exon and intron segments. Relevant membership function is defined thereof on the basis of an "error'' feature prevailing in the overlapping ("noisy") segment with mutational aberrations. The underlying heuristics are as follows.

As stated earlier, the locations of splice-junction may not so reliably distinct. However, in a canonical sense, the splice-junction *consensus* may follow certain rules as regard to introns and exons [4.11]. For example, the introns almost always begin with the residue set {**gt**} at 5′-end and ends with an {**ag**} at the 3′-end. But, inasmuch as the nucleotide sequence corresponds to a set of statistically permutated elements, {A, T, G, C}, numerous putatively occurring {**gt**} and {**ag**} locations (other than in the introns as indicated) may prevail and resemble such canonical patterns. This implies that relying on such canonical details alone may not reasonably and robustly show the presence of true splice-junctions. Further, in the event of point-mutations, stemming of aberrant splice-sites is inevitable [4.12]. As such, should a junction be recognized and prevailing of possible cryptic junction sites should be elucidated, it is necessary to analyze statistically prevailing long-range genetic information so as to determine the extent to which subsequences surrounding the splice-junctions differ from sequence segments of adjoining spurious analogs; hence, true *versus* aberrant (cryptic) splice junctions can be distinguishably identified. A feasible suite of analysis is as follows:

Evolutionary conservation of splice-junctions is invariably hampered with inevitable phylogenetic-specific mutations. If such mutations are (assumed) independent, any “noisy” change in the spatial DNA pattern of the sequence (at the splice-junctions) can be marked as a “spatial jitter” with a characteristic parameter called *spatial signal-to-noise ratio* (SSNR).

Splice-junctions with a spatial jitter as above correspond to fuzzy offsets of exons and introns at their junctions. That is, the spatially-jittered junction corresponds to an overlapping mix of codon and non-codon entities and hence constitutes a (fuzzy) universe. In other words, the splice-junction information has a fuzzy structure that can only be identified/specified in norms of linguistic descriptions. Such descriptions can be characterized by a membership (function) [4.1] and [4.12] of belongingness to the attributes of exon or introns.

Indicated here is an appropriate FIE that delineates fuzzy overlaps of codon/non-codon parts so as to elucidate the underlying cryptic (or aberrant) splice-junctions. This is done on the basis of SSNR defined with reference to the spatial-jitter. The SSNR is also adopted to represent the relevant membership function.

4.6.1. Spatial-jitter across Splice-junctions

Consider a small window(-length) accommodating a finite-number (say, 100) of putatively occurring base residues along a DNA sequence. Suppose this window traverses a splice-junction. With no *a priori* information available on the accurate disposition of the splice-junction, it can be initially assumed that the reading gathered thereof is a "blurred" information implying an overlap of exon/intron region with a fuzzy codon/non-codon transition. That is, a spreading function is assumed to prevail across the finite window-length. The resulting spatially-varying 1-D signal so gathered from the scan of the entire DNA sequence would resemble a set of random telegraphic waveform train constituted by changing statistical profiles of exons and introns (being scanned). The task in hand is then to detect the spatial transition sites, each delineating adjoining exon/intron (or intron/exon) segments despite of the noisy, blurred spatial information (of the transition site).

Suppose $\pi(x)$ represents an uncorrupted DNA sequence pattern metric computed along the variable x denoting the 1-D space of the sequence length. Associated signal component will assumed to be corrupted in the event of mutational changes in {A, C, T, G} had occurred along the sequence are encountered. Such mutation-specific effects can be modeled as a contribution of "noise", m(x) on the signal part, $\pi(x)$. Hence, the signal output of the window-reader can be modeled by either a spatial-domain convolution description, namely, $s(x) = \pi(x)*m(x)$ or, equivalently by a corresponding frequency-domain description, $S(f) = \Pi(f)M(f)$, where S(f), $\Pi(f)$ and M(f) are the Fourier transforms of s(x), $\pi(x)$ and m(x) respectively.

Consider an intron-exon splice junction illustrated in Figure 4.1. The upper figure (marked as (a)) is a crisp noise-free (uncorrupted) site with a splice-junction at x_o along the DNA sequence constituted by {A, C, T, G} residues. Should mutational corruptions have taken place, this crisp transition-boundary x_o becomes ($x_o \pm \Delta x$), where Δx denotes spatial jitter (marked as (b)).

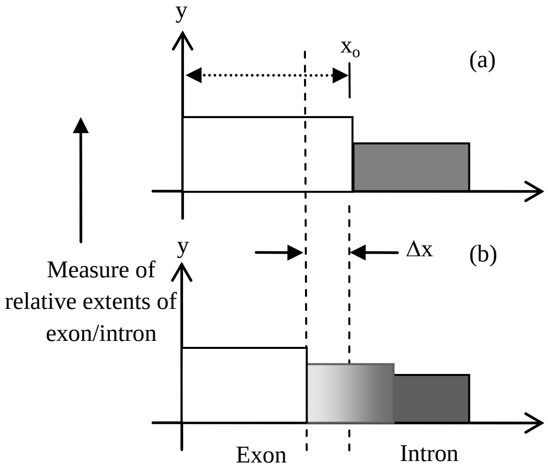


Figure 4.1: "Spatially-jittered" splice-junction manifesting as fuzzy exon/intron (or *vice versa*) transitional residues along the sequence.
(a) Unaltered (crisp) splice-junction;
(a) Fuzzy splice-junction with graded variation of divergence (distance) between the statistical features (specified as a measure on the ordinate, y) of exon/intron (or intro/exon) along the transition region. The abscissa (x) depicts a scale of residues along the DNA sequence.

Further in Figure 4.1, the y-axis depicts the measure/metric of (relative) statistical divergence of exon *versus* intron (or *vice versa*) prevailing at any point, x on the sequence. (This statistical divergence prevails due to the reason that exon has a distinct distribution of {A, C, T, G} constituents *vis-à-vis* the corresponding distribution in the intron segment).

The effect of (mutation-specific) corruption would make the splice-junction to become unclear of fuzzy, as shown in Figure 4.1 (b). In essence, Δx is a jitter variable superimposed on s(x) corresponding to crisp disposition of the splice junction x_o. The expected root-mean-squared (rms) jitter J_r at any splice-junction x_o can be expressed by the "noise power" imposed by the mutation errors.

In traditional communication theory, the term *signal-to-noise ratio* (SNR) is defined to specify the quality of an uncorrupted "signal (power) level" to the corrupting "noise power". Translating this concept, suppose the average length of intron-plus-exon is $\bar{x}$, corresponding "spatial SNR" (SSNR) with reference to the DNA sequence space (of Figure 4.1) can be defined as follows: SSNR = $(\bar{x})^2/J_r^2$.

4.6.2. Error probability of Splice-junction Prediction

Relevant to a "noisy" intron/exon (or exon/intron) transitions, the accuracy of locating the transition site, x_o is constrained by the probability of error associated with the estimation of x_o. In this context, within the specified blurring limits of jitter, the SSNR implicitly would predict the error probability of estimating the splice-junction.

Suppose a sequence of exon/intron (or *vice versa*) transitions (x^i_o's) prevail at locations indexed by i = 0, 1, 2, …, m. From these data, one can extract exon or intron widths (χ) as follows: $\chi_{i+1} = (x_{i+1} - x_i)_{E \text{ or } I}$ for all values of i = 0, 1, 2, …, m, where the suffix (E or I) denotes the measurement done on an exon or an intron respectively. In terms of the average length of consequent intron plus exon $(\bar{x})$ subspaces, the transition (split-junction) locations in the presence of mutation error-induced jitter can be expressed as follows: $(x_i)_{\text{Noisy}} = \sum_{j=0}^{i} k_j\bar{X} + \delta_i\bar{X}$, where k_j is an integer with k_o being zero; and, i = 0, 1, 2, …, m; further, δ is a dimensionless random variable, which in a simple case, has zero-mean Gaussian distribution with variance σ^2 = (1/SSNR). (This variance is invariant along the sequence length if the sequence statistics is assumed to be stationary).

Now defining a normalized variable, $\kappa_i = \chi_i/\bar{X}$, it can be estimated as $\kappa_i = K_i + (\delta_i - \delta_{i-1})$ with (i = 0, 1, 2, …, m); hence one can specify the probability of correct decoding of the splice-junction, $P_c(m)$ as the probability that $|\kappa_i - K_i| \le 0.5$.

Inasmuch as, $\kappa_i = K_i + (\delta_i - \delta_{i-1})$, the aforesaid probability can be restated as follows:

$$
\begin{aligned}
&P_c(m) = \text{Prob}\left\{|\delta_1 - \delta_0|\right\} \\
&|\delta_1 - \delta_0| \le 0.5....,|\delta_1 - \delta_{m-1}| \le 0.5
\end{aligned}
\tag{4.7}
$$

With the assumed Gaussian statistics for δ, the cumulative probability of correct decoding of the splice-junction, namely $P_c(m)$ can be deduced as follows:

$$
\begin{aligned}
Pc(\Delta x,\sigma)_{x_o} &= \sqrt{\frac{1}{2\pi\sigma}}\int_{-\infty}^{\Delta x} \exp\left[-\frac{1}{2}\left(\frac{\Delta x}{\sigma}\right)^2\right] d(\Delta x) \\
&= \frac{1}{2} + \frac{1}{2}\,\text{erf}\left(\frac{\Delta x}{\sqrt{2}\sigma}\right)
\end{aligned}
\tag{4.8}
$$

where, Δx with respect to an i^th junction is given by $\Delta_i x$ = $(\delta_i - \delta_{i-1})x$; and, erf(u) = $\frac{2}{\sqrt{\pi}}\int_0^u \exp(-u^2)du$. Further, the fuzzy-space in question enclaves the universe $\Im^m$ depicting an m-dimensional hypercube across the unit interval, **I** ∈ [– 0.5, + 0.5].

Equation (4.8) implies that the probability of correct detection (and hence error probability) of the splice-junction disposition is implicitly dependent on SSNR parameter. The plot of equation (4.8) is shown in Figure 4.2, where P_c is plotted as a function of $(x_o \pm \Delta x)/x_o$ with respect to a presumed, crisp splice-junction at x_o posing a transitional error-prone width Δx.

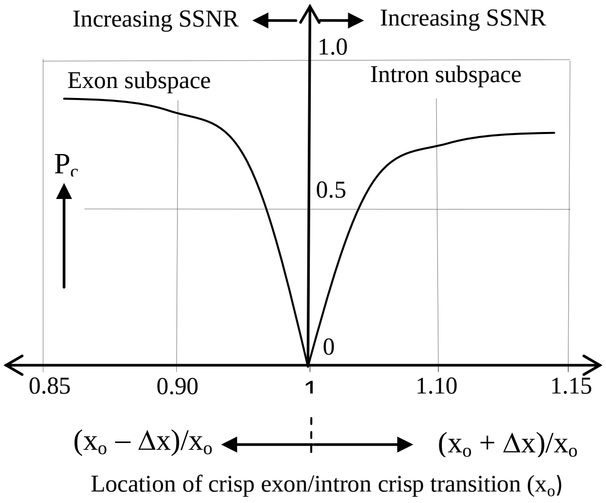


Figure 4.2: The probability of correct estimation of a splice-junction

This error-prone region depicts a subspace of overlapping exon/intron subspaces that smear the exact location of x_o. This unspecific (error-prone) subspace Δx is therefore, fuzzy imposing an imprecision on x_o. Relevantly, the generic description of P_c in this fuzzy subspace takes a membership attribute of vagueness *vis-à-vis* the position variable, x. The membership here depicts the belongingness to exon subspace or intron subspace. Hence described in the next section, are the underlying aspects of the fuzzy subspace in question with the object of ascertaining the splice-junction in the fuzzy subspace.

4.6.3. Fuzzy Splice-junction Prediction

Suppose a set of input values Δx_i are taken from the sequence and considered as non-specific or fuzzy. By denoting those segment values by $\{\Delta_i x\}_f$, corresponding $\{(P_c)_i\}_f$ can be written in terms of uncertain limiting-values of all the vectors in the bounding (lower and upper) interval, $\Delta x \in [\Delta x_L, \Delta x_H]$. Hence it follows that [4.1]:

$$\left\{P_C[\Delta x]_i\right\}_f \approx \left\{\left\{P_C[\Delta x_L]_i\right\}_f\right\} + \sum_{j=1}^{\alpha-1} (\rho_f)^{j-1}\left[\left\{P_C[\Delta x_L]_i\right\}_f\right]\Delta x^j/j! \qquad (4.9)$$

where, ρ_f (.) depicts the slope equal to $d(P_c)_i/d\Delta x_i$ and α is the number of interval-valued parameter for the range within $[\Delta x_L, \Delta x_H]$. Further, equation 4.9 denotes an algebraic sum of *addenda* computed *via* interval arithmetic and denotes the "width of the results". In other words, for the specified vector bounding-limits of $\{(P_c)_i\}_f$, namely, $\Delta x \in [\Delta x_L, \Delta x_H]$, an α-set of interval-valued parameters namely, $\{\mathbf{Q}\}$, $Q = Q_1, Q_2, \ldots, Q_\alpha$, prevails at or around x_o with no fuzzy attributes. Then relevant crisp-domain relation of $\{\Delta x\}$ *versus* $\{P_c\}$ can be written by a differential equation given by : $d^2P_c/d\Delta x^2 + (dP_c/d\Delta x)^2 = g(\Delta x)$ where g(.) is some arbitrary function of Δx.

In the event of overlapping fuzzy attributes existing at x_o, then the corresponding (fuzzy)-domain relation between $\{\Delta x\}$ *versus* $\{P_c\}$ can also be generalized by a stochastical discourse of P_c *versus* Δx expressed in terms of a fuzzy stochastical differential equation [4.1]. Further, in such exon-to-intron transition subspace (denoted as **F)** having fuzzy attributes, corresponding demarcation of exon/intron transition can be assumed to be at a centroid location (X_C) with a line-of-delineation through a centroid. This location refers to a defuzzified elucidation based on membership of belongingness of the site-of-interest in the fuzzy space. The procedure to find X_C is described below.

4.6.4 Centroid of Fuzzy Subspace at the Splice-junction

The SSNR and P_c considerations *versus* $(x_o \pm \Delta x)/x_o$ indicated before imply inherent statistical attributes of {A, T, G, C} population in the exon and intron regions across the splice-junction. The exon-side statistics encodes for genetic information (to make necessary protein) and the intron-side statistics is non-informative. In other words, suppose the probabilities of occurrence of the elements {A, T, G, C} in the exon are denoted by the set: $\{Q_A, Q_T, Q_G, Q_C\}$ with $(Q_A + Q_T + Q_G + Q_C = 1)$. Then, the associated errors for the elements of {A, T, G, C) are decided by the inequalities, $Q_A \neq Q_T \neq Q_G \neq Q_C$. Now, suppose the corresponding probabilities of occurrence in the intron are: $\{\Theta_A, \Theta_T, \Theta_G, \Theta_C\}$ with $(\Theta_A + \Theta_T + \Theta_G + \Theta_C = 1)$; then, the associated errors for the elements of {A, T, G, C } on intron-side are set by the condition that, $\Theta_A = \Theta_T = \Theta_G = \Theta_C = 0.25$. This is because the intron-side being non-informative, Laplacian hypothesis applies in presuming that all (four) elements are equally-likely to occur. Hence, with the distinction in the values of $\{Q\}_{A,\ T,\ G,\ C}$ and those of $\{\Theta\}_{A,\ T,\ G,\ C}$ relevant entropy/information-theoretic (IT) distances (that is, the statistical divergence or SD values) computed (for the exon and intron regions) would show distinction in the profiles of SD (in exon and intron regions) as illustrated in Figure 4.3. (This SD can be any one on the divergence measure such as KL or JS mentioned before. Illustrative measures are presented later in the results with reference to a real DNA structure).

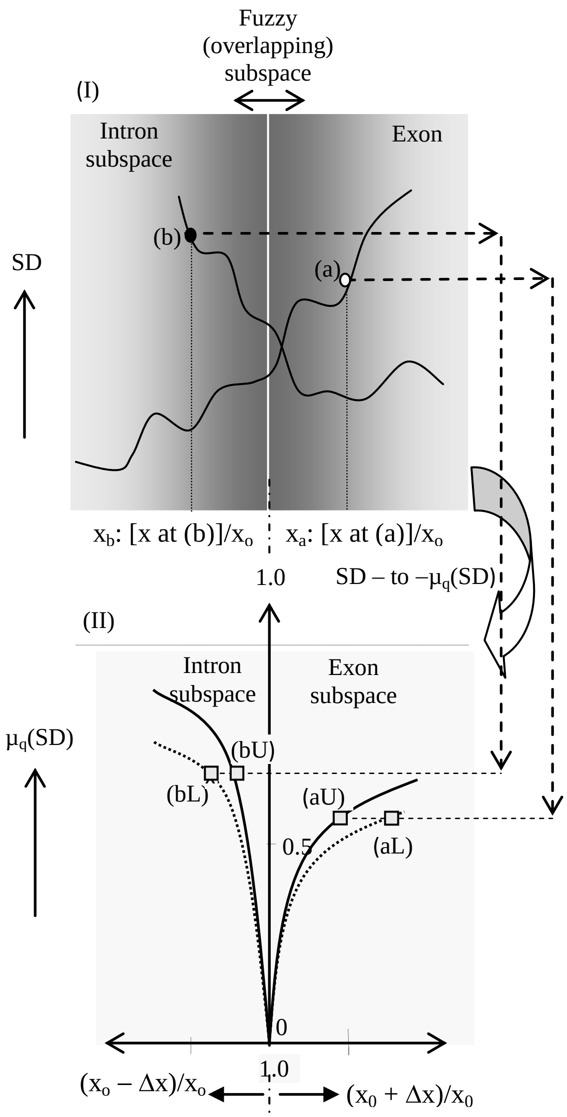


Figure 4.3: SD – to –μ_q(SD) mapping. The SD-value “a” maps to upper- and lower-limits of μ_q(SD) respectively as (aU) and (aL). Similarly, the SD-value “b” maps to upper- and lower-limits of μ_q(SD) respectively as (bU) and (bL).

I: $(x_o \pm \Delta x)/x_o$ *versus* SD curves in the intron and exon subspaces. Note the SD profiles are distinct in each region

II: $(x_o \pm \Delta x)/x_o$ *versus* membership function, μ_q(SD)

Following the considerations presented in [4.13] and [4.14], the expression for P_c, namely, $(1/2) + (1/2)\mathrm{erf}(\Delta x/\sqrt{2}\sigma)$, can be approximately written as: $L'_q(z)/L'_q(0)$ where $L_q(z)$ denotes the Bernoulli-Langevin function and the prime sign depicts the differentiation with respect to the argument $z = (\Delta x/\sqrt{}2\sigma)$. Explicitly, $L_q(z) = (1 + 1/2q)\coth\{(1 + 1/2q)z\} - (1/2q)\coth\{(1/2q)z\}$ where q represents an disorder entity associated with the statistics of the population concerned [4.16]. Shown in [4.16] is that the upper-bound of the isotropic disorder statistics is decided with q = ½ and the lower-bound (depicting an anisotropic disorder) is specified by q → ∞. Inasmuch as the statistics of exon-region would differ from that of intron-region, $q_E \neq q_I$. Further, as indicated in [4.14], the ratio $L'_q(z)/L'_q(0)$ denotes approximately the membership function μ_q for the region (fuzzy space or block, **F**:$\{x_i\}$) of interest with its fuzzy range (upper-to-lower) is decided by: q = ½ to q → ∞.

Hence, shown in Figure 4.3, is the mapping of computed divergence measures (SD) of intron and exon subspaces (across the slice-junction) into corresponding membership values, μ_q(SD) (with q = ½ yielding upper-bound values and q → ∞ giving the lower-bound values). For example, suppose a location x_a (in exon region) gives the SD-value equal to (a). Then, the value (a) maps on to the membership-plane with the entities (aU) and (aL) depicting respectively, the upper- and lower-bound values. Similarly, assuming a location x_b (in intron region) has an SD-value (b), this value maps on to the membership-plane as (bU) and (bL) denoting respectively the upper- and lower-limits. The steps as above can be elaborated as follows:

First, the chosen divergence measure (SD: KL or JS) is computed for the entire fuzzy domain **F** at each pointer-position within a chosen window-size. For this purpose, two subspaces $\mathbf{F}_{\text{Exon}}$ and $\mathbf{F}_{\text{Intron}}$ depicting respectively, the exon- and intron-side of the **F**-space are specified. Then, the computation of the SD-measures with exon statistics $\{Q\}_{A, T, G, C}$ in F_{Exon}-subspace and with intron statistics $\{\Theta\}_{A, T, G, C}$ in F_{Intron}- subspace is done with KL or JS algorithm.

The values of SD generated in each differential window (of F_{Exon}- and F_{Intron}-subspaces) accounts for the extents of codons and noncodons in the relevant fuzzy subspace. Corresponding to the window-specific pointer-positions along the sequence, the SD-score profile obtained across each differential block will be distinct for each subspace (exon or intron) in question. Next, the values of SD obtained are translated *via* membership function to provide descriptive details of belongingness in the fuzzy domain.

The translated values gathered can be subjected to a defuzzification process [4.13] and [4.16] in order to get the centroid position (of the pointer) that delineates the boundary of the two test fuzzy subspaces. Relevant local search follows the heuristics of "search and score" applied appropriately on the assigned membership values that describe the qualitative descriptions of overlapping and ambiguous codon/non-codon locales across the fuzzy site.

The boundary that marks the desired splice-junction being searched corresponds to a defuzzified location obtained *via* centroid-finding method. Towards the centroid, the fuzzy exon-domain profile and the fuzzy intron-domain would converge close a single membership value. Referring to Figure 4.3, the SD value (a) in the exon subspace yields mapped values (aL and aU); and, the SD value, (b) in the intron subspace maps into (bL, bU). Suppose the set {aL, aU} in turn projects on to x-axis at x_{aL} and x_{aU} respectively. Likewise, the set {bL, bU} projects on to x-axis at x_{bL} and x_{bU} respectively. Then, the mean position of x_{aL}, x_{aU}, x_{bL} and x_{bU} would correspond to the centroid being sought.

4.6.5 Simulation Experiments Using Real DNA Data

The efficacy of the efforts and procedure described above is illustrated with an example of real-world DNA sequence of dengue virus type 1 (NCBI Reference Sequence: NC_001477.1). Its CDS stretches from the nucleotide position 95 through 10273. Using the nucleotide population details, a moving-window based calculation of KL-measure is plotted in Figure 4.4 across the entire sequence length.

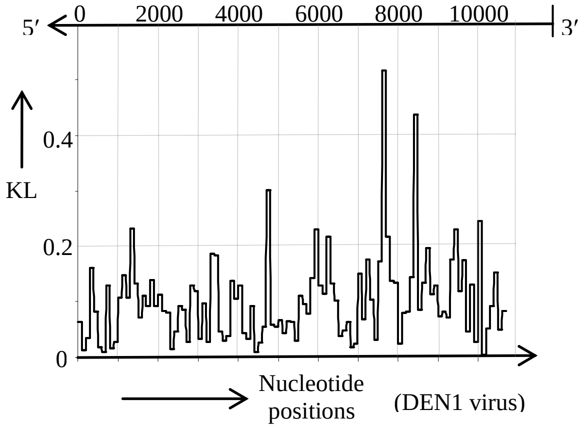


Figure 4.4: Nucleotide position *versus* computed KL-measure of the DNA sequence of dengue virus type 1 (NCBI Reference Sequence: NC_001477.1)

The data available in NCBI GenBank for example, shows a CDS stretch from the position 7574 through 10270 with an indication of a transition at 7574. Presented below in Figure 4.5 is an exclusive plot of KL-measure across this selected CDS regime at the transition locale around 7574. While the codon (exon)/non-codon (intron) transition is markedly seen, there is however a subspace of fuzziness, wherein an overlap of exon and intron regimes prevails indistinguishably (viewed with the simple KL-measure).

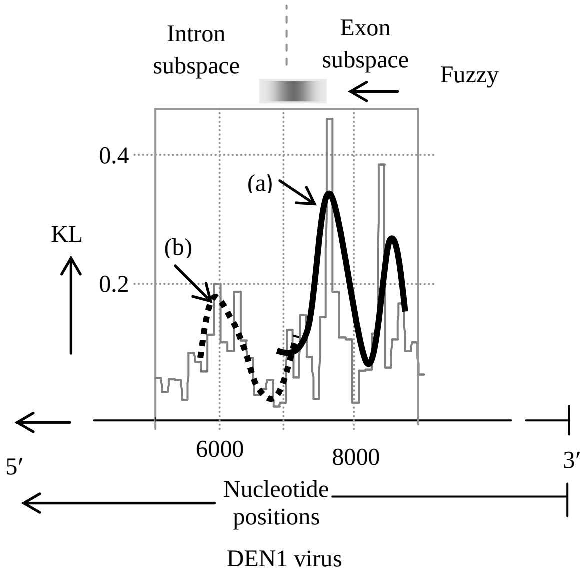


Figure 4.5: Nucleotide position in the limited range of 5000 to 9000 *versus* computed KL-measure of the DNA sequence of Dengue virus type 1 (NCBI Reference Sequence: NC_001477.1).

Therefore, by assigning membership attribute, the fuzzy inference engine algorithm (described earlier) can be invoked to decide on the location of the splice-junction in the fuzzy region. For this purpose, drawn in Figure 4.6 is the profile of membership values (μ_q) mapped from the computed KL-measures across the transition region of interest. There are two profiles: (a) depicts μ_q-values with q = ½ (meaning the upper-bound on the membership); and, (b) denotes μ_q-values with q = ∞ (meaning the lower-bound on the membership).

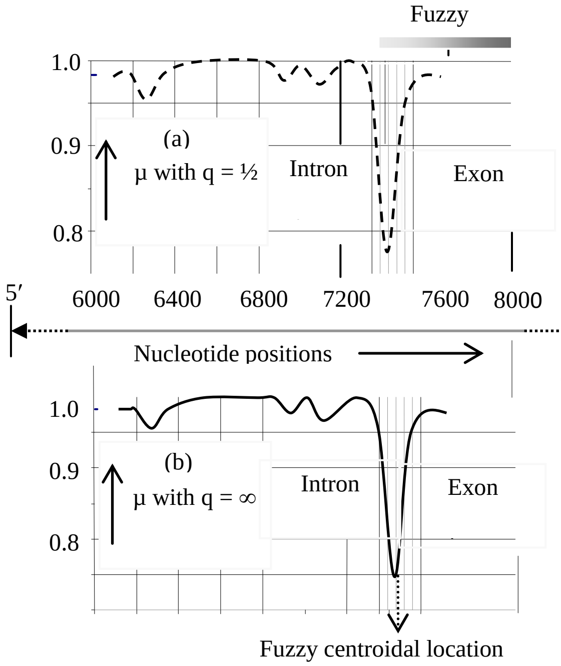


Figure 4.6: Membership profiles (μ_q) across the fuzzy transition region of interest.

(a) μ_q-values with q = ½ (meaning the upper-bound on the membership) *versus* nucleotide positions of the test DNA

(b) μ_q-values with q = ∞ (meaning the lower-bound on the membership) *versus* nucleotide positions of the test DNA.

From Figure 4.6, the location of the splice-junction buried in the fuzzy domain can be ascertained. This location corresponds to the centroid coordinate (x_C). This centroid position is featured by the upper- and lower-bound profiles of the μ-value. In view of the discussion presented earlier, x_C corresponds to the mean position of x_{aL}, x_{aU}, x_{bL} and x_{bU}. For the data presented in Figure 4.6, relevant computed results show that this centroid (x_C) is at 7401 (as against the crisp value indicated in NCBI GenBank as 7574). This centroid (7401) is the mean of: [$(x_{bL} + x_{bU})/2 = 7401$] and [$(x_{aL} + x_{aU})/2 = 7401$].

Depicted in Figure 4.7(a) are base residues reported around, for example splice-junction site, namely 7574. The present method predicts in addition, a cryptic set of 7370 and 7419 in the vicinity of the centroid 7401 determined. The selection of this set {7370, 7401} is based on the heuristics of [4.11] suggesting the intron's 3′-side preferential ending being **ag**. That is, the values 7370 and 7401 are picked around the centroid determined such that they are in conformance with the abutting of ag-residues.

In Figure 4.7(a)-(i), the intron-subspace ends with residue set {**ag**}at 7574 and is consistent with the canonical splice-junction consensus (as mentioned earlier) of [4.11]. Notwithstanding this canonical pattern, the mutational influences could have possibly induced aberrant splice-junctions. A scan through the test DNA indicates a cluster of sites between 7500 through 7700 exist at which the residues **a** and **g** occur together making it ambiguous on the decision that splice-junction alone can be the splice-junction of interest. However, following the fuzzy pursuit presented here, it enables pointing out that other cryptic splice-junctions such as 7370 and 7419 could reasonably be alternative splice-junction sites having adjacent **ag** residues as illustrated, for example in Figure 4.7(a)-(ii) with 7419 site.

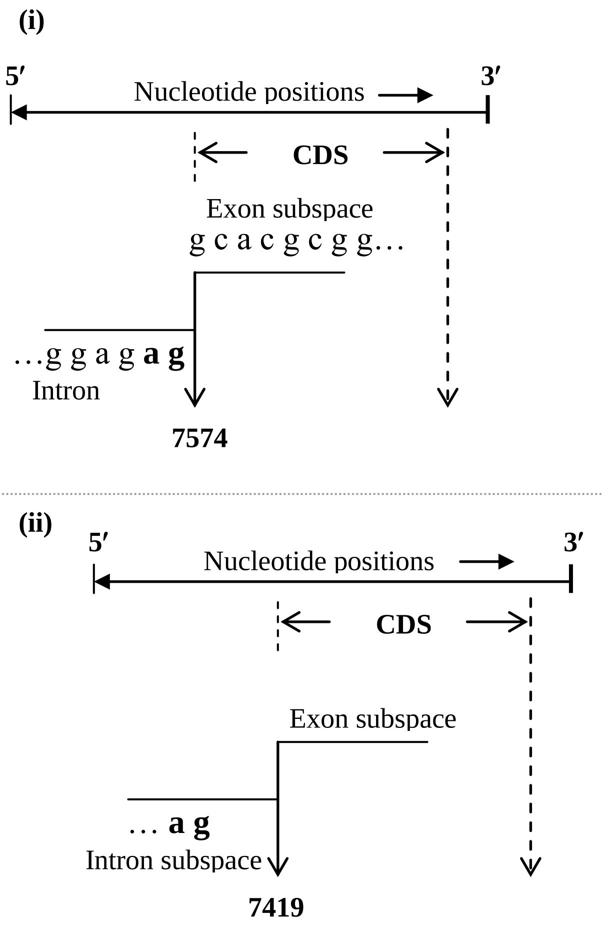


Figure 4.7(a): The details on nucleotides adjacent to the predicted splice-junctions: (i) As per [13]; and (ii) as per present method. (In both cases, the intron-subspace ends with a residue pair **ag** bases consistent with the canonical splice-junction consensus)

The complete list of aberrant splice junctions evaluated for the test viral DNA in the present study is presented in Table 4.1 and illustrated in Figure 4.7(b). Table 4.1 indicates the centroid values determined as well as cryptic transition sites predicted on the basis of the heuristics of [4.11]. It may be noted that the data available in NCBI Genbank portrays overlaps of CDS domains that eventually facilitate various protein structures as listed.

Table 4.1: Transition sites indicated in NCBI GenBank and the predicted sites as per the present method

CDS range data	Description	Transition site	Bounds of membership value		Centroid of UB and LB	Cryptic transition sites predicted**
			Upper-bound (UB)*	Lower-bound (LB)*		
95..394	Capsid protein	394	1, 401	301	352	350 354 394
94…436	Anchored capsid protein	436	301, 701	301, 701	501	515
710..934	Membrane glycoprotein	710	701	701	701	954
437..934	Membrane glycoprotein precursor	934/935	701, 1101	701, 1101	901	
935..2419	Envelope protein	2419/2420	1801, 2501	2801	2151	2160
2420..3475	Nonstructural protein 1	3475/3476	3301, 3801	3301, 3801	3551	3553
3476..4129	Nonstructural protein 2a	4129/4130	4001, 4301	4001, 4301	4151	4149, 4170
4130..4519	Nonstructural protein 2b	4519/4520	4301, 4701	4301, 4701	4501	4326, 4356 4452, 4505
4520..6376	Nonstructural protein 3	6376	6201, 6701	6201,6701	6451	6447,6462
6377..6757	Nonstructural protein 4a	6757	6701, 7001	6701, 7001	6851	6833,6857
6758..6826	2k protein	6826	6701, 7001	6701, 7001	6850	
6827..7573	Nonstructural protein 4b	7573/74	7201, 7601	7201, 7601	7401	7370,7419
7574..10270	Nonstructural protein 5	10270	10001, 10401	10001, 10401	10201	10202,10211

** The UB and LB values indicated correspond to the sites where minima of μ_q-plot (map) in the fuzzy domain of interest are observed, (for example, see Figure 4.6).

* The predicted site is based on locating a site in the vicinity of the centroid where the introns almost always begin with the residue set {GT} at 5′-end and ends with an {AG} at the 3′-end as illustrated in Figure 4.7

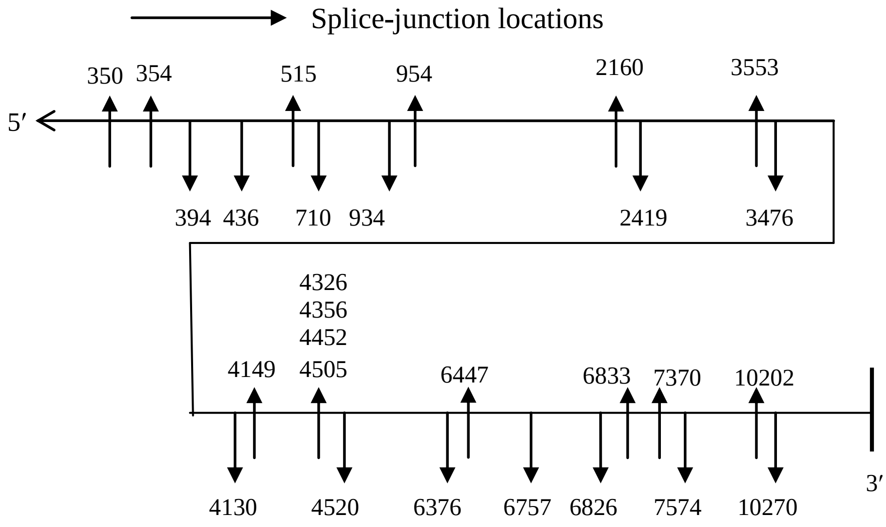


Figure 4.7(b): Summary of results on the locations of splice junctions. Downward arrows indicate values available in NCBI GenBank for DEN1 virus. Upward arrows indicated computed values that include details of cryptic sites in the fuzzy subspaces

The purpose of knowing correct and aberrant splice-junctions in the context of viral DNA (such as DEN1 virus) is pertinent to and implicates vaccine designs [4.18]. In general, a gene is first transcribed into pre-mRNA, which is a copy of genomic DNA containing exon and intron regions. Gene-splicing is an important form of protein diversity and has also regulatory functions and RNA-splicing is essential so as to regulate precisely the process that occurs after gene transcription and before mRNA translation (in which introns are removed and exons are retained). The sequences between the boundaries of introns (denoting regions of DNA or precursor RNA that are not represented in mature RNA, but reside between regions) and exons (depicting regions of DNA or precursor RNA represented in mature RNA) are not random. There are several splicing events that are possible eventually resulting in: Exon-skipping, intron-retention, cryptic splice-site usage and alternative 3′- and 5′-side splice-sites [4.12]. Further, in RNA splicing, the so-called splicing-variants may be formed prior to mRNA translation due to differential inclusion or exclusion of regions in the pre-mRNA structure. Also, a systematic analysis of splice-junction sequences in eukaryotic protein coding genes using NCBI GenBank databank has revealed a striking similarity among the rare splice-junctions [4.11] that do not contain **ag** at the 3′ splice site, or **gt** at the 5′ splice site.

As mentioned before, indistinct splice-junctions would result from deleterious effects of mutations that target the splice-sites causing variability in splicing patterns. Such deleterious effects eventually form a major source of protein diversity leading to a considerable extent of diverse proteomic functions that stem from a relatively small number of genes. Thus, changes in splice-site (alternative splicing) can induce different effects on the encoded proteins, not only in humans but also in viruses.

As regard to the viral leader sequences, there may be a splice donor site for generation of subgenomic messages, usually the *env* (viral envelope) transcript. In general, the role of RNA splicing is to generate a set of stable splice-junctioned sequences in viruses so that virus mimicry is enabled as a mechanism for the potential variability in envelope proteins, (which are susceptible for changes due to point-mutation and thus, avoid to be recognized by T-memory cells of higher organisms in vaccine trials).

The present study offers a systematic way of elucidating cryptic splice-junction sites in viral DNA structures, the knowledge of which can be profitably used in vaccine design efforts. The study is being extended to a variety of viruses in order to elucidate the underlying cryptic aspects of splice-junctions. Pertinent analytical framework and computational aspects are augmented with the details available in [4.19 - 4.21].

4.7 Information Redundancy: Application to Genomic Sequences

The second objective of this chapter refers to formulating a genomic sequence analysis with the information redundancy (IR) being the metric of the associated entropy and/or information details. Specifically, this IR approach is presented to elucidate the distinguishing features of four related sequences such as those of strains of a viral species. For example, discussed in Chapter III, are the details on a virus like dengue with its serovar DEN1, DEN2, DEN3 and DEN4. These viral strains could be phylogenetically related and are causative for similar, but, distinct dengue fevers. Therefore, it is of interest to know not only the similarity features between the genomic sequences of these strains, but also useful to identify their distinguishing characteristics. The similarity features are elucidated in detail in Chapter VII. Here, IR concept is adopted to study their distinguishing characteristics.

The redundancy factor (R) for a genomic sequence is given by the equation:

$$R_1 = 1 - \frac{H(r)}{[H(r)]_M} \tag{4.10}$$

Here, H(r) is the information content of the genomic sequence and is given by equation (4.2). The redundancy factor for each strain of dengue virus is determined and is denoted as R_1, R_2, R_3 and R_4 for DEN1, DEN2, DEN3 and DEN4 respectively. Figure 4.8 below shows the most important redundant regions of all the four serovar of dengue virus.

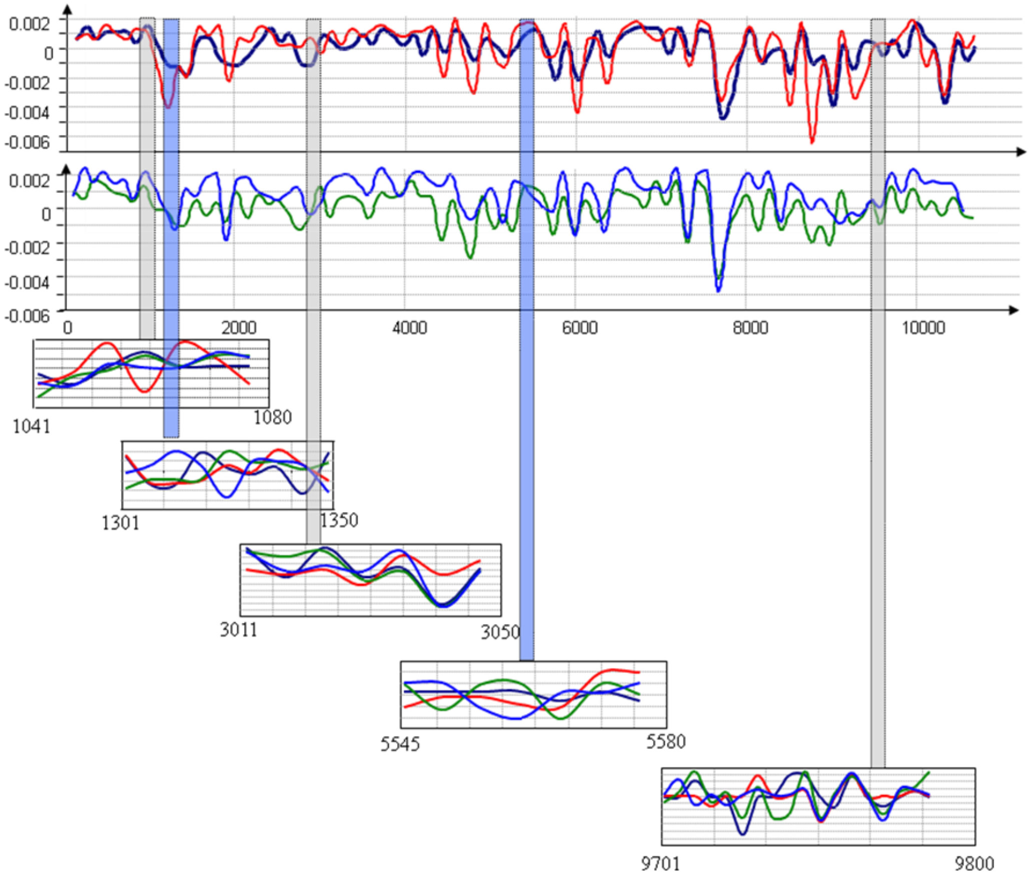


Figure 4.8: Highly redundant segments common to all the four serovar of dengue virus

Consider for example the genomic sequence DEN1 of dengue virus. To determine the information content in its sequence, first, the total number of occurrence of each member of the set {A, T, G, C} is determined in a given window (here the window size is 120). It is denoted as p_A, p_T, p_G and p_C for A, T, G and C respectively. Then, the information content (denoted by H_1) for each window is determined using modified Equation (4.2) as:

$$H_1 = (p_A \times \log(p_A)) + (p_T \times \log(p_T)) + (p_G \times \log(p_G)) + (p_C \times \log(p_C)) \qquad (4.11)$$

Now, consider a junk sequence with equal probability (0.25) of occurrence of all the four bases and with the same sequence length as that of DEN1. Then, for the same window size as above, the total number of occurrence of each member of the set {A, T, G, C} is determined and is denoted as q_A, q_T, q_G and q_C. The information content of this equiprobable junk sequence (denoting the introns) is given as:

$$H_{M1} = \left(q_A \times \log\left(q_A\right)\right) + \left(q_T \times \log\left(q_T\right)\right) + \left(q_G \times \log\left(q_G\right)\right) + \left(q_C \times \log\left(q_C\right)\right) \qquad (4.12)$$

Hence, by substituting Equations (4.11) and (4.12) into Equation (4.10), the redundancy factor (R_1) for DEN1 can be deduced as:

$$R_1 = 1 - \frac{H_1}{H_{M1}} \qquad (4.13)$$

Similarly, the redundancy factors for all the other strains of dengue virus can be determined. The result has been shown in Figure 4.8 above.

4.8 CpG Motifs in Viral ssDNA

In the DNA segments, a set of CG motifs constituted exclusively by short stretches of guanine (G) and cytosine (C) bases may prevail with an occurrence frequency of such CG nucleotides being higher than in other regions. Such CG-motif section is also called the *CpG island* where ‘p’ implies that C and G are connected by a phosphodiester bond.

The CpG islands, in essence are unmethylated regions that contain high concentration of C and G. The generally accepted definition of what constitutes a CpG island in a DNA sequence was proposed in [4.22] as being a 200-bp stretch of DNA with a (C + G) content of 50 % and an (observed CpG)-to-(expected CpG) ratio being higher than 0.6. A subsequent study [4.23] based on an extensive search on the complete sequence of human chromosome 21 and 22, stipulates that a DNA region with equal or greater than 500 bp having a GC-content exceeding 55 % and an (observed CpG)-to-(expected CpG) ratio of 0.65 could be more likely a true CpG island. The CpG sequences are relatively rare in human DNA but, are more commonly observed in the DNA of foreign organisms such as bacteria or viruses.

There has been a general research interest in understanding the presence of CpG oligonucleotide in such viral and bacterial genomes. Specifically, the human immune system has been studied as regard to the way it evolved in recognizing CpG sequences as early signs of infection and initiating an immune response on *ad hoc* basis. The extent of CpG abundance and/or deficiency specific to viral genomes of a certain species has also been of research interest. [4.24].

Thus, with reference to an ssDNA, the bioinformatic effort of interest in the present study refers to knowing the presences of CpG motifs along the genome sequence *via* compatible analytical and computational procedures. Identification and delineation of CpG islands in a test ssDNA can be done again by using the concept of entropy-based statistical divergence. Relevant details are presented below:

4.8.1 Locating the CpG Islands: Entropy-based Approach

Location of CpG islands in the test sequence implies finding the fragment of bases that constitute the motifs of CpG. In order to determine the presence of such islands, first the entropy-dictated by the occurrence statistics of three cases, namely, C-alone, G-alone and CG-jointly are determined. For this purpose, a “junk” sequence of length L bases is first constructed as follows: Considering a total sequence length of L nucleotide base locations, a uniformly-distributed set of L/4 random locations (epoch spaces) are generated and assigned for the base A (in the space set L). Likewise, three different ensemble sets of random locations are permutatively generated within the field L and assigned for C, G and T. Thus, the total length L bases is occupied randomly by A, T, G, C each with an equal probability of {$q_A = q_T = q_G = q_C = 1/4$). The sequence of length so generated can be dubbed as a "junk sequence" implying that the statistics of {A, T, G, C} does not bear any information (due to certainty of equal-probable occurrences of the elements in {A, T, G, C} consistent with Laplacian hypothesis on equally-likely occurrences of epochs). In contrast, considering an actual genomic sequence (such as a viral sequence), the elements of {A, T, G, C} would occur randomly with unequal probabilities as dictated by the encoded genetic information. That is, corresponding probabilities of occurrence of A, T, G and T, namely, p_A, p_T, p_G and p_C are such that, $p_A \neq p_T \neq p_G \neq p_C$; but, $(p_A + p_T + p_G + p_C) = 1$.

Now, taking a window across actual and junk sequences, the occurrence probability sets namely, {p_{C+G}, $p_{C,}$ p_G} and {q_{C+G}, $q_{C,}$ q_G} are determined for each window of nucleotides. Hence, the (observed CpG)-to-(expected CpG) ratio is specified by $R_{pw} = (p_{C+G})/(p_C \times p_G)$ and $R_{qw} = (q_{C+G})/(q_C \times q_G)$. Corresponding mutual entropy, say, in terms of Jensen-Shannon measure, namely, JS (Rpw, Rqw) is then determined *via* Equation (4.14).

$$(\mathrm{JS})_w = \left\{0.5 \times p_w \times \log_e (p/M)_w + 0.5 \times q_w \times \log_e (q/M)_w\right\} \text{ nats} \tag{4.14}$$

Pseudocode A describes the step-by-step approach pursued and the specific results obtained are illustrated in Figure 4.8.

Pseudocode A

Pseudocode describing computation of mutual entropy based identification of CpG islands in the ssDNA of the Parvovirus B19V

```
// Identification of CpG islands in B19V virus ssDNA

Initialize

Generate a junk random sequence of {A, T, C, G}
→   Generating junk random sequence of {A, T, C, G} of length 5000
      ← Junk random sequence represents a random sequence of {A, C, T, G) of length 5000 numbers such that, A = 1250, C = 1250, T = 1250, and G = 1250 epochs are generated putatively with equal probability of occurrence of A, C, T and G

// Step I

Calculate the occurrence probability of (C + G) per window-segment of 10 nucleotides in the junk sequence
→   In equally-spaced window size of 10 nucleotides, count total of (C + G) = $M_{C+G}$
    ← Then, the occurrence probability of (C + G): $q_{C+G} = M_{C+G}/10$
       ← Repeat this calculation over 500 windows that span the stretch of 5000 nucleotides

Calculate the occurrence probabilities of C and G per window-segment of 10 nucleotides in the junk sequence
→   In equally-spaced window of size 10 nucleotides, count C’s = $M_C$; and count G’s: $M_G$
      ← Then, the occurrence probability of C: $q_C = M_C/10$; and, the occurrence probability of G: $q_G = M_G/10$;
      ← Repeat this calculation over 500 windows that span the stretch of 5000 nucleotides


//Step II

Determine the ensemble average of ($q_{C+G}$, $q_C$ and $q_G$)↔$(q_{C+G})_{Av}$,$(q_C)_{Av}$ and $(q_G)_{Av}$
→   This refers to generating a number of junk random sequences of {A, T, C, G} of length 5000 and in each case iterating Step I, so that ensemble averages of the probabilities over the entire set of ensembles generated are computed to yield $(q_{C+G})_{Av}$,$(q_C)_{Av}$ and $(q_G)_{Av}$
```

```
// Step III

Determine the ensemble average of ($p_{C+G}$, $p_C$ and $p_G$)↔$(p_{C+G})_{Av}$,$(p_C)_{Av}$ and $(p_G)_{Av}$ corresponding to the actual {A, T, C, G} sequence of the ssDNA of B19V virus
        →    This refers to actual test sequence of B19V ssDNA
             sequence of length 5594 nucleotides
        →    Repeat procedures of Steps I and II to get ($p_{C+G}$, $p_C$
             and $p_G$)↔$(p_{C+G})_{Av}$,$(p_C)_{Av}$ and $(p_G)_{Av}$ corresponding to the
             actual sequence of the ssDNA of B19V virus


// Calculating the observed/expected probability ratios for junk
   and actual sequences

Define the (observed CpG)-to-(expected CpG) ratio as $R_{pw} = (p_{C+G})_{Av}/(p_C \times p_G)_{AV}$ and $R_{qw} = (q_{C+G})_{Av}/(q_C \times q_G)_{Av}$ for actual and junk sequences
        →    Determine $R_{pw}$ and $R_{qw}$ for each window-segment

// Determining the mutual entropy, say Jensen-Shannon (JS)
   measure for each window
 →JS (Rpw, Rqw) is then determined via equation (4.13).

Print
            ←  Window segments across 0 to 5000 nucleotides versus
               the JS measures obtained

Plot
    → Window segments (indexed as 0 to 5000) versus JS measure
      estimated in each segment is plotted as an x-y graph

Result/Output
    →  (Figure 4.8)
            ←  CpG islands correspond to window-segments wherein JS
               values namely, JS(Rpw,Rqw), are seen clustered and
               some of such values are in excess of 0.6

End
```

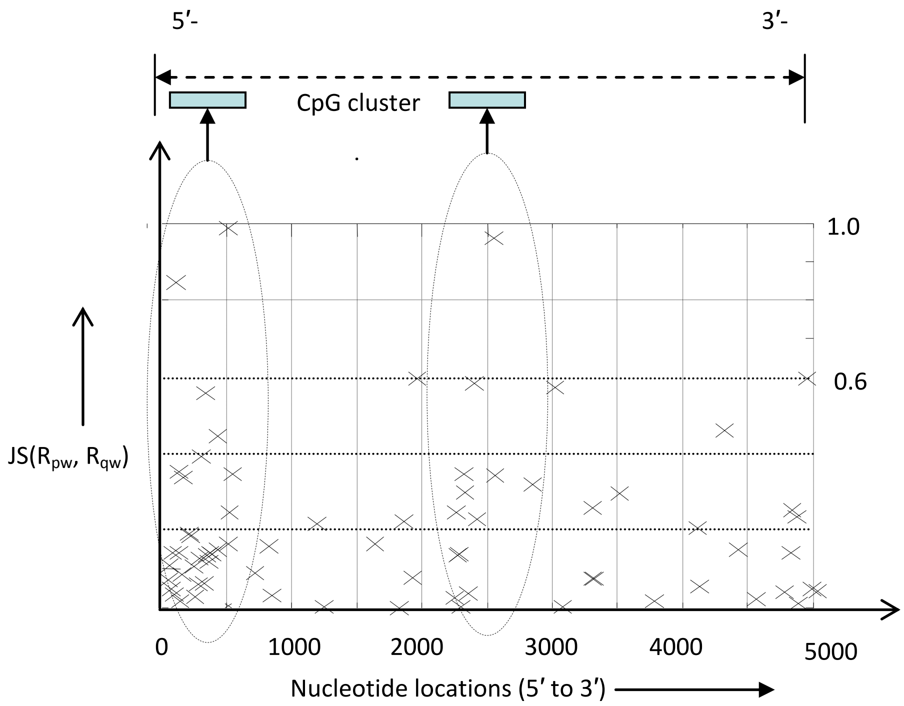


Figure 4.8: Locating CpG segments in the test sequence: The computed results marked as (×) correspond to normalised JS-measure evaluated with Rpw and Rqw ratios (described in the text). The clustered regions plus computed measure exceeding 0.6 depict plausible CpG islands as shown.

The concept of entropy is thus exercised in locating CpG island segments by considering the divergence of (C+G, C and G) populations in the test sequence *versus* the statistics of this population set in a simulated junk sequence, as shown in Figure 4.8.

4.9 Salient Results and Discussions

The salient discussions in this chapter thereof are concerned with the following:

- Various entropy concepts and statistical ordering of the residue structures of the sequence, so as to identify the parametric as well as the non-parametric divergence matrix compatible for discriminating informative and non-informative sub-segments in a sequence and elucidating similar/dissimilar features across a set of sequences

- Finding splice-junctions between codon-noncodon segments, specifically, to describe the fuzzy transitions at the splice junctions in terms of a spatial jitter algorithm

- Predicting splice-junctions in viral sequences is elaborated. The possibility of aberrant splice-junctions appearing in viral sequences as a result of mutation is indicated with DEN1 virus as a case study example

- Extending the concept of information-theoretic description of genomic statistics in terms of Shannon's information redundancy factor (R)

- The efficacy of information-theoretic approach to identify CpG islands is also presented. Relevant computational methodology and results are presented with respect to Parvovirus B19V ssDNA

Relevant results on DEN1-DEN4 viral sequences are presented in Chapter VII.

4.10 Closure

This chapter is written to outline the entropy considerations useful in deducing the segmental features of genetic statistics in genome structures. Relevant applications to viral genomes form the theme of the study addressed.

CHAPTER V

ENERGETICS-BASED VIRAL GENOMIC SEQUENCE ANALYSIS

5.1 Introduction

The most prevalent viruses in nature are single-stranded (DNA or RNA) viruses with the genetic material encapsulated in icosahedral-shaped capsid proteins [5.1]. They cannot reproduce by themselves, but infect the host cells with their genetic materials, enabling the (host) cellular machinery to produce more viruses. The viral ssDNA or ssRNA assumes invariably a hairpin format (for structural stability as has been explained in Chapter III) and this hairpin form consists of a base-paired stem-structure plus a loop sequence having unpaired or mismatched nucleotides. Knowing the profile of hairpin structures and the loops and bulges in viral genome is largely pertinent in: (i) Understanding virus replication process [5.2] and (ii) in drug synthesis applications, where a relevant compound being sought may act as a binding agent in a specific ss-DNA/-RNA inhibiting the replication of certain viruses [5.3].

As regard to the framework of overall folding of DNA structure into hairpin formats, exclusive research in bioinformatic perspectives is often needed (adjunct to wet studies) so as to ascertain their energetics profiles responsible for the folding and stability of the genome backbone. While a microbiological description of DNA hairpin bends emphasizes the biological importance of such bends, a distinct pursuit of research deliberates the physics of thermodynamics on the structural aspects of the folded single-stranded viral genomes [5.4], yielding details on the associated energetics profile of the genome structure. In this thesis, the energetics profile of two viruses has been discussed. The energetics profile of parvovirus B19V along with its associated entropic features is described in this present chapter. The application of the energetics profile of the various serovar of dengue along with its entropic features and data from spectral analysis in cohesively determining unique characteristics of each serovar has been discussed in Chapter VII.

The scope of the present study emphasizes the following objectives in their bioinformatic contexts: (i) Constructing an analytical framework to elucidate the thermodynamics-based energetics profile of a test ssDNA/ssRNA *vis-à-vis* its hairpin format; (ii) determining the entropy profile of the test ssDNA and characterize the underlying Shannon information of genetic expression in the strand; (iii) delineating codon and noncodons sections in the single-strand using the segmented features of the associated entropy; and (iv) elucidating the stability aspects of ssDNA *via* energetics and entropy details

5.2 Energetics of a Genomic Sequence: An Overview

As stated earlier in Chapter IV, biothermodynamics and entropy are like Siamese twins. One of the main steps in the viral life cycle which is required for replication of the virus is genome ejection into the host cell. The ejection of the viral genome from the capsid is due to very high internal pressure due to the electrostatic forces of the nucleotides [5.5], [5.6]. This pressure and electrostatic charge is partially responsible for the delivery of the viral genome into the host cell, thus making it central in the infection process. Also, as the overall folding of single-stranded DNA/RNA into hairpin structure (which is very important for its replication) and its stability depends on the energetics associated with the virus, the detailed energetics-profiling is imperative in determining the particular characteristics of the virus in question.

The general prospects of the analyses indicated above and related computations would eventually illustrate the integration of stabilized, hairpin-folded viral DNA structure into a host genome in the perusal of replication processes.

The bioinformatic contexts as above can be expanded specifically to include the following:

- The site in the ssDNA/ssRNA at which the hairpin bend would take place
- Possible sites in the folded structure where bulging (bubbles) may prevail
- The integrity of base-pair matching within the hairpin-folded sections of the loop and the stem
- The stability of folded-format of the ssDNA/ssRNA in terms of the associated energetics
- Correlating energetic and entropy profiles of the folded ssDNA
- Eventual use of genetic information (knowledge) in the folded structure of the ssDNA. (For example, understanding the mechanism of hairpin-bend (folding) implications in viral ssDNA *vis-à-vis* using relevant details for rational design of vaccines.)
- Delineating codon (or coding DNA sequence, CDS) and noncodon segments in the test sequence. While CDS codes for a protein's amino acid sequence and noncodon segments do not. Examples of such non-coding segments (or “junk” codons) are introns, repetitive sequences and sequences between genes.

5.3 The Watson-Crick Model Base-pairing and Energetics of ssDNA/ssRNA

The RNA (transcribed from the double-stranded DNA) and ssDNA sequences tend to become more compact by “folding” or “bending” themselves into a stabilized hairpin structure (mostly towards 3’ end) *via* nucleotide base-matching set by *Watson Crick (WC) pairing* of A ↔ T and G ↔ C. The RNA hairpin forms are widely addressed in the literature [5.7 - 5.9]. Analogous to RNA hairpin, it is hypothesized in this study that the hairpin structure of viral ssDNA is made of turn-around loop with the closing base-pairs and a stem. In addition, a bulge on a strand and/or an internal loop on both strands of the hairpin can be formed due to unpaired nucleotides [5.10]. However, associating assertively such a bulge and/or an internal loop composition with a viral ssDNA is an open-question for investigation.

The stem-part plus loop constitutes the so-called *stem-loop* (SL) structure. The stem primarily consists of WC base-pairs (bp) formed between two anti-parallel structures of the hairpin, 5'-through 3'-end. The stability (as indicated to prevail in RNA hairpins) is due to the so-called TNCG tetra-loop composition (where, N depicts any of the four bases A, T, G, C). Typically, **c**TNCG**g** is an example notation that includes the loop with a closing base-pair (**c** and **g** denoted by lower case bold font). The closing base-pair is the first base-pair set next to the loop at the commencement of the stem. The hairpin-loop sequences preferentially decide on the type of closing base-pairs. For example, the UNCG tetra-loops in RNAs prefer a **cg** closing base- pair over a **gc** version (as decided by nearest-neighbor effects).

Normally, a region of unpaired nucleotides may exist at the apex of the hairpin loop, which serves as the region where the directionality of the backbone reverses making the two anti-parallel strands of the stem. In a RNA hairpin a minimum of three nucleotides can make a bend or turn (consistent with stearic repulsion); however, loops of four nucleotides (mentioned above as tetra-loops) are more common. The hypothesis of the present study conforms to the both tri- and/or tetra-loop characterizations applied to viral ssDNA genomes.

Considering the viral ssDNA, it is yet to be ascertained confirmedly of stable and unstable aspects of hairpins and the existence of combinatorial base nucleotide selection in forming tetra- and/or tri-loops associated with closing base-pairs. Relevant folding kinetics expressed in terms of energetics and entropy consideration could be determinants of stability/instability in the folded viral ssDNA messages as discussed in Section 5.5.

5. 4 Free-energy Thermodynamics of nearest-neighbor (NN) WC Nucleotides

In addition to the favored (neg)entropy enabled by WC-pairing toward stability of the hairpin structures formed, such (stability) dynamics also relies on free-energy minimization specified by *nearest-neighbor* (NN) parametric attributes of base-pairs in the test sequence. That is, the stability in question conforms to the rules stipulated by each base-pair depending only on the most adjacent pairs with the associated total free-energy being the sum of each contribution of the neighbors [5.11]. The underlying considerations are as follows:

WC base-pairs are significant motifs, whose thermodynamic aspects can be well represented by the NN model that indicates the stability of the base-pairs being dependent on the identity of the adjacent pairs. Known generally as *individual nearest neighbor* (INN) model, it implies a preferential stacking of energetically conducive pairs with loop-initiation leading to an eventual hairpin structure. Totally, the associated conformational free-energy is constituted by paired and unpaired nucleotide stacking in the bend as well as by the nucleotides at the loop.

The free-energy increments of the base-pairs in the sequence are counted as stacks of adjacent pairs. For example, the consecutive CG base-pairs are worth about (− 3.3 kcal/mol) [5.12 - 5.15]. The loop-region formed normally has unfavorable increments called *loop initiation energy* that largely reflects an entropic cost expended in constraining the nucleotides within the loop. For example, the hairpin loop made of four nucleotides may have an initiation of energy as high as + 5.6 kcal/mol. Mostly the unpaired nucleotides in the loop contribute favorable energy increments. From the literature indicated above, a set of approximate conformational free-energy can be gathered on the basis of nearest-neighbor considerations.

Suppose the A, C, T, G bases as before are translated with 0, 1 scoring (with 1 indicating WC-pair and 0 denoting NWC-pairing). Hence, by considering − 3.3 kcal/mol as zero reference (depicting the lowest energy level), Table 5.1 is constructed, which depicts the data on relative energy values compiled from [5.12 - 5.15] for NN composition of four adjacent bases as shown. That is, indicated in Table 5.1, are four adjacent matched or unmatched WC-pairs along the hairpin-stem structure. By designating the matched pairs (WC) as 1 and the unmatched pairs (NWC) as 0, the details on possible four adjacent pairs *versus* the associated free-energy values are as listed in Table 5.1.

Table 5.1: Depiction of four WC- and/or non-WC neighbors and the corresponding (approximate) relative levels of free-energy

Note:

- Scoring scheme: WC neighbor- 1; NWC neighbor- 0
- NN-R: Neighbor on the right-side of centre element; CE: centre element; and, NN-L: Neighbor on the left-side of centre element
- EV: Energy values (in kcal/mol specified relative to – 3.3 kcal/mol depicting the lowest energy level and taken as zero reference level)

Four possible neighbors of WC (1) and non-WC (0) pairs				
NN-R	CE		NN-L	EV
1	1	1	1	0
0	1	1	1	0
1	1	1	0	0
0	1	1	0	0
1	1	0	1	1.2
1	1	0	0	2.2
0	1	0	1	2.7
0	1	0	0	6.6
1	0	1	1	1.2
1	0	1	0	2.2
0	0	1	1	2.7

0	0	1	0	6.6
1	0	0	1	4.3
1	0	0	0	8.9
0	0	0	1	8.9
0	0	0	0	10.0

Using the relative free-energy data-set shown in Table 5.1 as the metric and resorting to the sliding-window method, the profile of free-energy variation across the test sequence can be determined. Pseudocode A presents the steps for determining the energetics-profile of Parvovirus B19V.

Pseudocode A

Pseudocode describing computation of energetic feature of nearest-neighbors (bases) in the DNA sequence of the Parvovirus B19V

```
Initialize
//   Computation refers to ascertaining the hairpin-bend
     structural features of the ssDNA strand of B19V virus


Input
         →  In this computation, only nucleotides from base-number
            5401 to 5594 (at the 3′-end) are considered
              →  Relevant test sequence of {A, T, G, C} from 5′- 3′
                 is posted as a string
              →  The test sequence is converted into a 1 × n matrix
                 with each letter of {A, T, G, C} representing
                 matrix element
                    ←  This is named as window 1 (W1)
```

```
          → The reverse (3’-5’) of the test sequence is
            considered (omitting the dangling end) - This is
            named as window 2 (W2)
Compare the elements residing in W1 and W2
          ←   Each element in W1 matrix is compared against the
              W2 matrix in the same column
                → Comparison implies looking for (C ↔ G or
                  G ↔ C) or (A ↔ T or T ↔ A) match

Perform scoring in binary format
  If (C ↔ G or G ↔ C) or (A ↔ T or T ↔ A) are seen across W1 and
  W2, the score is indicated as 1
  or else, the score is set as 0
        ← This scoring is continued across the window-segments
          being compared and 1’s and 0’s generated is stored in
          a $1 \times n$ matrix

Construct the nearest-neighbor (NN) concept based energetic
profile for the nucleotide stretch 5401 to 5594
          ←  Energetic values are assigned by considering two
             centre elements (of 1-0 map) and one neighbor on
             each side of this central pair. The allocation is as
             in Table 1.
          →  The energy value assigned is stored in a matrix


Print
        ← Window segment 5403 to 5593 versus the EV obtained


Plot
  → Window segments (indexed as 5403 to 5593) versus EV value
    estimated in each segment is plotted as an x-y graph


Result/Output
  → (Fig. 5.2b)

End
```

5.5 Application of Entropy and Energetics Profiles of ssDNA

The use of entropy-based segmentation method and energetics method are addressed side-by-side in this section as regard to finding the specific structural details of genomes. It refers to characterizing the sub-regions of genomic sequences such as loops and bulges. A brief description about the application of entropic considerations to a viral ssDNA is discussed below followed by the results due to both methods being compared.

5.5.1 Entropy Considerations

Pertinent to a hypothetical hairpin structure shown in Figure 5.1, suppose the occurrence probability of Watson-Crick (WC) base-pair matching, that is, A matching T (A ↔ T) or G matching C (G ↔ C) across the inverted stem parts is designated as ‘p’; and the probability of occurrence of corresponding mismatches is designated as ‘q’ (so that, $q = p - 1$). Between the event spaces of p and q, exists a cross-entropy (or mutual information) in Shannon’s sense. Such cross-entropy details can be assessed by understanding the probabilistic features of the test ssDNA sequence bearing the inherent genetic information. That is, viewed in terms of the probabilistic occurrence profile of {A, C, T, G}, the associated Shannon information or entropy can be ascertained as follows:

Considering the test hairpin structure of an ssDNA, its WC-pair statistics in the stem region (along with any nicked base-pairs that may exist in the stem and/or in the loop) is first determined. In essence, relevant analytical and computational pursuits involve elucidating the statistical divergence of WC *versus* non-WC (NWC) pairs occurring in the sequence pattern; that is, a strategy in terms of the statistical occurrence profiles that segregate WC and NWC entities is evolved with the objective of primarily locating the site at which bending takes place. By considering the nature of WC/NWC pairs manifesting as matched or unmatched inversions of {A, T, G, C} on both sides of the site where hairpin bending occurs (as illustrated in Figure 5.1), the associated analytical framework can be conceived on the basis of statistical divergence of WC *versus* NWC populations along the hairpin structure. That is, the extent of bases forming WC pairs across the stem region is specified by a (statistical) "distance" of WC entities with respect to (the statistics of) those bases, which do not pair as WC entities (that is, remain as NWC pairs).

The following sliding-window approach is indicated to determine p and q values consistent with the definitions of p and q mentioned above. Suppose the WC base-pairs (AT, CG, GC, TA) are each assigned a fitness 1, and all other pairs are treated as mismatches with an assigned fitness of 0. Apart from the WC-mirrored sets, namely, A ↔ T and G ↔ C, the set of pairs {G ↔ T and T ↔ G} is also regarded as stable pairs (Weise et al., 2008); hence, in all, the set {A ↔ T, G ↔ C, G ↔ T} is designated as a set of *canonical base-pairs* that can be accounted for in assigning the fitness 1.

Suppose the total length of the sequence is constituted by L bases. Considering a hypothetical hairpin structure, a sliding-window of size 'w' (containing a total of w pairing and non-paring nucleotide counts) is specified along the stem and its reversed part (of the hairpin bend on the 3'- side) as shown in Figure 3. Within each window, the count of 1's, (meaning matching WC-pairs) is denoted as 'm'; likewise, the count of non-matching pairs is denoted as 'n', so that, (m + n) = w; hence, (m/w) = p and (n/w) = q for the window-segment chosen. Thus, for each window size w sliding across the hairpin structure, the set $\{p, q\}_W$ is determined.

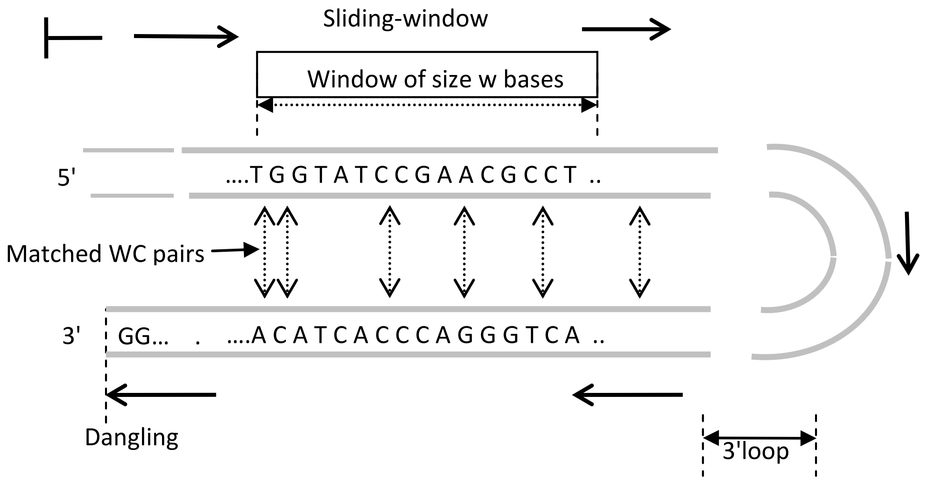


Figure 5.1 Illustration of sliding-window procedure to compute p and q values across a hypothetical hairpin structure

The statistical-distance that measures the cross-entropy (in each window segment, w) can be obtained *via* statistical divergence metrics by knowing p_w and q_w values (namely, the estimated p and q values in each window segment). Expressing in nats unit of information (negentropy), the statistical distance measures (for the window of size w) indicated above can be defined explicitly as follows:

$(KL)_w = \{p_w \times \log_e(p/q)_w + q_w \times \log_e(q/p)_w\}$ nats

(5.1)

$(JS)_w = \{0.5 \times p_w \times \log_e (p/M)_w + 0.5 \times q_w \times \log_e(q/M)_w\}$ nats (5.2)

where, $M = 0.5 \times (p_w + q_w) = 0.5$.

$(B)_w = -\log_e(\rho_w)$ nats

(5.3)

where $\rho_w = (p_w \times q_w)^{1/2}$

(Apart from KL-, JS- and B-measures, there are also a number of other statistical divergence metrics available as reported in [5.16], [5.17]. They can also be used in the algorithmic exercise in hand).

The analysis pertinent to a given test hairpin structure as above, thus evaluates the statistically co-evolving positions of WC pairs across the folded structure and decides the associated cross-entropy or mutual information. The computation involves divergence measures in terms of the {p, q}-based discrimination of WC and NWC. The complete procedure for determining the entropic profile is summarized in Pseudocode B.

Pseudocode B

Pseudocode describing computation of relative entropy feature of the DNA sequence of the Parvovirus B19V

```
Initialize
//   Computation refers to elucidating the relative entropy
     feature of the DNA sequence of the Parvovirus B19V

Inputs
   → Test sequence (of B19V virus) made of {A, T, G, C} is
     posted as a string of nucleotides numbered from 1 to 5594
     from 5′-end to 3′-end
        → This test sequence is converted into a 1 × n matrix
          with each letter of {A, T, G, C} representing the
          matrix element

Define the sliding window
   → Select a window-size (10 nucleotide bases)
          → An index is assigned to flag the start of the
            window and another is assigned to flag closing the
            window
                 → A pair of windows are defined:
```

```
← Window 1 (W1) is on main string at 3′-
  end with the start-window direction
  indexed along 5′ to 3′
← Window 2 (W2) is an auxiliary
  (sliding- window) constructed with
  start end in reverse direction sliding
  from 3′ to 5′

Perform widow-sliding along the sequence
    → Window lengths of W1 and W2 are varied by
      setting start and end indices as necessary.
    → Initial size is set as 5 and in linear
      multiples of 5.

Compare the elements residing in W1 and W2
    ← Each element in W1 matrix is compared against the
      W2 matrix in the same column
        → Comparison implies looking for (C ↔ G or
          G ↔ C) or (A ↔ T or T ↔ A) match

Perform scoring in binary format
  If (C ↔ G or G ↔ C) or (A ↔ T or T ↔ A) are seen across W1 and
  W2, the score is indicated as 1
  or else, the score is set as 0
      ← This scoring is continued across the window-segments
        being compared
          ← In each W1-W2 comparison, 1′s are counted and
            added; likewise 0′s are counted and added

Store the scores
  → Score values for each window length are stored

Perform computation of the probabilities of occurrence of 1′s and
0′s in each of the configured/compared W1-W2 window segments
    → For each window segment of 2 × 5 characters (and its
      multiples), the probability (p) of 1 is computed as
      follows:
```

$$p = [\text{Total counted score of 1's}] / [(2 \times 5 \times N\ (=1, 2, 3,..., 560)]$$

```
    → Likewise, the probability of 0 (q) is:
```

$$q = [\text{Total counted score of 0's}] / [(2 \times 5 \times N\ (=1, 2, 3,..., 560)]$$

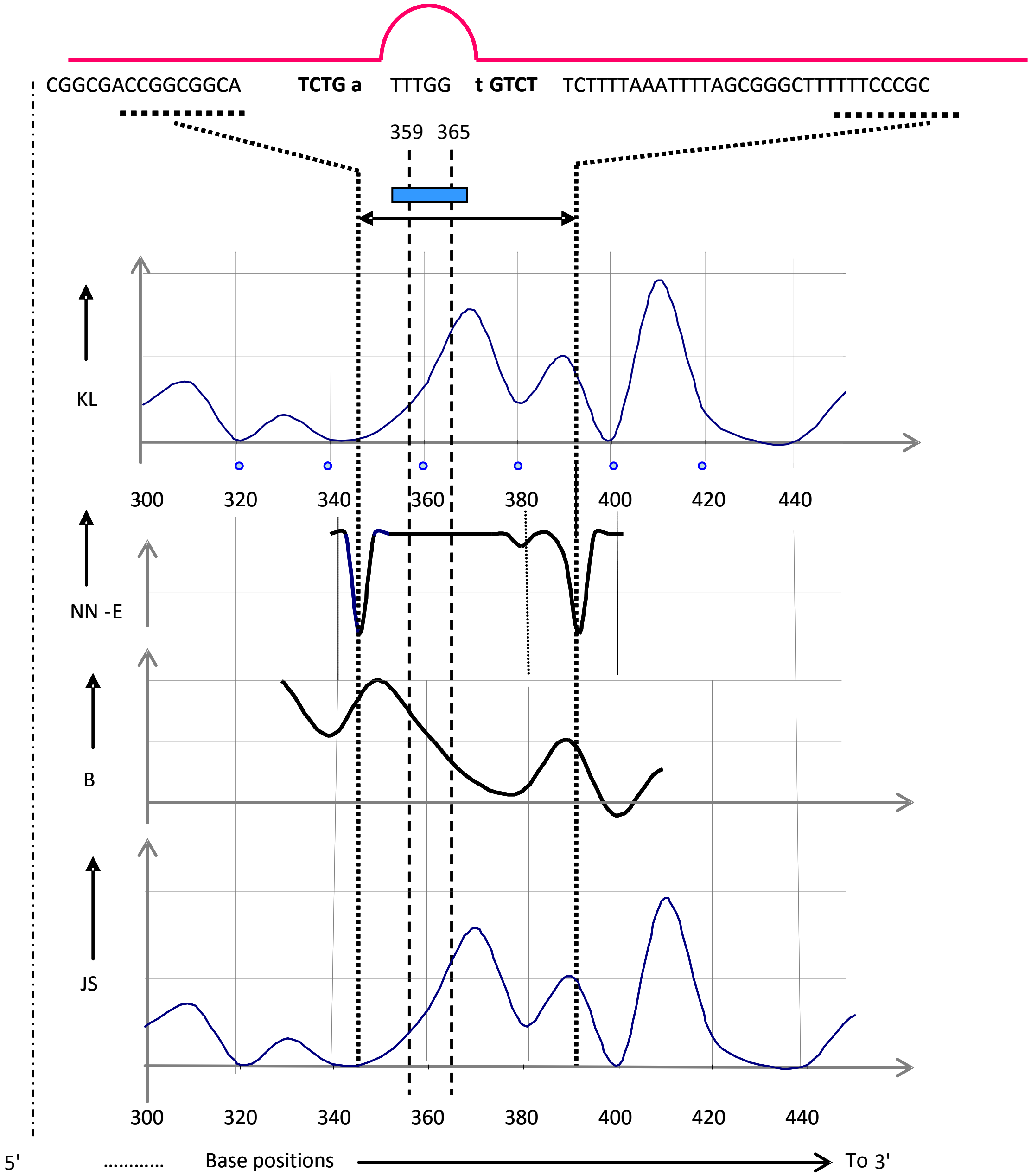
```
    ←or, q = (1 - p)
```

```
Compute the statistical divergence on: p versus q
  → The statistical divergence (distance) between the main and
    the sliding window is determined via  say, Kullback-Leibler
    (KL) measure (Equation 5.1)
      ← KL divergence estimation of each window segment is
        done with the following algorithm:

        $\leftarrow \mathrm{KL} = [\mathrm{p}\log_e(\mathrm{p}/\mathrm{q}) + \mathrm{q}\log_e(\mathrm{q}/\mathrm{p})]$ nats
      ← Similarly JS- and B-divergence estimations of each
        window segment is done with the s algorithms of
        Equations (5.2) and (5.3)

Print
  → Window (of length 2 × 5) segment is indexed as 1 to 560
    commencing at 5′-end and terminating at 3′-end along the x-
    axis; and the KL-measure estimated in each segment is
    depicted along y-axis for each x-axis value

Plot
  → Nucleotide base position-segments (indexed as 1 to 560
    across 5′-end to 3′-end) versus KL-measure estimated in
    each segment is plotted as an x-y graph

Result/Output
  → (Fig. 5.2a)

End
```

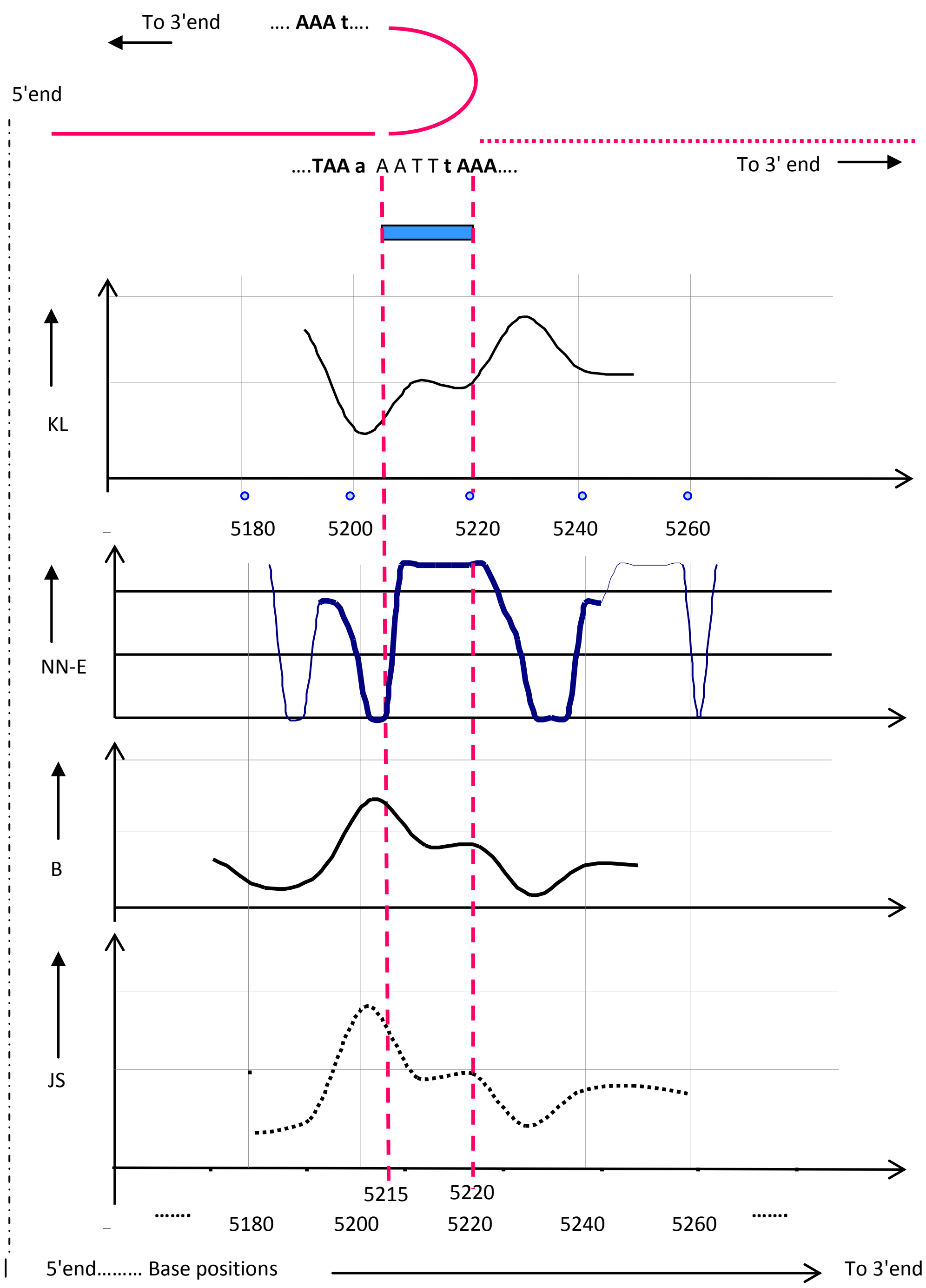
5.5.2 Result of Application of Entropy and Energetics Profiles of a ssDNA

Inspired by existing investigations on RNA folding [5.10] and results available thereof, the present study was motivated to provide an exclusive entrée to viral ssDNA hairpins. It is a bioinformatic attempt to ascertain various structural, kinetic and thermodynamic aspects of a test viral ssDNA, so as to understand the stability and control of the underlying gene expression.

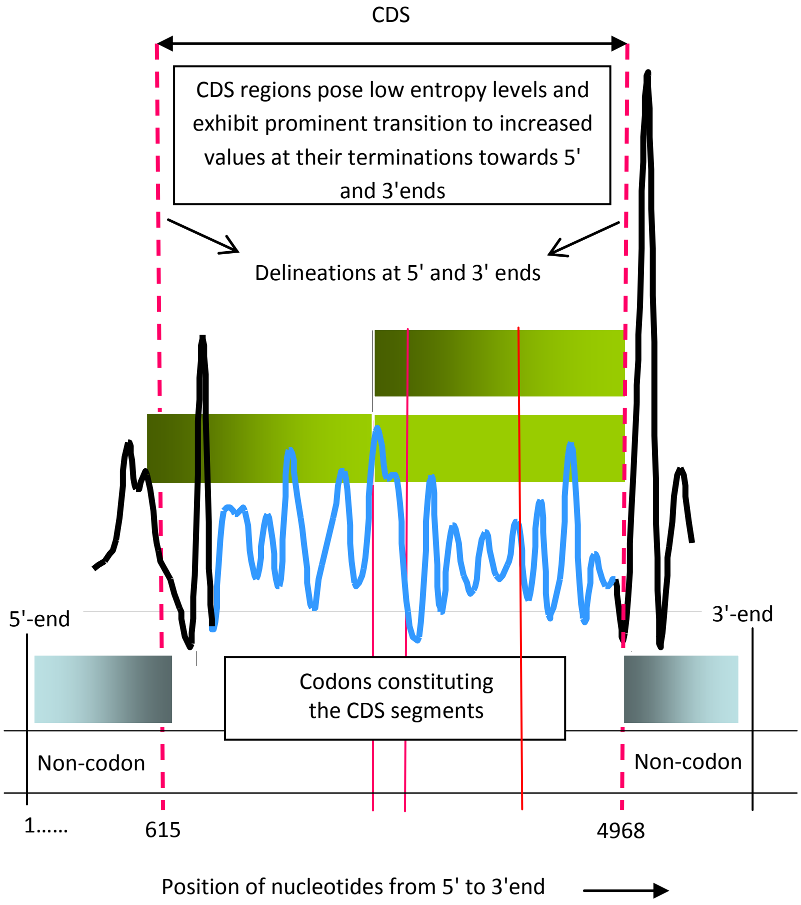
RNA sequences, (which are single-stranded) have been widely addressed in the literature [5.10] as regard to their structures, kinetics, thermodynamics and biological functions. Specifically, how a RNA folds back on itself (forming a complex structure) has been studied in terms of the resulting hairpin composed of a stem-part (with WC base pairing) and a loop-section, wherein the backbone reverses its directionality. Relevant studies also include descriptions of the hairpin structures with the associated diversity in their stem, loop and closing base-pairs. In relevant contexts, the kinetics and thermodynamics of hairpin-folding, folding transition-states and the co-operativity of folding have been investigated. However, similar information on viral ssDNA is rather sparse. Relevant outcomes of the study are presented in Figures 5.2 and 5.3; and, the results obtained and the associated inferences are detailed below.


CGGCGACCGGCGGCA
TCTG a
TTTGG
t GTCT
TCTTTTAAATTTTAGCGGGCTTTTTTCCCGC
359
365
KL
300
320
340
360
380
400
420
440
NN -E
B
JS
300
320
340
360
380
400
420
440
5'
....
Base positions
To 3'

(a)


To 3'end
.... AAA t....
5'end
....TAA a A A T T t AAA....
To 3' end
KL
5180
5200
5220
5240
5260
NN-E
B
JS
5215
5220
....
5180
5200
5220
5240
5260
....
5'end......... Base positions
To 3'end

(b)


CDS
CDS regions pose low entropy levels and exhibit prominent transition to increased values at their terminations towards 5' and 3'ends
Delineations at 5' and 3' ends
5'-end
3'-end
Codons constituting the CDS segments
Non-codon
Non-codon
1……
615
4968
Position of nucleotides from 5' to 3'end

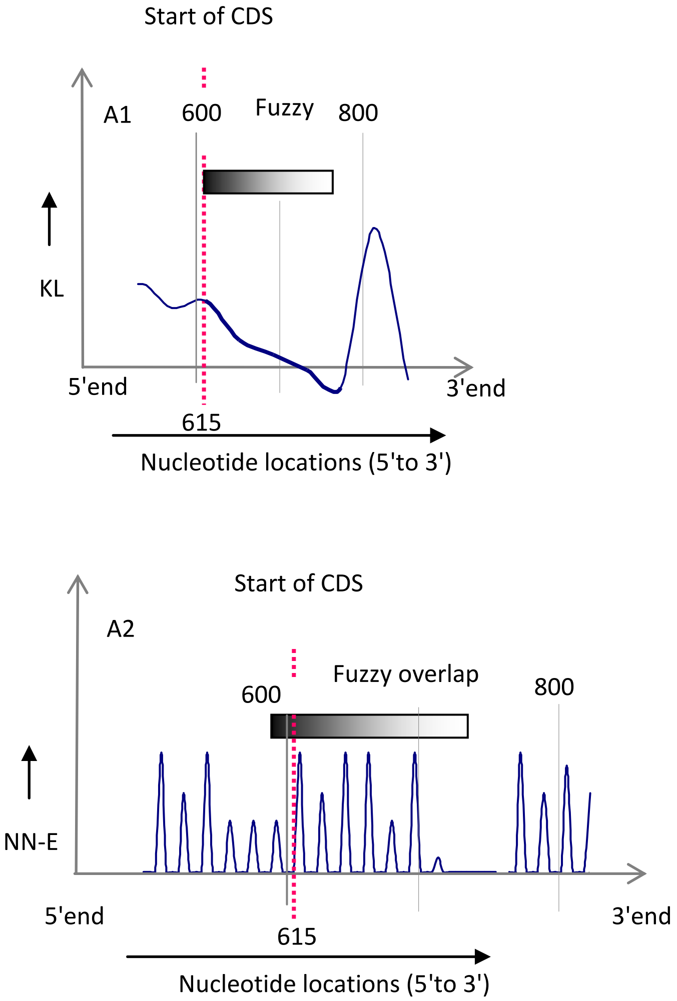
(c)

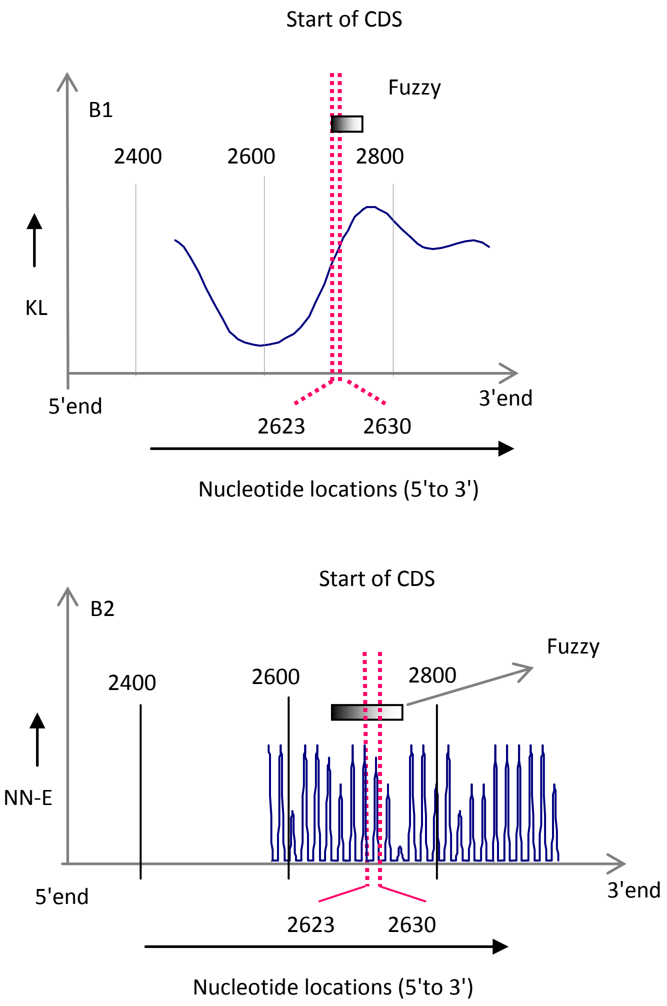
Figure 5.2: Illustrations of loop and bulge formations

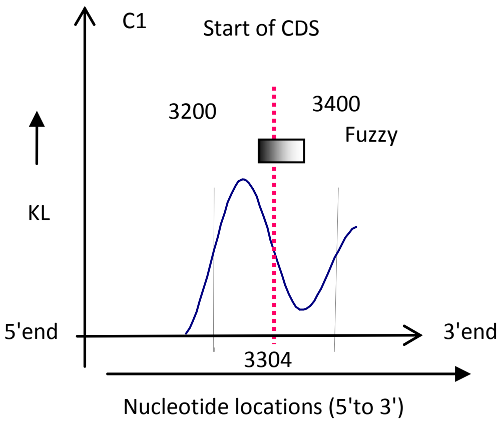
(a) Formation of a loop (bulge) at the 5′-end: The normalized results plotted conform to entropy measures of: KL-, JS- and B-metrics as well as NN-based energetics values (NN-E) evaluated in the vicinity of 5′-end. (The sequence of bases and base locations indicated at the closing-ends of the bulge are taken from the GenBank data [5.18] on the test genome)

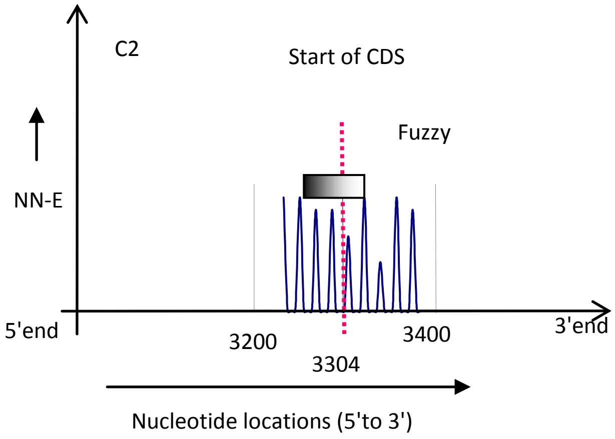
(b) Formation of a hairpin-loop at the 3′-end: The normalized results plotted conform to entropy measures of: KL-, JS- and B-metrics as well as NN-based energetics values (NN-E) computed in the vicinity of 3′-end. (The sequence of bases and base locations indicated at the closing-ends of the hairpin-loop are taken from the GenBank data [5.18] on the test genome).

(c) Overall representation of CDS in the test sequence. (Details as in Figure 5.3)


Start of CDS
A1
600
Fuzzy
800
KL
5'end
3'end
615
Nucleotide locations (5'to 3')
Start of CDS
A2
Fuzzy overlap
600
800
NN-E
5'end
3'end
615
Nucleotide locations (5'to 3')


Start of CDS
B1
Fuzzy
2400
2600
2800
KL
5'end
3'end
2623
2630
Nucleotide locations (5'to 3')
Start of CDS
B2
Fuzzy
2400
2600
2800
NN-E
5'end
3'end
2623
2630
Nucleotide locations (5'to 3')


C1
Start of CDS
3200
3400
Fuzzy
KL
5'end
3'end
3304
Nucleotide locations (5'to 3')


C2
Start of CDS
Fuzzy
NN-E
5'end
3200
3400
3'end
3304
Nucleotide locations (5'to 3')

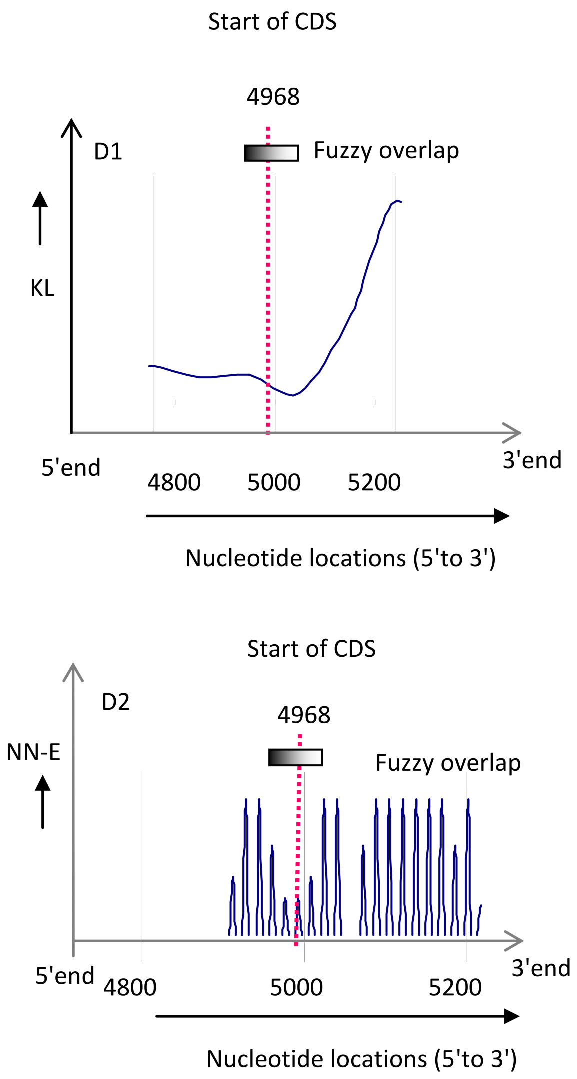


Figure 5.3 Locating the CDS segments in the test sequence: Delineation of codon/noncodon boundaries.

The set of normalized results plotted conform to entropy measures of KL-, JS- and B-metrics and NN-based energetics values (NN-E) computed across the stretch of 5′-end to 3′-end. (Note: (i) The sequence of bases and base locations indicated are transition regions as per GenBank data [5.18] on the test genome); and, (ii) at each transition site, the details are enlarged and presented as Figures A1 through D2 where the transitions shown are overlaps of codon and noncodon denoting fuzzy transitions (emphasized as shaded blocks).

A1, B1, C1 and D1: KL-measure based data showing entropy changing from low-to-high or high-to-low at the transition sites.

A2, B2, C2, and D2: NN-energy based data showing energy level changing from low-to-high or high-to-low at the transition sites.

Summary of results is as follows:

(i) The feasibility of loop (bulge) formation at 5′-end with inverted palindromes (flanking the bulge) is analyzed using the statistical distance (divergence) measures computed with WC matching and WC non-matching probabilities across the stem prior to and after the bulge. Figure 5.3(A1) illustrates the results obtained.

(ii) The divergence between WC matching pairs *versus* non-matching pair across the hairpin stem at 3′ end is determined and the results are illustrated in Figure 5.3(B1).

(iii) The codon and noncodon sections in the test sequence are delineated to ascertain the CDS of the sequence. Again, the occurrence statistics of {A, T, G, C} is discriminated against that of a junk sequence *via* divergence measures. Figures 5.3(C1) and 5.3(D1) illustrate the outcome.

(iv) The variation of (scored) energy levels along the sequence is ascertained as illustrated in Figures 5.3(A2) 5.3(B2), 5.3(C2) and 5.3(D2). In each case, the existence and extent of minimum energy levels depicts the stability feature.

5.5.3 Inferential Remarks on the Results

With the results described above, the following inferences can be made:

- As regard to Figure 5.2 (a), where the formation of a bulge at 5′-end is indicated, the GenBank data [5.18] on the test sequence implies the bulge being at the base positions 359 to 365. The present analysis shows a well-defined, almost symmetrical minimum energy (NN-E) sites on both sides of the bulge site. Also, the entropy feature (especially the B-measure) also indicates symmetrical maximum entropy levels abetting the bulge bilaterally. Thus, the algorithms indicated and computations performed enable finding of loop/bulge formation at the 5′-end with details on the relevant site in conformance with GenBank data

- Likewise, considering the 3′-end, GenBank data projects the hairpin bend at 5215-5220. The computed results of the present method again clearly show this loop sites with two minimum energy (NN-E) valleys on its sides in Figure 5.2(b). The entropy features also provide the information on the loop-site with KL-, JS- and B-measures changing distinctly from low-to-high or high-to-low values.

- Considering Figures 5.2(c) and Figure 5.3, it can be inferred that the procedure of this study advocated again gives confirming results on codon-noncodon transitions along the test sequence. The results are validated with GenBank data. However, unlike the crisp transition locations specified in GenBank data, the delineating locales are rather fuzzy. Such observations are consistent with the results in [5.19].

5.6 Conclusion

In a nut-shell, pursued in the underlying research of this chapter are simultaneous considerations of Shannon-entropy and thermodynamics-related energetics features across a test, single-stranded viral genome. In Chapter VII, along with these two approaches, spectral domain analysis for ssRNA has also been pursued. The eventual scope is to uniquely identify and formally distinguish the characteristic pattern(s) of the test ssDNA/ssRNA. The method pursued thereof enables robust determination of the sites-of-interest in the genome pointing out regions of possible loops and/bends that may occur toward stability. Use of relevant analyses is viewed in the context of possible vaccine-designs.

CHAPTER VI

FOURIER SPECTRAL CHARACTERISTICS OF VIRAL GENOMES

6.1 General

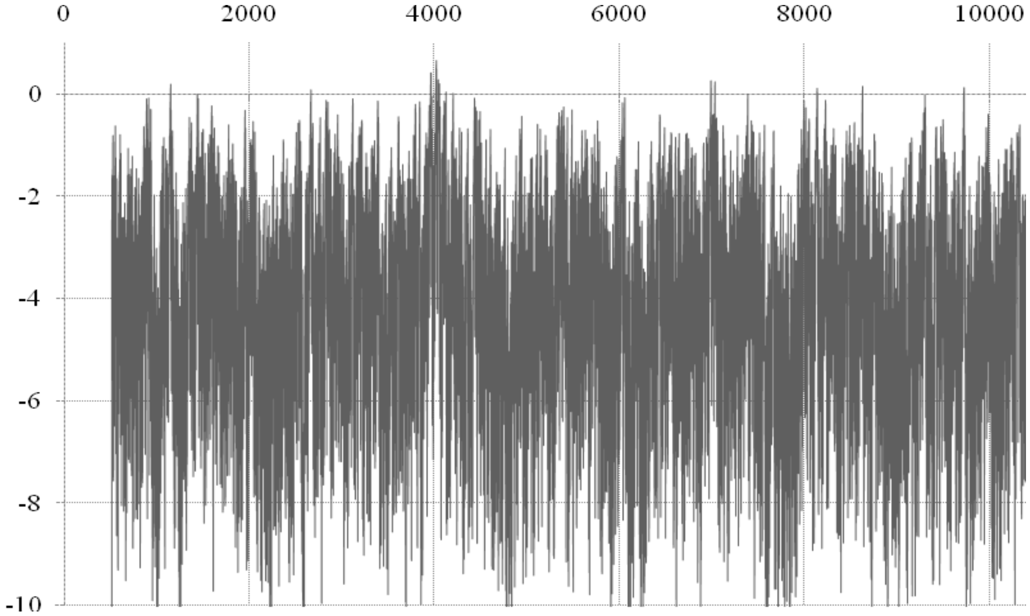
A vast amount of critical information is contained in the genomic sequences and requires multiple analysis techniques to locate and interpret the data. The standard approach is to represent genomic sequences in the form of character sequences of which each character can be one out of a finite number of entities (4 in case of DNA and RNA and 20 in case of amino acid). That is, in the case of DNA or RNA, the character string consists of the elements of the set {A, T (or U in case of RNA), G, C}; in the case of proteins, the character string consists of the 20 amino acids. Genomic information is thus inherently discrete in nature as there are finite numbers of elements in the set that comprise the genomic sequence. DNA molecules thus store the digital information that constitutes the genetic blueprint of living organisms. This fact suggests that we may interpret the DNA sequence as a discrete-time sequence that can be studied using standard techniques from the field of digital signal processing. Although DNA sequences are truly symbolic signals, numerical assignments are required to be made for analysis purposes. Such assignments can be made only after careful consideration. Various methods for numerical assignments have been briefly discussed in Section 6.3

Digital Signal Processing (DSP) comprehends the representation, transformation and manipulation of digital signals as well as the information associated to them. In this context, signals are usually physical magnitudes that vary in time or space, and digital signals are those represented as sequences of numbers, as in the case of time series. However, this symbolic approach severely limits the method for mathematically and computationally handling the data. The possibility of finding a wide application of DSP techniques to analyze the genomic sequences arises when these are converted appropriately into numerical sequences, for which several rules have been developed.

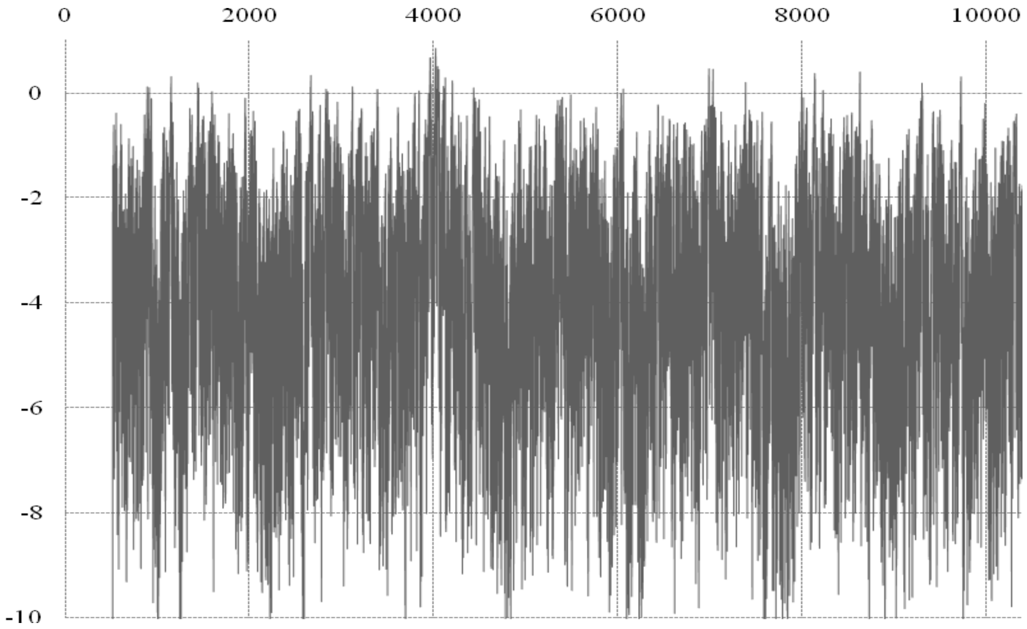
It is expressed in [6.1] that an obstacle may be faced in extracting useful information content in a genomic sequence. Considering still sparsely known dependence between nearby bases and their occurrence statistics across the genomic sequence (in Markov's sense), it is expressed in [6.1] that extraction of useful information in a genomic sequence may not be fully done with the basic statistics of the constituents. As such, it is argued that the Fourier-transform may be adopted considering the fact that, real and imaginary parts of Fourier coefficients are all independent random variables and as such, they may yield two distinct sets of fortifying details on the associated statistics with augmented information. Simulation experiments using some DNA strings extracted from the GenBank database [6.2 - 6.5] are shown to confirm the said assertion of [6.1].

The use of Fourier methods for bio-sequence analysis is also described in (6.6) and [6.7], where Fourier expansions are described as the "image" of the spatial sequence with relevant comparison efforts. Further, detection of similarities between DNA sequences is illustrated in [6.8] by enforcing Fast-Fourier transform (FFT) to ascertain the correlation between DNA sequences using complex-plane encoding. In [6.9], the Fourier transform method is adopted to distinguish coding and non-coding sub-sequences in a complete genome. A comprehensive review on genomic signal-processing method is addressed in [6.10] and [6.11] where the local texture information in genome structures is extracted by Fourier spectral mapping of test sequences. In extended contexts of using Fourier transform methods in the analysis of genomics and proteomics, digital signal-processing (DSP) techniques are also proposed in the literature [6.12] where, spectrograms are shown to be powerful tools for DNA sequence analyses providing local frequency information. Specifically, the so-called *short-time Fourier transform* (STFT) appears to provide a useful localised measure of frequency content in the spatial sequence pattern. This STFT obtained in [6.12] is consistent with traditional discrete Fourier transform (DFT). It is applied to genome sequence analysis using sliding-window technique. In [6.12], the parametric feature of the test sequence considered refers to the so-called *electron-ion interaction potential* (EIIP) values assigned to the nucleotides. They denote a numerical representation of the genome sequence, which can be subjected to a Fourier transform. That is instead of viewing the parametric values of EIIP along the sequence depicting implicitly the statistics of genetic information, one can obtain the Fourier transform of the numerical sequence and the resulting spectrum can be analyzed towards determining the associated genomic information.

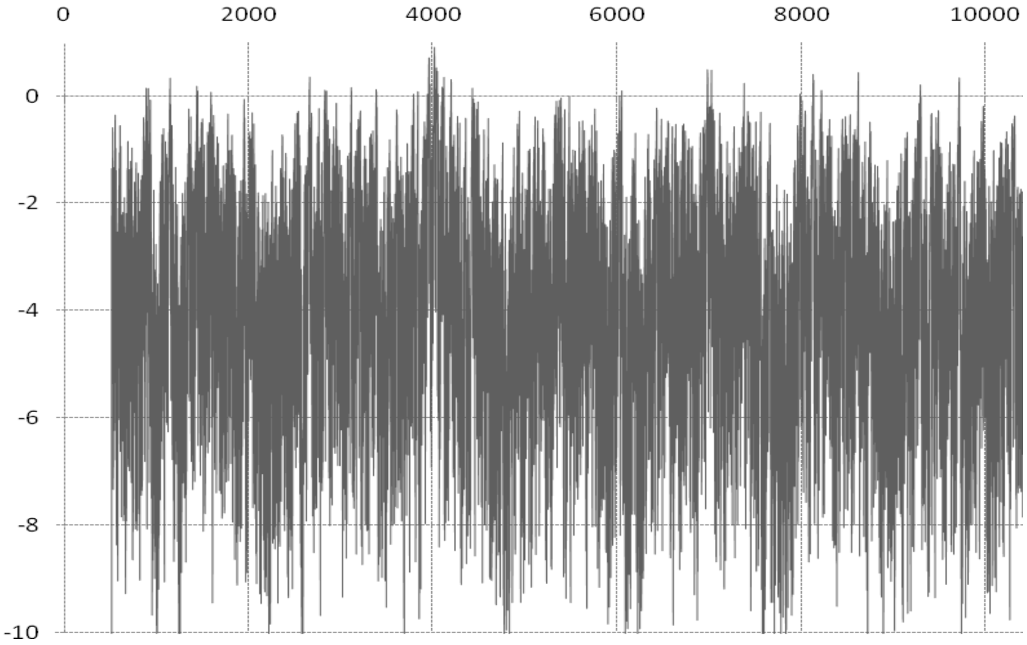
For a clear Fourier spectrum to be generated, a reasonable sized sample is required. Larger samples give relatively poor resolution, making it tedious to locate features exactly using only the Fourier transform. Smaller sample provides better resolution, but, makes it difficult to locate the local distinguishing features. Figure 6.1 (a), (b), (c) and (d) below shows the subtle variations in the complete Fourier spectrum of DEN1 serotype of dengue virus using various window sizes obtained with EIIP data. (More details on EIIP are furnished in Chapter VII). The x axis denotes the base locations. To exactly view the difference in resolution due to the differing window size, consider a threshold to be zero. Figure 6.2 (a) – (d) shows the difference in resolution of DEN1 due to varying window sizes.



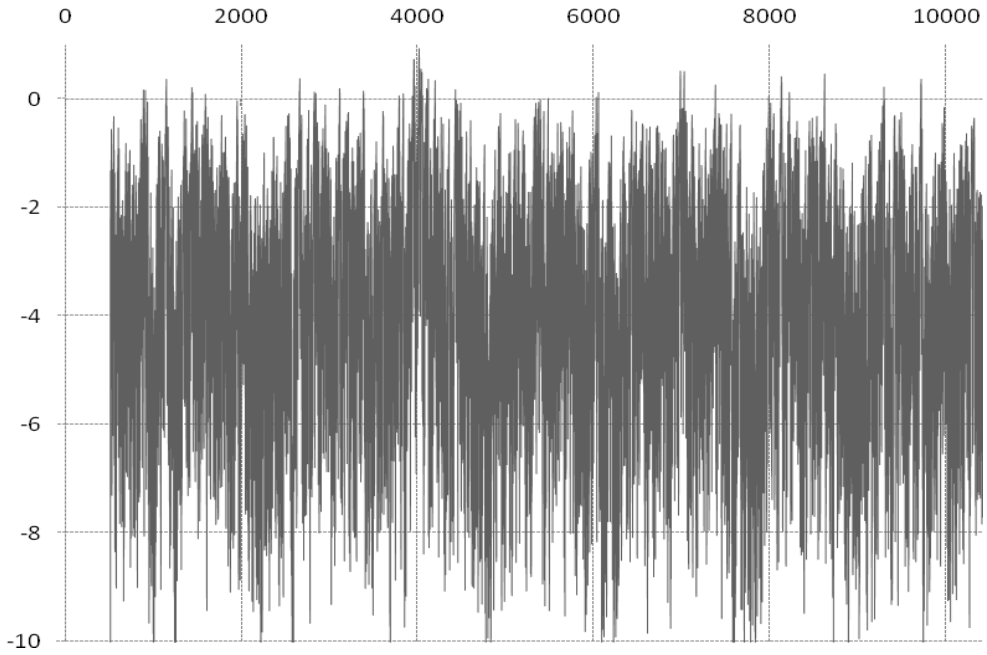
(a)


0
2000
4000
6000
8000
10000
0
-2
-4
-6
-8
-10

(b)


0
2000
4000
6000
8000
10000
0
-2
-4
-6
-8
-10

(c)



(d)

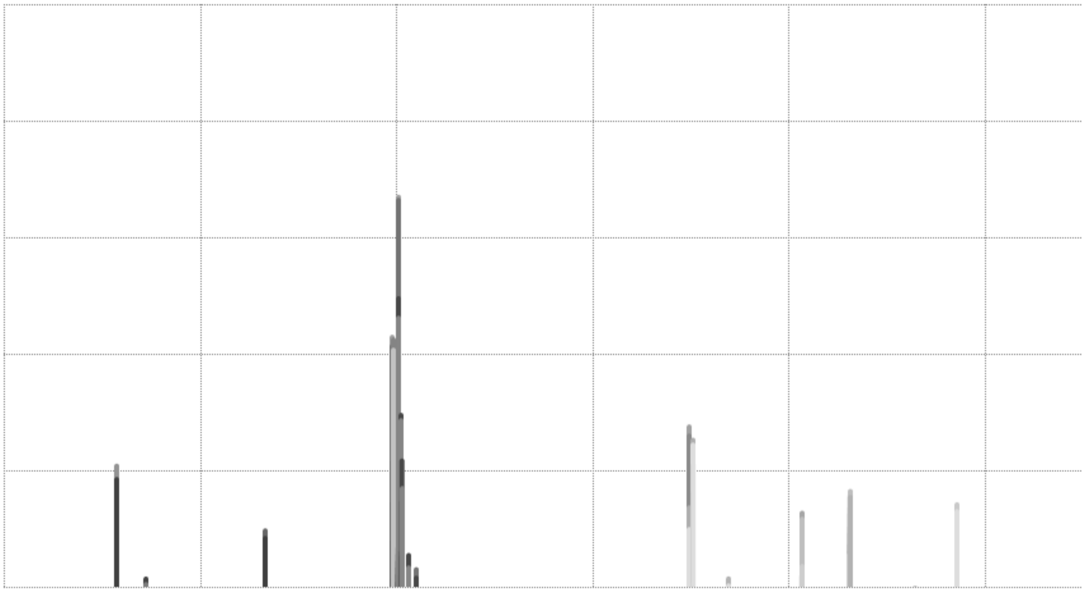
Figure 6.1: Spectrum of DEN1 serotype of dengue virus sequence details pertinent to EIIP values computed using various window sizes

(a) Spectrum of complete DEN1 sequence details pertinent to EIIP values computed with window size 120

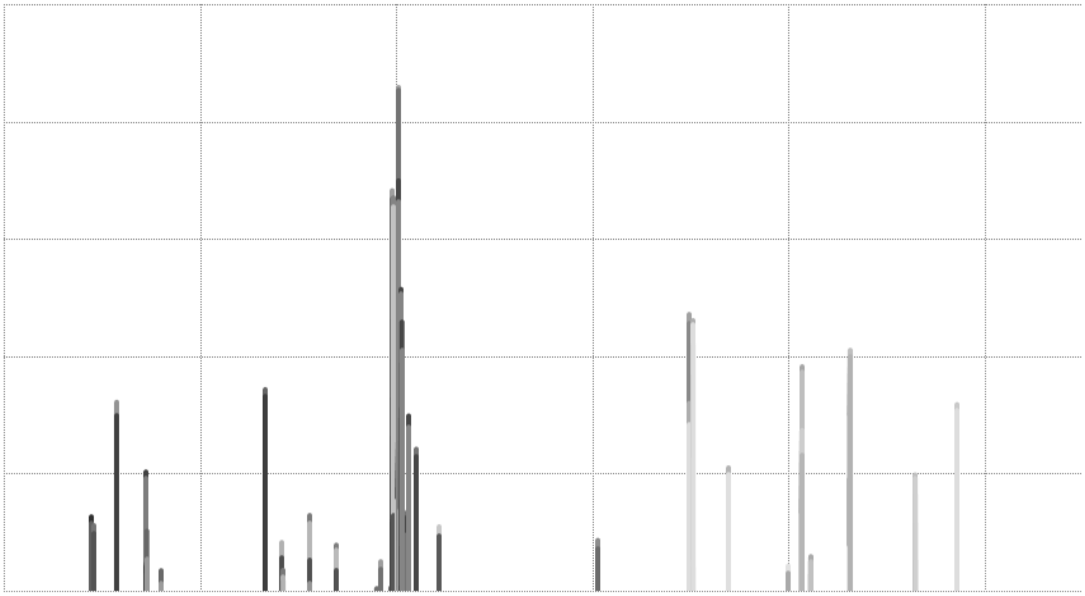
(b) Spectrum of complete DEN1 sequence details pertinent to EIIP values computed with window size 180

(c) Spectrum of complete DEN1 sequence details pertinent to EIIP values computed with window size 255

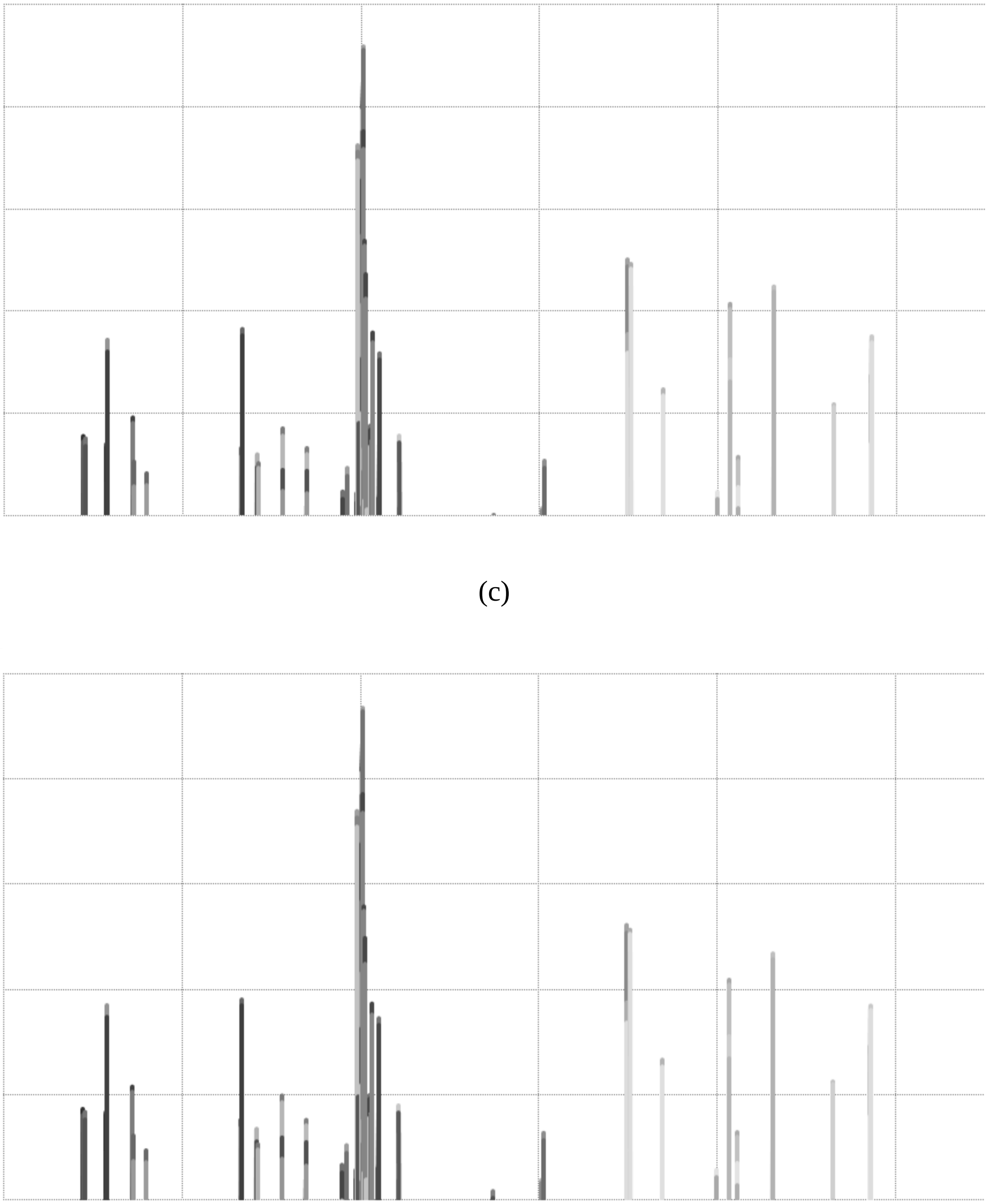
(d) Spectrum of complete DEN1 sequence details pertinent to EIIP values computed with window size 510



(a)



(b)



(c)

(d)

Figure 6.2: Difference in resolution of DEN1 details pertinent to EIIP values computed due to varying window sizes

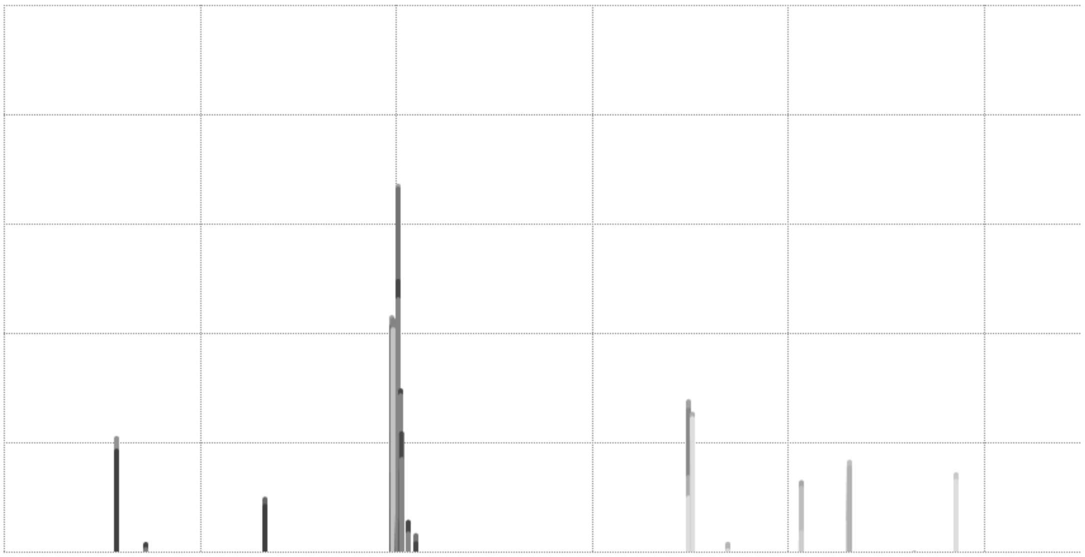
(a) Spectrum of DEN1 sequence details pertinent to EIIP values computed with window size 120

(b) Spectrum of DEN1 sequence details pertinent to EIIP values computed with window size 180

(c) Spectrum of DEN1 sequence details pertinent to EIIP values computed with window size 255

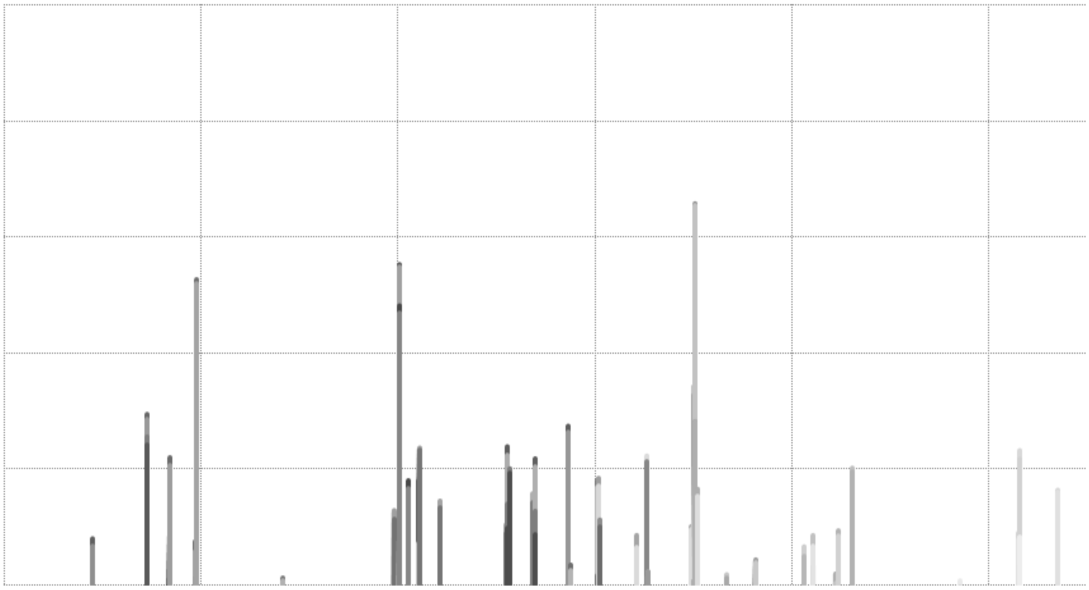
(d) Spectrum of DEN1 sequence details pertinent to EIIP values computed with window size 510

The genomic sequence of the different serovar of dengue virus has a homology of 60-80% amongst themselves. By spatial analysis of the spectrums of all the four serovar, the similarity/dissimilarity between these serovar can be identified. Figure 6.3 below shows the distinguishable features of all the four serotypes for the range 0 to 1. In the present study, a window size of 120 has been used as it provides reasonable resolution as well as indicates the local features and is consistent with the window size used in the other analytical methods.

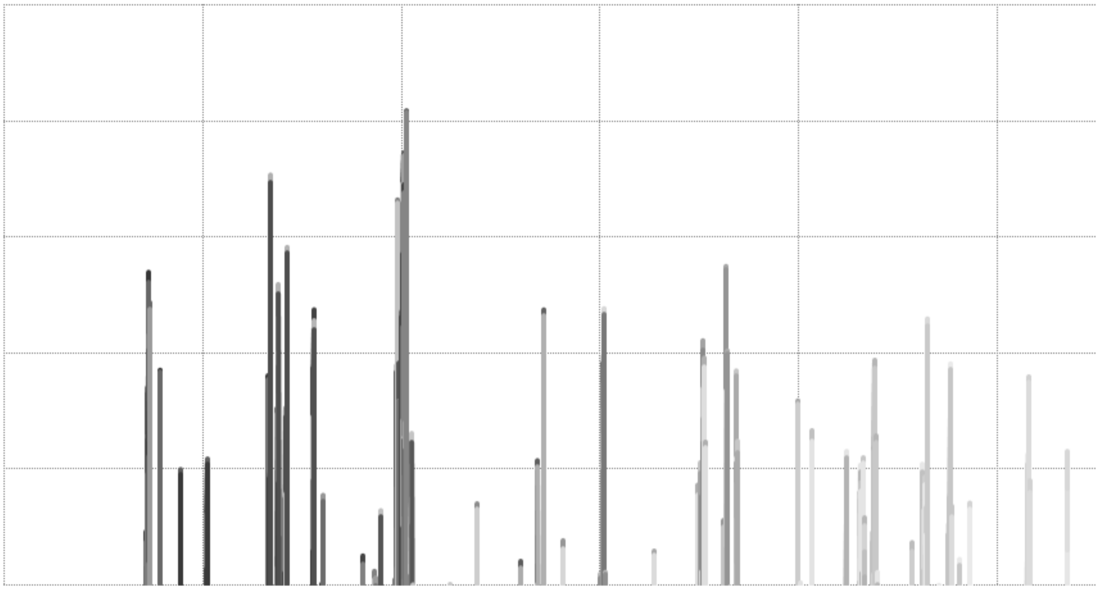


(

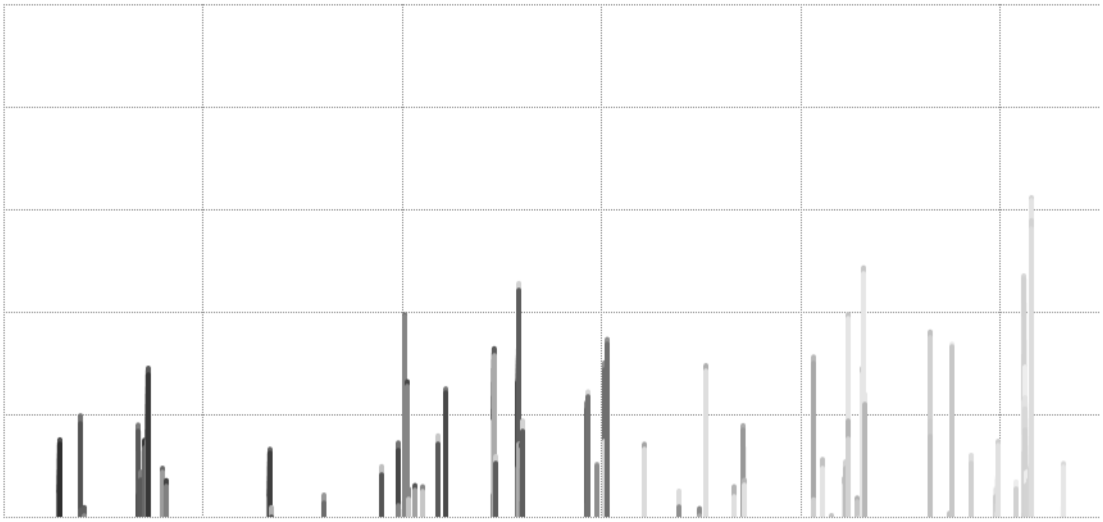
(a)



(b)



(c)



(d)

Figure 6.3: Distinguishable features of all the four serotypes of dengue virus. Details pertinent to EIIP values computed

(a) Spectrum of DEN1 sequence details pertinent to EIIP values computed with window size 120

(b) Spectrum of DEN2 sequence details pertinent to EIIP values computed with window size 120

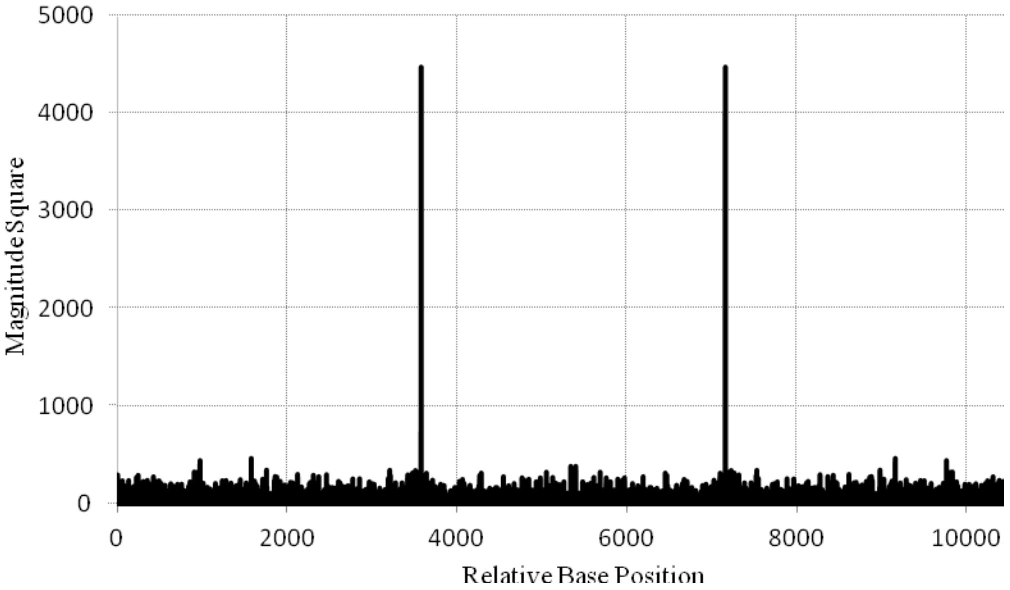
(c) Spectrum of DEN3 sequence details pertinent to EIIP values computed with window size 120

(d) Spectrum of DEN4 sequence details pertinent to EIIP values computed with window size 120

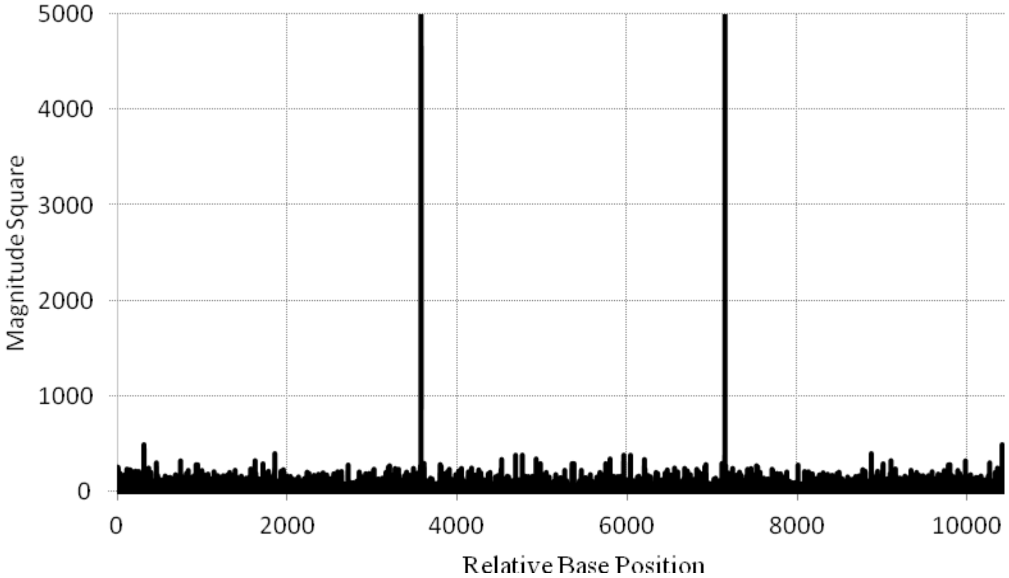
6.2 Periodic Property of Genomic Sequence

As seen previously in Section 6.1, if the characters in the genomic sequence are represented by numerical values, then signal processing algorithms can be easily applied to extract useful information. However, it is important that systematic and proper values are assigned to each of the characters in the genomic sequence and not just some random value.

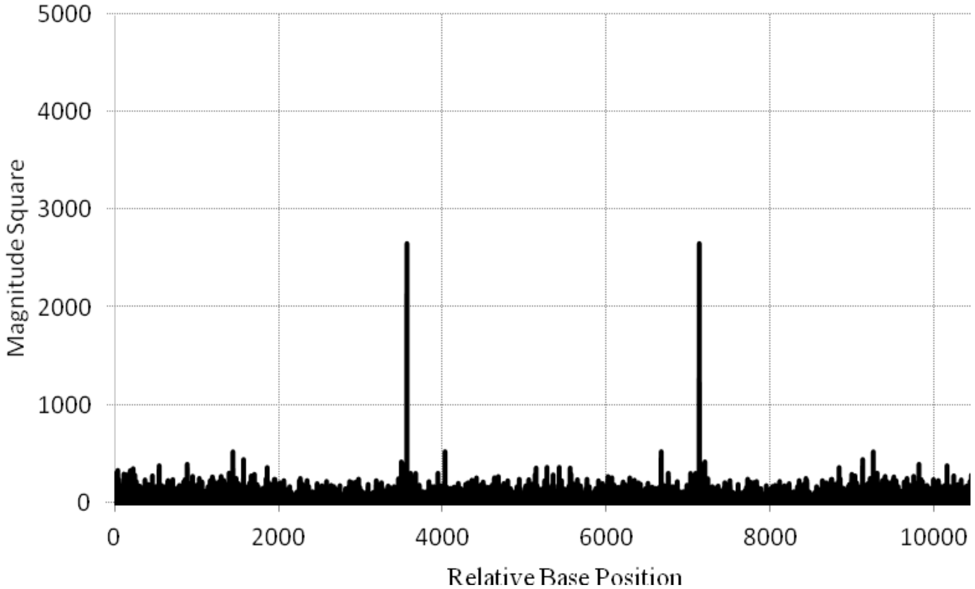
Periodicity of DNA sequences has been examined by various methods including autocorrelation function analysis, Fourier spectrum analysis, DNA walking, entropy, Hurst index estimation, de-trended fluctuation analysis, wavelet translation and mutual information function. One of the important properties of any genomic sequence is the 'period-3 property' mentioned by Annastasiou in (6.10). By virtue of the characteristic information in the gene sequence bearing the triplets (= 4^3 = 64) constituted by the nucleotide set {A, T, G, C} across exon regions, it is shown in (6.10) that the corresponding Fourier domain power-spectrum of a prokaryotic DNA has a strong peak at a frequency k = N/3. It corresponds to a spectral component with a period 3 and N represents the discrete integer of the set denoting the sequence (in the spatial domain). Figure 6.4 shows the period 3 property of all the four strains of dengue virus using the method described by Annastassiou [6.10].



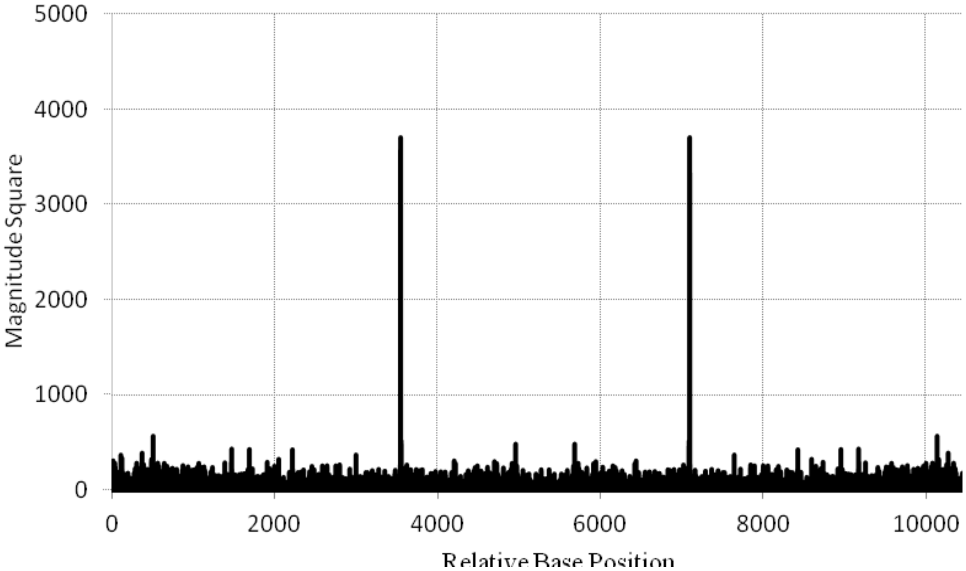
(a)


5000
4000
3000
2000
1000
0
Magnitude Square
0
2000
4000
6000
8000
10000
Relative Base Position

(b)


5000
4000
3000
2000
1000
0
Magnitude Square
0
2000
4000
6000
8000
10000
Relative Base Position

(c)



(d)

Figure 6.4: Period 3 property of all the four serotypes of dengue virus

(a) Period 3 property of DEN1 serotype
(b) Period 3 property of DEN2 serotype
(c) Period 3 property of DEN3 serotype
(d) Period 3 property of DEN4 serotype

Should non-informative introns excessively interrupt the sequence (as in the case of eukaryotic DNA), the power-spectrum would then tend to become flat. In other words, the power spectrum of the DNA is an implicit indicator of the mutual information content (or, the conditional entropy of, say a segment of a DNA (exon) being informative subject to the presence of non-informative segments (introns) present in the domain of interest). Thus, homologous DNA sequences are implicitly specified by their power spectrum attributes as discussed in [6.1] and [6.10].

Furthermore, to reflect the differences in coding structure of nucleotides, use of power spectral analysis of DNA sequences has been elaborated in [6.13], with reference to the classification of bacteria; and, the power spectra of the underlying DNA sequences are presented as self-organizing maps. Such power-spectrum approach is regarded as intuitive as well as effective in reducing the dimension of the complete DNA sequences of a clustered set of species. Relevant details are again fortified in a number of studies [6.14 - 6.21]. Apart from such archival set identified above, there also prevails an exclusive thesis on Fast-Fourier transform (FFT) analysis applied to DNA sequences due to Hanson [6.22].

6.3 Numerical Representation of Genomic Sequences

The numerical representation determines which features of the genome are highlighted by the analysis. For example, Figures 6.1 and 6.3 (a) represent the spectrum of DEN1 serotype of dengue virus. However, they use different numerical representation and highlight different features of the same viral serotype. Thus, the translation can be performed in a number of ways, a few of which has been discussed below.

One of the earliest methods was proposed by Silverman and Linsker [6.23]. They represent each base as the vertex of a tetrahedron in three dimensional spaces and the genome sequence is transformed into an array composed of three dimensional vectors. A Fourier transform is then performed on each of the three sequences made up of a directional component from the sequence vectors. The resulting spectrum is the sum of the three Fourier transforms. Tiwari et al. [6.24] uses four binary strings to represent the occurrence of each base in the nucleotide sequence, summing the individual spectra to give an overall sample spectrum for the genome. Using this method, a repeating period of 3 has been found in coding regions [6.24-6.26]. Fourier analysis has also been employed as one of a number of weighting factors in the determination of introns splice sites.

In formulating the spectral domain comparison of genomic sequences of test viral strains, the FT-based algorithmic and computational efforts pursued in the present study essentially follow the procedure due to [6.12], wherein the genomic sequences are represented by numerical sequences and the FT of the spatial-sequence is determined. The numerical values used thereof conform to the so-called *electron-ion interaction-potential* (EIIP) values assigned to the nucleotides. Implicitly, such an EIIP-based numerical sequence leads to an information spectrum method (ISM) of transforming a genomic sequence for decomposition *via* Fourier series. This Fourier-transformed sequence inherently contains physico-chemical details attached to the biological functions of the genomic structure. As such, elucidating the Fourier series of numerically-formatted genome leads to detecting code/frequency pairs that are specific to the genomic sequence *vis-à-vis* its biological profiles. This method is insensitive to the location of the motifs and therefore warrants no prior alignment of the sequence with its counterparts. The EIIP values assigned for nucleotide bases are listed in Table 6.1.

Table 6.1: EIIP values of nucleotide bases [6.12]

Base	**EIIP value**
A	0.1260
T	0.1335
G	0.0806
C	0.1340

For example consider the following character sequence: {A, A, A, G, T, A, G, C}. The corresponding EIIP values for the above sequence is {0.1260, 0.1260, 0.1260, 0.0806, 0.1335, 0.1260, 0.0806, 0.1340}.

6.4 Spatial Domain Analysis of Genomic Sequence

As mentioned earlier, in, the Fourier transformation in spatial domain can be applied the context of biological sequence analysis to identify protein coding genes in a linear sequence. In essence, given a function of a spatial variable, its Fourier transform identifies different frequency sinusoids and their amplitudes contained in that function.

Suppose an analog numerical signal f(x) with x of varying amplitude (specified for example, a sequence of EIIP values assigned to each base of the genomic sequence) is considered. Its STFT can be defined as follows:

$$f(x) \leftrightarrow F(f) = \sum_{m=1}^{R=120} f(x-m)\ w(m)\ \exp^{-m\omega} \qquad (6.1)$$

In evaluating equation (1), the whole test sequence is divided into N equally-spaced intervals or frequency (presently, N has been arbitrarily taken as 120 to match the window size improvised in other methods mentioned earlier). As already explained in 2.3, the frequency k in equation (1) is equal to N/3. Since, $\omega = 2\pi k/N$, with the substitution of value of k, $\omega = 2\pi/3$. The subsequence spanning the window is denoted by w[m] with the window length (size) being R and $R \leq N$. (taken as 120). The computational procedure of elucidating Fourier spectral details of a test sequence is presented in Chapter VII.

6.5 Conclusion

In this chapter, the use of spatial spectral domain analysis as a tool for finding the underlying features of a genomic sequence has been discussed. Various methods for assigning numerical values to the genomic sequences have been discussed. It has been observed that depending on the numeric value assigned as well as the selected window size, different properties of the genomic sequence can be invoked and explored by applying Fourier transform to the numeric sequence.

CHAPTER VII

A METALEARNING APPROACH TO EXPLORE VIRAL GENOMICS: A MODULAR FRAMEWORK OF DATA-MINING

7.1 Introduction

A workflow structure of computation, wherein the user facilitates necessary feedback toward final classification and decision conforms to a metalearning approach. Developed in this chapter is a bioinformatic inference methodology that allows a metalearning framework to systematically classify and elucidate the "best candidates" which prevail commonly among a set of viral serotypes. Relevantly, the present chapter refers to a bioinformatic method of analyzing a set of genomic sequences in order to cohesively elucidate their common and differential features in terms of the associated entropy, energetics and spectral characteristics. Such a diverse and distinct set of analyses would robustly enable distinguishing and comparing test sequences in terms of their global as well as local genomic details. For example, considering the diversity that manifests as distinct variations within the subspecies of a virus (or bacteria) [7.1] designated as *serovar* or *serotypes*, comparing and contrasting them as regard to their genomic features can help in the efforts of seeking a single and/or distinct rational vaccines for the serogroup in question. In acquiring such genomic feature details of the viral strains using multiple procedures with information collected independently from different pursuits, will offer clarity on the distinctions to be made and similarities to be identified for possible use in vaccine design applications.

Hence, proposed here is a strategy to view a given set of genomic sequences in the framework of three perspectives, namely their associated entropy, energetics and spectral characteristics. These three independent methods would complement each other in robustly identifying the distinctions as well as the similarity features among the test sequences in question; and, implementing such three-prong analyses essentially forms the scope of the present study.

7.2 Scope of the Present Study

With the availability of entropy, energetic and spectral-domain methods narrated in Chapters IV, V and VI, it is attempted in this present chapter to invoke and apply all these three techniques cohesively in analyzing a set of genomic sequences. The reason for this cohesive and collective approach is to gather comprehensively as much of feature details as possible that would distinguish and differentiate subtly the test genomic sequences. In contrast, suppose a single technique is alone deployed. Then, the associated analysis may not show all the distinguishing feature details as needed.

As indicated earlier, the test genome being considered here for the study refers to that of a virus, and the set of sequences analyzed correspond to the serotypes of this test virus. Strains of a single virus, though are known to be distinct from each other, may possess certain common subtle features [7.2], which remain dormant, mostly unseen and may not be indicated explicitly, if a single type of analysis is restrictively envisaged.

Hence, formulated as a scope of this study is a three-prong methodology that robustly evaluates the subtle similarity/dissimilarity features between the genomes of a given set of viral strains; and, it is surmised that such cohesive feature details collected across the test serovar may possibly show directions toward designing common and/or distinct vaccines for the diseases caused by the test virus.

In the present study, the test dengue virus, takes four forms of serovar, namely DEN1, DEN2, DEN3 and DEN4. Relevant GenBank data [7.3 - 7.6] provides complete genomic details of these strains and, analyses are pursued in conformance with the proposed cohesive efforts of using entropy, energetic and Fourier transform techniques.

7.3 Methodology/Application of Analytical Frameworks

In seeking genomic details that indicate common and differential features of the serotypes of a virus, it is attempted in this study *via* three distinct methods mentioned earlier, namely entropy, energetics and Fourier transform techniques. Outlines on the implementation of these approaches are presented below:

7.3.1 Entropy-based analysis

The statistical aspects of nucleotide sequence entropy discussed earlier implies probabilistic occurrence of nucleotides positioned spatially along the sequence length. Inasmuch as the DNA sequence is a mix of codon (exons) and non-codon (introns) parts, the appearance of nucleotide the elements in the set {A, T, G, C} along the sequence-space will be inherently random. However, considering the exon regions exclusively, the associated randomness implies negative entropy yielding Shannon information concerning the underlying genetic code toward protein making and it remains fostered in the associated coding sequence segments (CDS) of the genomic structure.

In the context of non-viral DNA sequences, the coding and non-coding regions have been comprehensively studied as regard to their entropy details in Chapter IV. It is attempted here to apply similar analytical considerations to the four strains of dengue virus under discussion. For this purpose, specifically the so-called entropy segmentation method and its variations in statistical divergence sense are used. The procedure is described below.

Distinguishable characteristics of non-informative introns and informative exons enable delineating the associated splice-junctions using various statistical divergence methods (7.7) of entropy segmentation. Hence, considered here, is the well-known Kullback-Leibler (KL) divergence measure, which delineates the exon/intron boundaries and provides an information profile of the test sequence in question.

It is applied to the genomic sequence of all the four strains of dengue virus so as to distinguish them in the entropy-plane. Relevant algorithmic steps are outlined in Pseudocode A.

Pseudocode A

```
%    Pseudocode on computing the relative entropy features of
     the RNA sequences of test viral serotypes

Initialize

//Computation refers to elucidating the relative entropy features
of a given set of RNA sequences of, for example, dengue virus
strains
```

Input

→ Test nucleotide sequence from 5′-end to 3′-end (of each test strain of the dengue virus), is posted as a string (for example, numbered from i= 1 to 10735 for DEN1)

→ Each test sequence is converted into **(1 × n)** matrix with each letter of {A, T, G, C} representing the matrix element

Construct: A hypothetical random sequence of {A, T, C, G} with uniform probability distribution of occurrence of each element

→ A hypothetical nucleotide sequence is generated with the statistics of A, T, G, C as follows: $p_A = p_T = p_C = p_G = 0.25$

→ This hypothetical nucleotide sequence denotes a non-informative 'junk' chain of {A, T, G, C} and used as a common reference sequence

→ The hypothetical sequence matrix constructed is again of size (1×10735) corresponding to the same length of DEN1 characters

Next

// Step I

→ Specify a sliding-window accommodating 120 nucleotides

→ Calculate the occurrence probabilities of (A, T, G and C) per window-segment (with 120 nucleotides) in the hypothetical sequence

→ Count the number of A's in the window = M_A

← Then, the occurrence probability of A: $q_A = M_A/120$
Similarly, evaluate q_T, q_G and q_C per window

// Step II

Calculate the occurrence probabilities of (A, T, G and C) per window-segment (with 120 nucleotides) in the actual test sequence (say DEN1)

→ Repeat procedures of Step I to obtain $\{p_{A1}, p_{T1}, p_{G1}, p_{C1}\}$ per window for DEN1

```
      →  Similarly, evaluate the sets {$p_{A2}$, $p_{T2}$, $p_{G2}$, $p_{C2}$}, {$p_{A3}$,
         $p_{T3}$, $p_{G3}$, $p_{C3}$} and {$p_{A4}$, $p_{T4}$, $p_{G4}$, $p_{C4}$} for DEN2, DEN3 and
         DEN4 respectively

// Step III
Compute the statistical divergence: Invoke Kullback-Leibler
entropy – Information-theoretic (IT) measure
    → The statistical divergence (distance) between the test and
      hypothetical sequences is determined via Kullback-Leibler
      (KL) measure
    → Define a window size of 120 bases
        ← KL divergence estimation of each window segment of 120
          bases is done as follows: (The window is identified by
          an index k = 1, 2,…, 90)
        ← Suppose, KL1 refers to the estimation pertinent to DEN1
          $(KL1)_k$ =  $\{[(p_{a1} \times \log(p_{a1}/q_a) + (q_a \times \log(q_a/p_{a1})] +$
                  $[(p_{t1} \times \log(p_{t1}/q_a) + (q_t \times \log(q_a/p_{t1})] +$
                  $[(p_{g1} \times \log(p_{g1}/q_g) + (q_g \times \log(q_g/p_{g1})] +$
                  $[(p_{c1} \times \log(p_{c1}/q_c) + (q_c \times \log(q_c/p_{c1})]\}_k$
Next
    → Compute KL2, KL3 and KL4 likewise using Step III for each
      window (k = 1, 2,…, 90) DEN2, DEN3 and DEN4 respectively
End
```

7.3.2 Energetics-based analysis

The energetics profile of DNA/RNA structures, as discussed earlier in Chapter V, refers to the framework of structural stability of a sequence, that conforms to the rules stipulated by the chemistry of bonding concerning each base-pair as well as it depends on the disposition of the most adjacent pairs. Relevantly, it relies on the associated total free-energy (in the thermodynamic sense) depicting the sum of the contributions from the neighbors (7.8). That is, the INN-model takes into consideration two neighboring bases (about a pair of centre elements); and, an energetics value (EV) for each pair is assigned depending on the neighbors on the immediate right- and left-side of the centre pair. Compiled data on the energetic value for each pair of centre element with reference to their adjacent neighbors is presented in Table 5.1 in Chapter V. Low value of EV implies that the chemistry of the elements is self-selected in the sequence so as to assume a minimum potential energy profile toward thermodynamic stability.

The analytical basis of prescribing an EV-profile along the test sequence is summarised in the Pseudocode B.

Pseudocode B

```
%     Pseudocode to compute the energetics feature associated
      with nucleotides dispositions as individual nearest-
      neighbors (INNs) in a test RNA sequence

Initialize

   → Test nucleotide sequence (say DEN1) from 5′-end to 3′-end is
     posted as a string (numbered from i= 1 to 10735 for DEN1)


// Step I
Construct the INN-concept based energetic profile for the test
sequence

   ← At each nucleotide position, energetic values are assigned
     by considering a pair of bases, designated as centre
     elements (CE) and their neighbors (NN-L and NN-R) on each
     side of this central pair, using the EV allocations as in
     Table 5.1
           →    A moving-window is set to slide past each
                nucleotide; and, at each window position, the
                minimum energy of the CE with respect to their
                neighbors is noted down using Table 5.1.
            →   The energy values observed are stored in a matrix
                EV1 (of size 1×10732) for DEN1
Next

    → EV2, EV3 and EV4 are corresponding matrices obtained
      likewise for DEN2, DEN3 and DEN4 respectively using Step I

End
```

7.3.3 Fourier-transform FT based analysis

As mentioned earlier in Chapter VI, given a function of a spatial variable, its Fourier transform (FT) identifies different frequency sinusoids and their amplitudes contained in that function. Relevant FT-based algorithmic and computational efforts pursued in formulating the spectral domain comparison of genomic sequences of test viral strains, essentially follow the procedure due to (7.9), wherein the genomic sequences are represented by numerical sequences and hence, the FT of this spatial-sequence of numerals is determined. The numerical values used thereof as mentioned in Chapter VI conform to the so-called *electron-ion interaction-potential* (EIIP) values assigned to the nucleotides. Implicitly, such an EIIP-based numerical sequence leads to an *information spectrum method* (ISM) of transforming a genomic sequence for decomposition *via* Fourier series. This Fourier-transformed sequence determined from EIIP values inherently contains physico-chemical details attached to the biological functions of the genomic structure. As such, elucidating the Fourier series of this numerically-formatted genome (in terms of EIIP values) leads to detecting the code/frequency pairs that are specific to the genomic sequence *vis-à-vis* its biological profiles. This method is insensitive to the location of the motifs and therefore warrants no prior alignment of the sequence with its counterparts. The EIIP values assigned for nucleotide bases are listed in Table 6.1 in Chapter VI. The computational procedure of elucidating Fourier spectral details of a test sequence is presented in Pseudocode C.

```
                              Pseudocode C
%    Pseudocode to compute spatial frequency characteristics of
     a test RNA sequence using STFT method

Initialize

%    Identify EIIP values, which refer to the chemistry-
     specified electro-ion interaction potential for each base
     as listed in Table 6.1

Input

  →  Test nucleotide sequences from 5′-end to 3′- end (say, for
     example, DEN1) is posted as a string (numbered from i= 1 to
     10735)

// Step I
Construct the spatial frequency spectrum for the test sequence of
DEN1 based on EIIP values using STFT

  ←  Construct a string of EIIP values replacing {A, T, G, C} on
     the sequence

  ←  CALL EIIP values from Table 6.1 and assign them
     appropriately to each base encountered along the entire
     sequence of DEN1 and store it in a matrix

  ←  The short-time Fourier transform (STFT) of the test
     sequence of EIIP-string is obtained by applying Equation
     (1) with a frequency $\omega = 2\Pi/3$ and suitable window size
     (say, $R = 120$)

  ←  Apply Step I to determine similarly, the spectra of DEN2,
     DEN3 and DEN4

End
```

The flowchart illustrating the algorithmic steps of Pseudocodes A, B and C are given below in Figure 7.1:

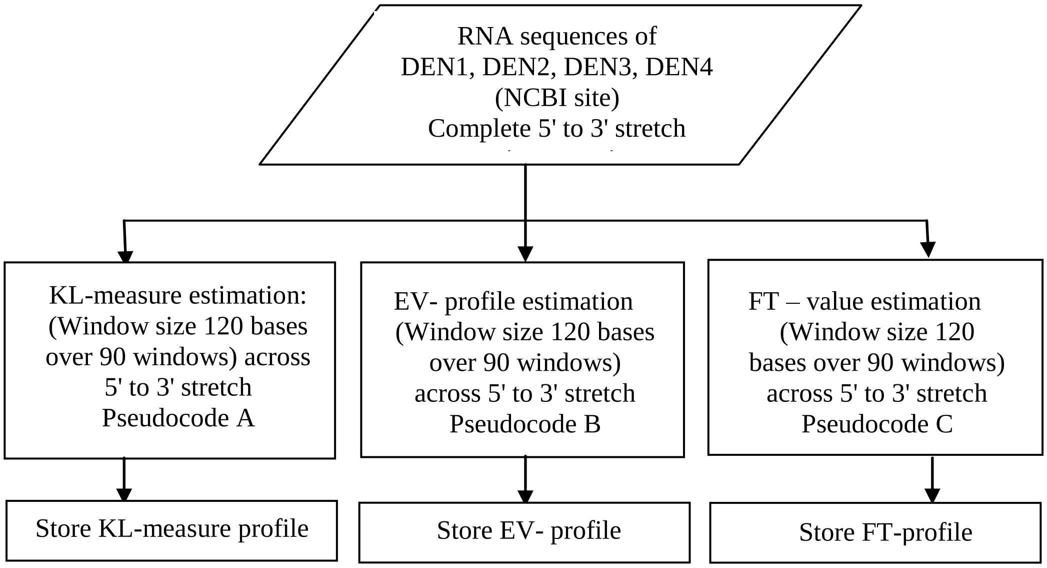

RNA sequences of
DEN1, DEN2, DEN3, DEN4
(NCBI site)
Complete 5' to 3' stretch
KL-measure estimation: (Window size 120 bases over 90 windows) across 5' to 3' stretch Pseudocode A
EV- profile estimation (Window size 120 bases over 90 windows) across 5' to 3' stretch Pseudocode B
FT – value estimation (Window size 120 bases over 90 windows) across 5' to 3' stretch Pseudocode C
Store KL-measure profile
Store EV- profile
Store FT-profile

Figure 7.1: Flowchart illustrating the algorithmic steps of Pseudocodes A, B and C

7.4 Extracting common and differentiating features between a set of sequences

Having formulated three distinct methods based on entropy, energetics and Fourier spectrum characteristics of a nucleotide sequence as discussed above, the results are compiled with reference to the four sequences pertinent to RNA sequences of DEN1, DEN2, DEN3 and DEN4. Using these data, the motif segments that indicate common features seen along 5′-end to 3’-end stretches of all these multiple (four) sequences in question are ascertained. Relevant procedures deployed are outlined in Pseudocodes D1 and D2.

Pseudocode D1

```
%   Pseudocode on the procedure to determine the set of motifs
    forming the finger-print of four RNA sequences pertinent to
    DEN1, DEN2, DEN3 and DEN4

Initialize

→   Identify and designate the test viral strains
      →   Dengue viral strains: Dengue 1 virus (DEN1), Dengue 2
          virus (DEN2), Dengue 3 virus (DEN3) and Dengue 4 virus
          DEN4

→ Perform Collection of RNA sequence details from NCBI site at:
          : http://www.ncbi.nlm.nih.gov/genome/10308

        →DEN1: 5′ agttg…tct 3′  (10,740 bases)
        →DEN2: 5′ agttg…tct 3′  (10,720 bases)
        →DEN3: 5′ agttg…tct 3′  (10,710 bases)
        →DEN4: 5′ agttg…tct 3′  (10,650 bases)

→   Perform: Analyses concerning: Extraction of entropy,
    energetics and Fourier spectral details

      →   Computations are done as per the following algorithms:

//Step I

Call Sub pseudocode A, B and C

  → Recall the test RNA sequences
          ←   For each RNA sequence stretch from 5′ end to 3′
              end:
          →   Apply subroutine outlined as in Pseudocode A for
              the calculation of KL-measure
          →   Apply subroutine outlined as in Pseudocode B to
              determine the EV profile
          →   Apply subroutine outlined as in Pseudocode C to
              get the EIIP-based spectral profile
          ←   Perform the calculations as above considering a
              window size of 120 bases and across all windows
          →   Find the mean of KL_measures obtained across all
              windows from 5′-end to 3′-end:
              →KLM1
        →    Find the mean of min_Energy values (EV profile)
             across all windows from 5′-end to 3′-end:
             →MEM1
        →    Find the mean of FT_amp values (EIIP-based
             spectral profile) across all windows from the 5′-
             end to 3′-end:
             → FTM1

←    Assuming that the net characteristics of the sequence is
     intrinsically linear and equally-weighted by KLM, MEM and
     FTM, combine the three measures, namely, (KLM, MEM and FTM)
     to obtain a summed average as follows:
        → DEN1-AV = (KLM1 + MEM1 + FTM1)/3

←    Obtain likewise DEN2-AV, DEN3-AV and DEN4-AV respectively
     for DEN2, DEN3 and DEN4

→    For each window indexed as n (n = 1, 2, 3, … , 90) across
     5′-end to 3′-end, determine normalized (linearly)combined
     measures for DEN1 as follows:
             D1 = [Average of $(KLM1 + MEM1 + FTM1)_n$]/ DEN1-AV

→      Similarly, for DEN2, DEN3 and DEN4 respectively obtain,
             D2 = [Average of $(KLM2 + MEM2 + FTM2)_n$]/ DEN2-AV
               D3 = [Average of $(KLM3 + MEM3 + FTM3)_n$]/ DEN3-
             AV
               D4 = [Average of $(KLM4 + MEM4 + FTM4)_n$]/ DEN4-
             AV

End

Next
//Step II

→    Find the overall mean of D-values concerning all the four
     strains of the virus: $D_M$ = (D1 + D2 + D3 + D4)/4

     ←    Compute: Demeaned values for each window by
          subtracting the value of $D_M$ from each element of the
          matrices D1, D2, D3 and D4

             → For each window (n = 1, 2, … , 90)
                  DiffD1 = (D1 - DM)
                  DiffD2 = (D2 - DM)
                  DiffD3 = (D3 - DM)
```

```
                    DiffD4 = (D4 - DM)
Next
//Step III

→     To find logistic regression on DiffD1, DiffD2, DiffD3 and
      DiffD4

%     Note: Logistic regression [7.10] and [7.11] (also known as
      logistic model or logit model) is a method to predict the
      probability of occurrence of p(z) of an entity (z) by
      fitting data to a logistic (logit) curve. This logistic
      function is a nonlinear sigmoid (S-shaped) with z being the
      independent explanatory variable and p(z) is the dependent
      variable. With large values of z ($z \rightarrow \infty$), p would tend to 1
      as an asymptote. Otherwise, $p(z) = 1/[1+\exp(-z)]$; and, the
      variable z is set by a linear regression of the associated
      independent variables.
      In the present case, z denotes DiffD1, DiffD2, DiffD3 or
      DiffD4 all obtained by a linear combination of KLM, MEM and
      FTM measures evaluated on each sequence.

Do

→     Logistic regression $(LRk)_{k=1,2,3 \text{ and} 4}$ for each window (n = 1, 2,
      …, 120) using the logit function:

      GOTO DEN1 sequence
            →Find: $LR1_n = 1/ (1 + \exp[DiffD1])_n$

      GOTO DEN2 sequence
            →Find: $LR2_n = 1/ (1 + \exp[DiffD2])_n$

      GOTO DEN3 sequence
            →Find: $LR3_n = 1/ (1 + \exp[DiffD3])_n$

      GOTO DEN4 sequence
            →Find: $LR4_n = 1/ (1 + \exp[DiffD4])_n$

End
Next

→     Find the overall mean: LRM = (LR1 + LR2 + LR3 + LR4)/4

      ←     Demean LR1, LR2, LR3 and LR4 for each window with
            respect to LRM to obtain:

             LRD1 = |LR1 - LRM|
                 LRD2 = |LR2 - LRM|
                 LRD3 = |LR3 - LRM|
                 LRD4 = |LR4 - LRM|
```

← Find logarithm of each element of **LRD1, LRD2, LRD3** and **LRD4** to obtain respectively: **LLRD1, LLRD2, LLRD3 and LLRD4 {**For example, $\mathbf{LLRD1_n} = \log_e(\mathbf{LRD1})_n$}

% *Note:* By taking logarithm of the data as above, it will filter out most of the scattered outlier values

Plot:

→ $\mathbf{LLRD1}_{n\ =\ 1,2,3..}$ values (across 120 bases in each window) *versus* window number indexed as (120 × n) with n = 1, 2, …, 90

→ **Plot LLRD2, LLRD3** and **LLRD4** values likewise (across 120 bases in each window) *versus* window number indexed as (120 × n) with n = 1, 2,…, 90.

Next

//Construction of motifs

% *Note:* **This refers to deciding motif sections between the four multiple RNA sequences of DEN1, DEN2, DEN3, DEN4 using the computed data on** $\mathbf{LLRD1}_{n\ =\ 1,2,3..}$**,** $\mathbf{LLRD2}_{n\ =\ 1,2,3..}$ **,** $\mathbf{LLRD3}_{n\ =\ 1,2,3..}$ **and** $\mathbf{LLRD4}_{n\ =\ 1,2,3..}$

→ **Select** Regions in the plots of **LLRD1, LLRD2, LLRD3** and **LLRD4 (***versus* n) as per the following criterion:

→ **Mark** the Sections along the sequences wherein the bases of all the four strains indicate almost similar (normalized) LLRD measures ≥ 0.6 within a span of say, at most 25 bases. Suppose the motif sections observed are identified as k = 1,2,…,7 (See Appendix).

% *Note:* The chosen upper threshold of ≥ 0.6 means that all the four sequences are almost identical to a probabilistic extent of 0.6 or better. Figure 7.2 shows the seven selected segments from 5' to 3' of the multiple sequences as per the above criterion

→ **Write** For each of the selected motif span at k = 1,2,…7, the prevailing nucleotide sequence is noted down:

→**Label** them as $\{(\mathrm{MSEQ})_{\mathrm{DEN}n,\ n=1,\ 2,\ 3,\ 4}\}_{k,\ k=\ 1,\ 2,\ \dots,\ 7}$

```
        → Add 10-15 bases before the start and after the end of each of the selected region (MSEQ) to account for stray end-effects
End
```

Pseudocode D2

```
%   Pseudocode on the procedure to determine the aligned and/or unaligned AA representation of the selected seven motif sections

Input
→   For each of the selected motif span at k = 1,2,…7, the nucleotide sequence labeled as {(MSEQ)DENn, n = 1, 2, 3, 4}k, k = 1, 2, …,  and 10-15 bases added before the start and after the end of each of the selected region (MSEQ) to account for stray end-effects

→   Find the regular expression for each end-corrected $(MSEQ)_{1, 2, …, K}$ section using PRATT tool:
    → (Available at: http://web.expasy.org/pratt/)
→   Determine: Position specific scores on each segment of k = 1,2,…,7

%   Note: For motifs specified by a segment (of k = 1, 2,…, 7), the position-specific scoring implies probability information of nucleotides at each position of the ungapped multiple sequence aligned and obtained via PRATT tool. Using such scoring considerations, the regular expression for each segment (k = 1, 2,…, 7) is translated into  a set of four nucleotide sequences constructed on the basis of the positional frequencies of each residue, so that the degree of sequence conservation at each position of the multiple alignment is duly taken into consideration.
    For example, given below is a regular expression obtained using PRATT and the corresponding four sequences of nucleotides for the window k = 1
```

Range	Aligned/ Unaligned	Regular Expression
1041- 1080	Aligned	A-[AG]-[AG]-[AG]-[AC]-A-A-[AG]-C-C-[AC]-A-C-[ACG]-[CT]-T-G-G-A-[CT]-[AT]-T-[AT]-G-A-[AG]-C-T-[CGT]-[ACT]-[ACT]-[AGT]-A-A-[AG]-A DEN1: AAGACAAACCAACACTGGACATTGAACTCTTGAAGA cgga DEN2: aa AAAACAAACCAACATTGGATTTTGAACTGATAAAAA ca DEN3: AGAACAAGCCCACGCTGGATATAGAGCTTCAGAAGA ccga DEN4:gccc AGGGAAAACCAACCTTGGATTTTGAACTGACTAAGA
	Unaligned	A-x(4)-A-A-x-C-C-x-A-C-x(2)-T-G-G-A-x(2)-T-x-G-A-x-C-T-x(4)-A-A-x-A

List

All PRATT-specified regular expressions $(RE)_k$ for each sequence k = 1, 2,…, 7 (Appendix)

Next

Perform: RE-to-nucleotide translation. A "select" set of nucleotide translations of the regular expressions obtained

% *Note:* Explanation:

- Each RE would, in general, provide multiple nucleotide translations
- However, at any position in a translated version, each base is seen with certain position-specific probability of occurrence across the columns of four sequences
- Based on such probabilistic considerations only, a restricted (select) set of translations are listed as per the procedure described below:

→**Procedure:** Consider for example, a partial regular expression:

← RE: … A-[CT]-G-A-A-C …

← It can be translated back to a set of four sequences as follows:

Set X: DEN1: ACG AAC
DEN2: ACG AAC
DEN3: ATG AAC
DEN4: ATG AAC

Set Y: DEN1: ATG AAC
DEN2: ATG AAC
DEN3: ACG AAC
DEN4: ACG AAC

In making the above sets {X} and {Y}, the implication of the equal probability of occurrence of C or T in [CT] is invoked.

Likewise, suppose in another example, if -[CTG]- is encountered at a position. Then, the possible combinations for position-specific translations will be:

```
  C       C         C             T       T      T
  C   OR  C    OR   C       OR   T  OR   T  OR   T
  C            C          C              T       T
T
 C          T         G              T       C       G
```

and so on.

→ Correspondingly, translated sets of {X}, {Y} etc. are framed for each $(RE)_{K = 1, 2, …, 7}$

→ Then, select numbers of sets are alone chosen as explained below:

→ For each $(RE)_k$, obtain the open-reading frames (ORF) and ascertain the ORF using the tool Sequence Manipulation Suite: ORF Finder (Available at: http://www.bioinformatics.org/sms2/orf_find.html

% (*Note:* Set the value of the parameter 'Only return ORFs that are at least 3 codons long instead of the default value '30'. Further, only forward reading frames are considered inasmuch as dengue is +ve single-stranded RNA virus)

List

→ All ORF's obtained are tabulated in terms of amino acids (AA) for each $(RE)_{k=1, 2, …,7}$

Next

→ For each subsequence numbered k = 1, 2, …, 7 and depicted in terms of AAs, manually select the AA sequence part found most commonly among all the ORF's obtained for each motif section, k = 1, 2, …, 7

→ Tabulate all such sorted (manually selected) AA segments obtained depicting a total of 36 motif sections (in the present analyses of dengue virus). The results on the identified motifs concerning DEN1 through DEN4 serotypes are presented in Table 7.1

END

Table 7.1: Manually selected sets of motifs of the four multiple sequences of DEN1 to DEN4

Range	Aligned/ Unaligned	Fitness	Regular Expression
1041-1080	Aligned	131.0950	A-[AG]-[AG]-[AG]-[AC]-A-A-[AG]-C-C-[AC]-A-C-[ACG]-[CT]-T-G-G-A-[CT]-[AT]-T-[AT]-G-A-[AG]-C-T-[CGT]-[ACT]-[ACT]-[AGT]-A-A-[AG]-A DEN1: AAGACAAACCAACACTGGACATTGAACTCTTGAAGA cgga DEN2: aa AAAACAAACCAACATTGGATTTTGAACTGATAAAAA ca DEN3: AGAACAAGCCCACGCTGGATATAGAGCTTCAGAAGA ccga DEN4:gccc AGGGAAAACCAACCTTGGATTTTGAACTGACTAAGA
	Unaligned	79.2310	A-x(4)-A-A-x-C-C-x-A-C-x(2)-T-G-G-A-x(2)-T-x-G-A-x-C-T-x(4)-A-A-x-A
1301-1350	Aligned	136.9046	A-C-[AC]-[CG]-[CT]-[ACGT]-C-A-C-[AT]-[AC]-[AT]-G-G-[AG]-G-A-[AC]-[ACG]-[AC]-[ACG]-C-A-x(1,2)-C-A-x(0,1)-G-T-[ACGT]-G-G-A-A-A-T-G-A-[ACG] DEN1: tagtc ACCGTACACACTGGAGACCAGCAc-CAaGTTGGAAATGAG acca DEN2: tgata ACACCTCACTCAGGGGAAGAGCAtgCA-GTCGGAAATGAC ac DEN3: tcatt ACAGTGCACACAGGAGACCAACAc-CAgGTGGGAAATGAA acgc DEN4: ttgta ACAGTCCACAATGGAGACACCCAtgCA-GTAGGAAATGAC
	Unaligned	94.9112	A-C-x(4)-C-A-C-x(3)-G-G-x-G-A-x(4)-C-A-x(1,2)-C-A-x(0,1)-G-T-x-G-G-A-A-A-T-G-A
3001-3050	Aligned	123.1960	G-G-[ACT]-C-C-[AT]-[AGT]-T-[ACGT]-T-C-[AT]-C-A-[AG]-C-A-C-A-A-[CT]-T-A-[CT]-[AC]-G-x(0,1)-C-C-[AC]-G-G-[ACG]-[CT] DEN1: atgga GGACCAATATCTCAGCACAACTACAGaCCAGGAT att DEN2: tcgct GGACCAGTGTCTCAACACAACTATAGaCCAGGCT a DEN3: tagct GGTCCTATCTCACAACACAACTACAGgCCCGGGT accac DEN4: atgcg GGCCCTTTTTCACAGCACAATTACCG-CCAGGGC
	Unaligned	87.0711	G-G-x-C-C-x(2)-T-x-T-C-x-C-A-x-C-A-C-A-A-x-T-A-x(2)-G-x(0,1)-C-C-x-G-G
3601-3650	Aligned	115.7843	C-[ACT]-[ACT]-[AGT]-G-G-x(1,3)-C-x(1,3)-T-[ACG]-[AT]-C-[AC]-T-[GT]-[GT]-A-[AGT]-[AGT]-G-A-[CT]-[ACT]-T-[AG]-[ACG]-[CGT]-[ACG]-[AC]-[AG]-[ACG]-[ACG]-[CT]-[ACG]-[ACT]-[GT]-[ACGT] DEN1: tctca CAATGGga-CaatTGACATGGAATGATCTGATCAGGCTATGT atca DEN2: ttgat CACAGGgaaCa--TGTCCTTTAGAGACCTGGGAAGAGTGATG gt DEN3: CTCAGGg--CaaaTAACATGGAGAGACATGGCGCACACACTA ataat DEN4: atcat CCTGGGaggCc--TCACATGGATGGACTTACTACGAGCCCTC
	Unaligned	43.8706	G-G-G-x(0,2)-A-x(2,4)-T-x(2)-C-x-T-x(2)-A-x(2)-G-A-x(2)-T
5541-5590	Aligned	148.7139	G-[ACGT]-T-C-[AG]-T-G-G-A-A-[CT]-[AT]-C-[AC]-G-G-x(2,3)-T-x(0,1)-G-A-[AC]-T-G-G-[AG]-T-[ACT]-A-C-[ACGT]-G-A-[CT]-T-[AT]-[CT]-[ACG]-[AC]-[AT]-G-G DEN1: tgaaa GATCATGGAACTCAGGctaT-GACTGGATCACTGATTTCCCAGG t DEN2: tgaac GTTCGTGGAATTCCGGacaT-GAATGGGTCACGGATTTTAAAGG ga DEN3: GCTCATGGAATTCAGGcaaT-GAATGGATTACCGACTTCGCTGG gaaaa DEN4: ggaaa GGTCATGGAACACAGGgt-TcGACTGGATAACAGACTACCAAGG
	Unaligned	103.2513	G-x-T-C-x-T-G-G-A-A-x(2)-C-x-G-G-x(2,3)-T-x(0,1)-G-A-x-T-G-G-x-T-x-A-C-x-G-A-x-T-x(5)-G-G

9701-9750	Aligned	159.5583	T-T-T-C-x(0,1)-A-[CT]-[ACG]-A-[AG]-[ACT]-T-[ACG]-[AT]-T-[CT]-A-T-G-A-A-[AG]-G-A-[CT]-G-G-[ACGT]-[AC]-G-[ACG]-[AGT]-[ACT]-[AG]-[ACT]-T-[ACG]-G-T-[GT]-G-T-[GT]-C-C DEN1: cacca TTTCcACCAGCTGATTATGAAGGATGGGAGGGAGATAGTGGTGCC DEN2: acca TTTCcATGAGTTAATCATGAAAGACGGTCGCGTACTCGTTGTTCC a DEN3: TTTC-ATGAATTGATCATGAAAGATGGAAGAAAGTTGGTGGTTCC ctgca DEN4: c TTTC-ACAAGATCTTTATGAAGGATGGCCGCTCACTAGTTGTTCC atgta
	Unaligned	103.7513	T-T-T-C-x(0,1)-A-x(2)-A-x(2)-T-x(2)-T-x-A-T-G-A-A-x-G-A-x-G-G-x(2)-G-x(5)-T-x-G-T-x-G-T-x-C-C
9751-9800	Aligned	135.4156	A-[CT]-G-A-A-C-T-[AGT]-[AG]-T-[AT]-G-G-[ACGT]-A-G-[AG]-G-C-[AC]-[AC]-G-A-[AG]-T-[AC]-T-C-[ACGT]-C-A-[AG]-G-G-[AC]-G DEN1: ccaag ATGAACTTGTAGGTAGGGCCAGAGTATCACAAGGCG DEN2: ccaag ATGAACTGATTGGCAGAGCCCGAATCTCCCAAGGAG c DEN3: ccagg ACGAACTAATAGGAAGAGCAAGAATCTCTCAAGGAG cggga DEN4: ccagg ATGAACTGATAGGGAGAGCCAGAATCTCGCAGGGAG ctgga
	Unaligned	95.9112	A-x-G-A-A-C-T-x(2)-T-x-G-G-x-A-G-x-G-C-x(2)-G-A-x-T-x-T-C-x-C-A-x-G-G-x-G

Figure 7.2 depicting the seven motifs (indicated via A, B and C) that constitutes the fingerprint of the test multiple sequence set is shown below. Figure 7.3 and 7.4 are the flowcharts for the pseudocodes D1 and D2 respectively.

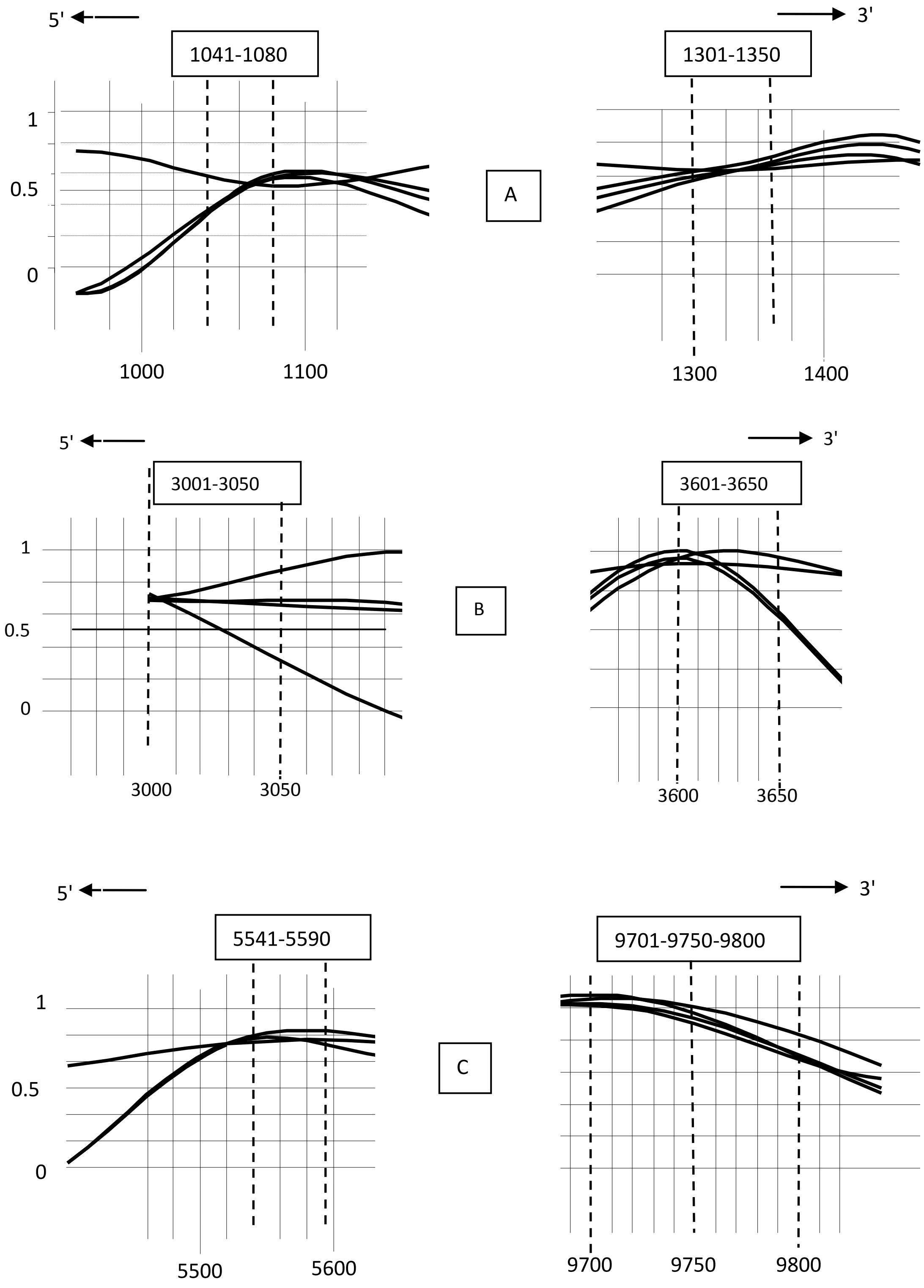


Figure 7.2: Selected (seven) motif segments across the multiple (four) RNA sequences of the test serovar in the entire 5'-to-3' stretch. They exhibit almost similar features

(measured ≥ 0.6 in the scale 0-1) estimated by the cohesive combination of entropy, energetics and Fourier spectral methods

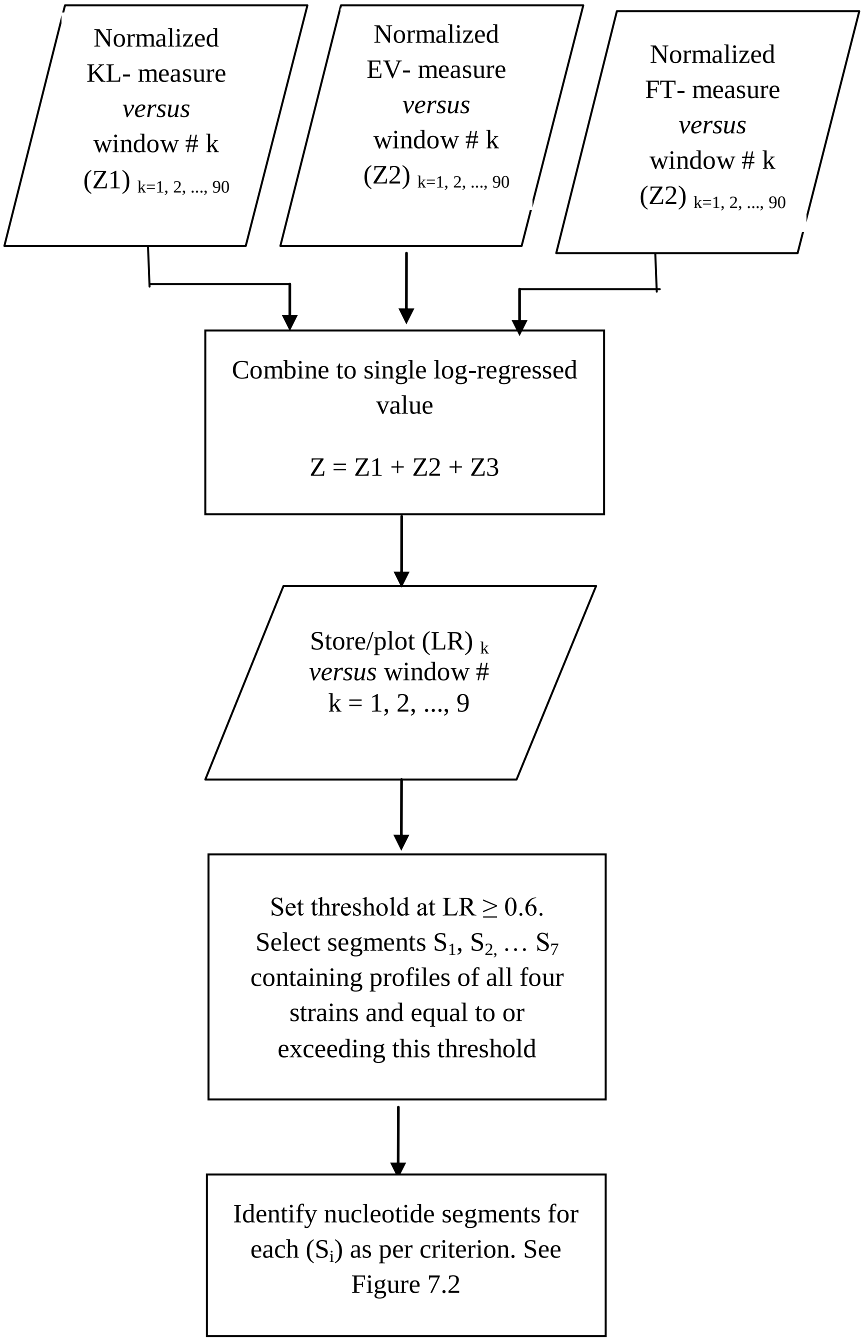

Normalized
KL- measure
versus
window # k
(Z1) k=1, 2, ..., 90
Normalized
EV- measure
versus
window # k
(Z2) k=1, 2, ..., 90
Normalized
FT- measure
versus
window # k
(Z2) k=1, 2, ..., 90
Combine to single log-regressed value
Z = Z1 + Z2 + Z3
Store/plot (LR) k
versus window #
k = 1, 2, ..., 9
Set threshold at LR ≥ 0.6.
Select segments S1, S2, … S7 containing profiles of all four strains and equal to or exceeding this threshold
Identify nucleotide segments for each (Si) as per criterion. See Figure 7.2

Figure 7.3: Flowchart presentation of Pseudocode D1

Flowchart shown in Figure 7.4 describes the salient aspects of the Pseudocode D2.

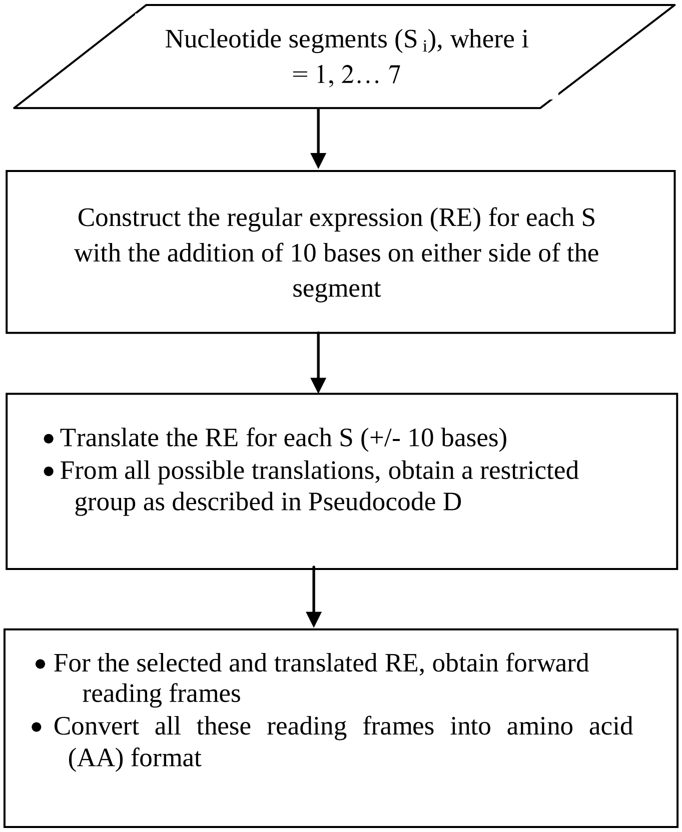

Nucleotide segments (S_i), where i = 1, 2… 7
Construct the regular expression (RE) for each S with the addition of 10 bases on either side of the segment
• Translate the RE for each S (+/- 10 bases)
• From all possible translations, obtain a restricted group as described in Pseudocode D
• For the selected and translated RE, obtain forward reading frames
• Convert all these reading frames into amino acid (AA) format

Figure 7.4: Flowchart presentation for Pseudocode D2

7.5 Finding common proteins among the test viral strains

The method of ascertaining a set of proteins that may prevail most commonly in all the four viral strains using the selected set of motifs as above is described in Pseudocode E below.

Pseudocode E

```
%    Pseudocode on the procedure to find the common proteins across all the four viral strains using the selected set of 36 motif sections listed in Table 7.2

Initialize

→    List the 36 motifs from Pseudocode D in AA format
→    For each motif the selected set of AA segments are subjected to BLAST search to ascertain the underlying homology and similarity. Relevant blastp compares an AA query sequence against a protein sequence database

List

→    Protein sequences obtained as output from BLAST search and corresponding expected value (E-value) estimated from BLAST for each AA sequence

Next

→    Tabulate the AA sequence, protein type and E-values obtained from the BLAST search. A low E-value indicates that a score has high confidence level
→    Filter out all those sequences that post high E-values
→    Prepare a condensed set of details as in Table 7.2 with reference to DEN1, DEN2, DEN3 and DEN4.

END
```

Table 7.2: Probabilistically most common motif sections (denoted in terms of AAs) and possible proteins synthesized by them across all the four dengue

serotypes. (Only 13 out of 36 are listed: See the note presented beneath the table)

Identified motif section: Highly probable amino acid (AA) sequences translating into proteins	Translated protein type from the AA sequence	Basis of selection priority	Typically present in:
KPTLDIELMK	envelope protein [dengue virus]	1	Dengue virus
KPTLDFELMK	polyprotein [dengue virus]	1	Dengue virus
TVHTGDQCK	polyprotein [dengue virus 1 and 3]	1	Dengue virus
TVHTGDQRK	polyprotein [dengue virus 1 and 3]	1	Dengue virus
SQHNYRPG	polyprotein [dengue virus 1 and 2]	1	Dengue virus
QDELIGRARISQG	polyprotein [dengue virus 2, 3 and 4]	1	Dengue virus
FHELIMKDGRELVV	polyprotein [dengue virus 1 and 3]	1	Dengue virus
FHELIMKDGSELVV	polyprotein [dengue virus 1, 2 and 3]	1	Dengue virus
QTNIGF*	FER-1-like protein 5 [Homo sapiens]	2	Human protein
QTNTGF*	nuclear pore complex protein Nup98-Nup96 isoform 1 [Homo sapiens]	2	Human protein
SQNLTR	dual specificity phosphatase 22 variant [Homo sapiens]	2	Human Protein
SQNLSR	Chain A, crystal structure of The human phosphatase (Dusp9)	2	Human Protein
SQNLAR	unnamed protein product [Homo sapiens]	2	Human Protein

Note:
(a) The analysis pursued also indicates existence of other 23 possible motifs in the four dengue viral sequences. But, they are not observed in the **BLAST** search to express proteins prevalent in either in dengue viral strains and/or in human. As such, they are not listed here as viable entities for vaccine design considerations on the dengue viral strains.
(b) Selection priority: The priority 1 or 2 indicated refers to the order of observations with low-end of E-values seen in BLAST tool results
(c) The results posted are obtained by using the protein **BLAST (blastp)** option of the **BLAST** tool

Presented in Figure 7.5 is a flowchart summary on Pseudocode E.

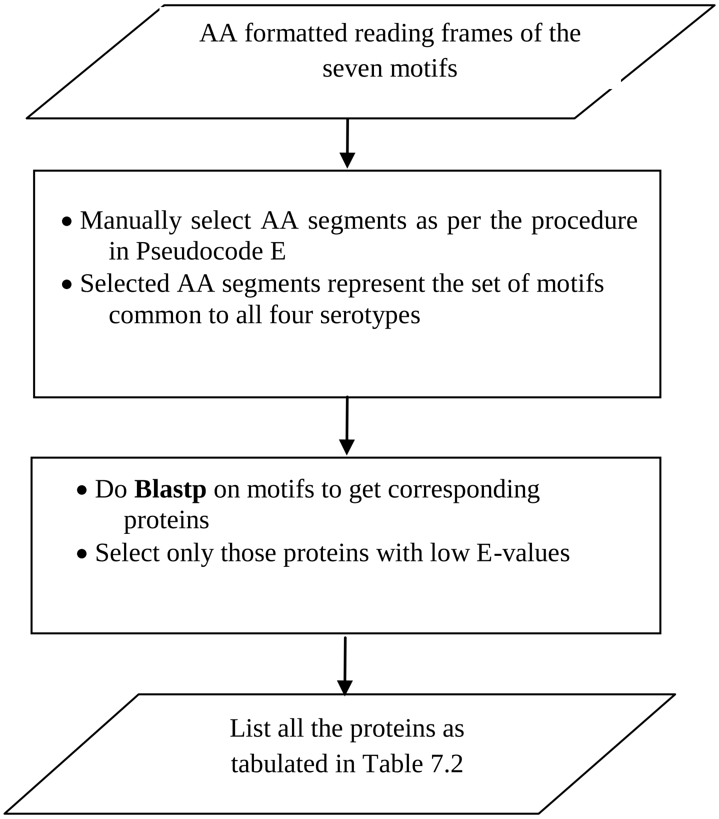

AA formatted reading frames of the seven motifs
• Manually select AA segments as per the procedure in Pseudocode E
• Selected AA segments represent the set of motifs common to all four serotypes
• Do Blastp on motifs to get corresponding proteins
• Select only those proteins with low E-values
List all the proteins as tabulated in Table 7.2

Figure 7.5: Flowchart presentation for Pseudocode E

7.6 Results

The thematic objective of the present study as stated earlier refers to applying cohesively the three methods of information-theoretic (entropy), thermodynamic-kinetics (energetics) and Fourier-spectral methods to DNA/RNA structures in order to obtain a comprehensive, featured portal that identifies and classifies distinguishable details buried across them. Currently, when this approach is specifically addressed to the serogroup of dengue virus, the resulting data obtained are listed in Table 7.1.

7.7 Viable use of the present study

The gene expression in a virus morphs to different patterns at the molecular (DNA/RNA) level across its different serotypes. Such discernment features in single viral diversity indicate the need for distinct vaccine designs *vis-à-vis* differentiable pathology of the strains concerned. Diverse vaccines can be synthesized by considering the distinct underlying DNA signature features of each serotype of a given virus. For example, relevant DNA feature can be determined in terms of the expression seen in each viral DNA/RNA structure observed as: (i) CDS, CpG, TATA box features; (ii) sites of homology specified by the spatial-spectrum (in Fourier domain); (iii) long-range correlation of coding/noncoding segments and (iv) individual nearest-neighbor energetic-interactions plus stability-seeking bends/loop formation (as in the case of single-strand DNA or in RNA sequences).

7.8 Closing Remarks

The present study is motivated by the interest in seeking a strategy towards comprehending the inner details of genomic information *via* a framework of multiple analyses applied simultaneously on a test sequence. Hence, a three-prong approach based on entropy, energetics and Fourier-transform methods is advocated and applied to RNA sequences of dengue viral strains. The collective details obtained thereof indicate most probable set of common protein-translated motif sections among the four serovar sequences in question. These can be viably used as indicators or biomarkers in the efforts of finding a common vaccine for the serotypes in question, coping with the diversity across a single virus.

Disclaimer: The list of proteins, motif sections etc. deduced in this paper (Table 7.1) and indicated as possible useful entities in designing a common vaccine for dengue serovar is based on analytical methods, computational procedures and various assumptions specified in making selective ensembles of data as necessary. No wet-lab studies were conducted to supplement or cross-validate the details furnished.

CHAPTER VIII

VIRAL GENOMIC SEQUENCES AND VACCINE DESIGN CONSIDERATIONS

8.1 Introduction

Disease prevention is of utmost importance for public health. It is always better to prevent than to treat the disease. A way to prevent a disease is to administer the susceptible individuals with what is known as the “vaccine”. Such vaccines can protect both, the people who receive them as well as those with whom they come in contact. A vaccine is any preparation used as a preventive introduction of an antigen (microorganism or an agent of disease into an host organism) to confer immunity against a specific disease, usually employing an innocuous form of the disease agent, such as a killed or weakened bacteria or virus, which can stimulate antibody production. Vaccine antigens are not strong enough to cause disease but still are strong enough to make the immune system to produce antibodies against them. Thus, the goal of any vaccination is the induction of an appropriate and effective immune response in the vaccinated person. However, it is still unclear what precisely exactly constitutes an effective immune response for many diseases so that an appropriate design can be attempted towards the prevention of the disease.

As discussed in earlier chapters, the gene expression in a virus morphs to different patterns at the molecular (DNA/RNA) level across its different strains. These discernment features offer a viable opportunity to conceive a set of distinct vaccine designs usable to prevent the differentiable pathology likely to be caused by the strains of the virus concerned. In this study, it is hypothesized such diverse vaccines can be intelligently synthesized by considering the underlying DNA signature features of the various strains of a given virus. Essentially, the unique expressions seen in each viral DNA/RNA structure as regard to its CDS, CpG, TATA box etc., sites of homology can be the entities of interest in vaccine design. These entities compared to those specified by the spatial-spectrum (Fourier domain) details, long-range correlation of coding/noncoding segments and nearest-neighbor energetic-interactions and related stability-seeking bends/loop formation (in the case of single-strand DNA or in RNA sequences). For example, the computed data on the distinguishable features pertinent to the RNA structures of dengue virus serovar as ascertained in Chapter VII can be used in conceiving a smart/ rational vaccine design.

8.2 What is Immunity? An Overview

Immune response is the result of a series of biochemical reactions that provides protection mechanism for the body from potentially dangerous pathogens and foreign substance. Two types of immunity exist, innate and acquired. Innate immunity is always present and consists of an intricate system by which infection is recognized with the production of antimicrobial/viral activities and recruitment of neutrophils and other phagocytic cells to the site of infection so as to kill or neutralize the invading pathogens [8.1-8.3]. Such activities are triggered by the presence of cell-surface and internal receptors (what are known as Toll-like receptors or TLRs) that recognize certain pathogen-associated molecular patterns present in or on the surface of pathogens. Acquired immunity is induced in response to the invading pathogens; but it is dependent on the innate immune system to facilitate the presentation of pathogenic antigens in order to further trigger either the production of antibodies or stimulating the cellular immunity [8.4].

The cellular part of immune system includes certain types of white blood cells (WBCs) called *lymphocytes*, (which in turn consist of B-cell and T-cell) and antibodies. Antigens are particles or protein molecules on the surface of pathogens that induce production of antibodies by the immune system. The cells of B series (including plasma-cells) produce antibodies. Antibodies attach to a specific antigen and make it easier for the immune cells to destroy them. A part of the antibody known as the paratope recognizes an antigen, or better its critical part (amino-acid sequence) known as epitope. The T-cells attack antigens directly and help in controlling the immune response. These T-cells are either helper (Th) or cytotoxic T(c) versions and they specifically kill host cells in which a pathogen resides [8.5]. They also release chemicals, known as *cytokines*, which control the entire immune response, including production of antibodies from B-cells.

The biochemical reactions during immune response produce and select particular epitopes from antigenic material or antigen presenting cell (APC). The epitope is a peptide that can be recognized by a T-cell and elicit an immune response against the foreign body. T-cell immunity is essential for the induction of long-term protective immunity against infectious disease agents [8.6-8.8], and the paratope part of the antibodies binds to the epitope by a lock and key mechanism [8.9]. The epitope can be continuous or discontinuous [8.10] as has been indicated in Figure 8.1. The lock-and-key feature of paratope-epitope structure is presented in Figure 8.2.

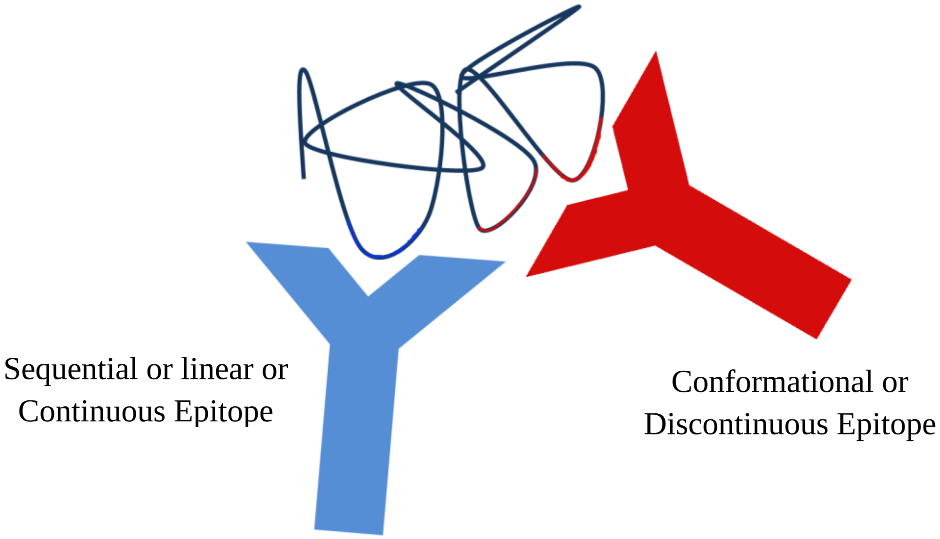


Figure 8.1: Types of Epitopes

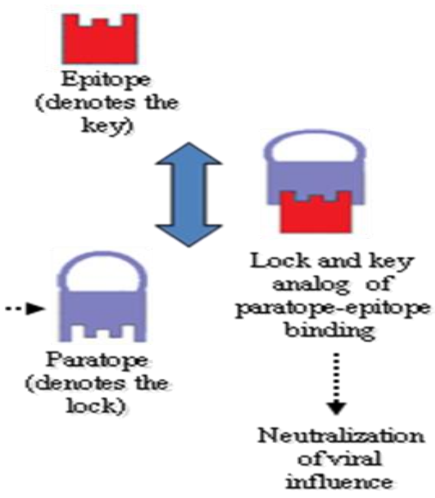


Figure 8.2: Lock-and-key mechanism of epitope-paratope

Once B-cells and T-cells are formed, a few of these would multiply and provide a "memory" for the immune system [8.11] on the invading pathogen. This allows the immune system to respond faster and more efficiently when the body is exposed subsequently to the same antigen; and, in many cases it will prevent the host from getting sick due to the invading pathogen or at least minimize the severity of the related infection.

The possibility of immune response of a person who acquired immunity to a certain pathogen is shown in Figure 8.3.

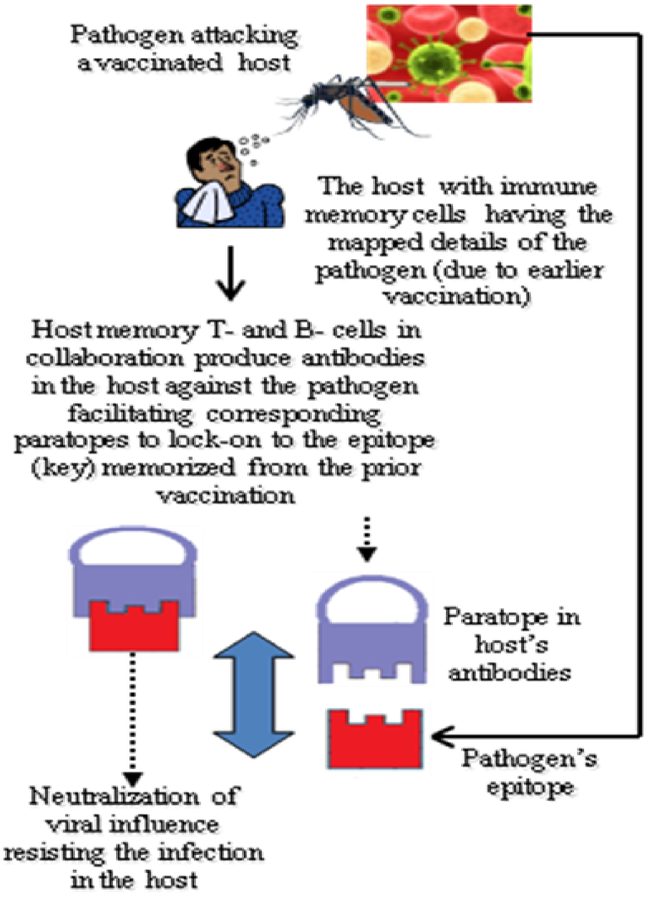

Pathogen attacking a vaccinated host
The host with immune memory cells having the mapped details of the pathogen (due to earlier vaccination)
Host memory T- and B- cells in collaboration produce antibodies in the host against the pathogen facilitating corresponding paratopes to lock-on to the epitope (key) memorized from the prior vaccination
Paratope in host's antibodies
Pathogen's epitope
Neutralization of viral influence resisting the infection in the host

Figure 8.3: Immune response generated by a previously vaccinated person's body (host) to a pathogen

8.3 Vaccine and Vaccine Designs: A Review

Consistent with the immunity response considerations as indicated above, a vaccine is an antigen that prepares an immune system for future protection against some pathogen without causing severe symptoms in the host. Vaccines can be of various forms such as: An organism (bacteria or virus), a protein (or peptide), or a nucleic acid sequence [8.12], [8.13]. Thus, the goal of vaccine design is to create an artificial means to produce immunological memory of a particular pathogen without the risk of developing the disease in the host and offer a resistance against future possible infections.

The science of *vaccinology* started with Edward Jenner's systematic investigations into the protective effect of cowpox against smallpox in the late 18th century [8.14], [8.15]. However, the momentum was provided by Louis Pasteur [8.16], [8.17] for further research. Empirically vaccines were designed on the basis of Pasteur's principle: "isolate, inactivate and inject" [8.18] and is being practiced till date. The traditional vaccine design approach is to derive all possible vaccines from a protein sequence and test each of them for an immune response experimentally. The vaccines are designed in specialized laboratories and tested *invitro*. However, this strategy is not only expensive, time-consuming and often unpredictable, but also it may fail in producing effective vaccine solutions against some pathogens, especially those with very high antigen variations [8.19].

To reduce the high costs associated with traditional vaccine development, *in silico* based methods of vaccine design is proving to be useful. With the advent of modern technologies plus bioinformatic tools and databases developed for proteomics, comparative genome analysis and interpretation of whole-genome sequences, computer-based techniques for designing or predicting epitopes for effective vaccine is increasingly gaining momentum. It can be used to assign putative gene functions to each open-reading frame (ORF) on the basis of homology to known proteins. Systematic identification of potential antigens of a pathogen using this information without the need for cultivation of the pathogen, is termed as 'reverse vaccinology' [8.20].

Vaccine history teaches that advances in vaccines are closely tied to the development of new and improved technologies. By understanding how vaccinology has evolved, innovative strategies—informed by novel technologies—can be pursued and applied towards the development of safe and effective vaccines against chronic infections, highly variable pathogens, or non-infectious diseases such as cancer, lupus, Alzheimer’s disease etc. Hence, a brief overview on the conventional as well as on recent novel strategies indicated for designing vaccines is discussed in following sub-sections.

8.3.1 Live, Attenuated and Inactivated Vaccine

The live attenuated vaccines consists of the whole attenuated bacterial or viral particles. This attenuation is achieved by growing generations of the pathogens in cells in which they do not reproduce very well. As they evolve to adapt to the new environment, they become weaker with respect to their natural host, human beings. Its major advantage is in the fact that they contain the entire possible antigenic spectrum as that of the pathogen, they induce most effective and long lasting immunity against the infection. However, on rare occasions and especially in immune-compromised or immune-deficient host, these pathogens introduced through vaccines can prove fatal. Thus, this conventional method, while successful against some pathogens (especially against certain viruses like chicken pox, mumps etc.), failed to provide a solution for many of those pathogens (especially bacteria which has thousands of genes) for which a vaccine is not yet available.

Inactivated vaccines are prepared by killing the disease-causing pathogen with chemicals, heat or radiation. These vaccines are very stable and safer than live vaccines as the dead pathogen can’t mutate back to their disease-causing state. However, most inactivated vaccines stimulate a weaker immune system response as compared to live attenuated vaccines and so it takes several doses or booster shots, to maintain immunity.

8.3.2 Subunit and Conjugate Vaccine

Subunit vaccines include only the epitopal or antigenic part that best stimulate the immune system. As they do not contain all the other molecules that make up the pathogen, the chances of adverse reactions to the vaccine are lower. Hepatitis B vaccine is an example of subunit vaccine.

Conjugate vaccines are a special type of subunit vaccine used for immunization against bacteria that possesses an outer coating of sugar molecules called polysaccharides. A polysaccharide coating on bacterial surface hides the antigens, so that the immature immune systems of infants and younger children can’t recognize or respond to them. In preparing conjugate vaccine, the antigenic material inside the polysaccharide coating of the bacteria is replaced by an antigen that an infant’s or young child’s immune system can recognize. Thus, the infant’s or child’s immune system responds to the polysaccharide coatings and defends against the disease-causing bacteria.

8.3.3 DNA and Recombinant Vector Vaccines

A novel class of vaccines that specially deserves to be mentioned is based on the immunization with "naked" DNA was introduced in late 1990’s. It’s main strength is that it highly reduced or eliminated the disadvantages associated with conventional vaccines designed using live attenuated or inactivated pathogen [8.21]. Direct inoculation of plasmid DNA encoding sequences of viral proteins results into the synthesis of the proteins causing immune responses in the host. Several advantages are associated with DNA immunization, e.g., cheap to produce, heat stability, amenable to genetic manipulation, mimic viral infection, and no risk of reversion to pathogenicity. Some concerns remain regarding their safety, e.g., the possible integration of plasmid DNA into host chromosomes [8.22].

Recombinant vector vaccines are experimental vaccines similar to DNA vaccines, but they use an attenuated virus or bacteria to introduce antigenic DNA to cells of the host's body. “Vector” refers to the virus or bacterium used as the carrier. This DNA vaccines do not use DNA as immunogen (as it is proposed in [8.12] and [8.13]) but just a critical DNA sequence inserted in plasmid in order to fabricate sufficient amoubt of protein (epitopal) copies for given epitope. This is the crucial distinction between the two concepts. A detailed review of the history of vaccinology is provided in [8.23].

8.4 Rational Design of Vaccine

Along with the above mentioned ways to design vaccines in biology laboratories, reverse vaccinology has fast gained acceptance [8.24]. Technological breakthroughs in molecular biology in the past decade have led to the generation of immense amounts of data and research at the genomic or proteomic level. This has given rise to an increase in the popularity of immune-informatics [8.25] and computational research methods, which are often faster than biological experiments as it reduces the time required for the identification of critical candidate vaccines and provides new solutions for those vaccines which have been difficult or impossible to develop. The two techniques central to this new discipline of rational vaccine design are epitope mapping and reverse vaccinology. Rational vaccine design uses computers for large scale screening of potential vaccines based solely on genomic and proteomic information [8.26] by using intelligent algorithms and tools for setting up a standardized approach. Today, the possibility of using genomic information allows us to study vaccine development *in silico*, without the need of cultivating the pathogen. The whole concept has been summarized in Figure 8.4.

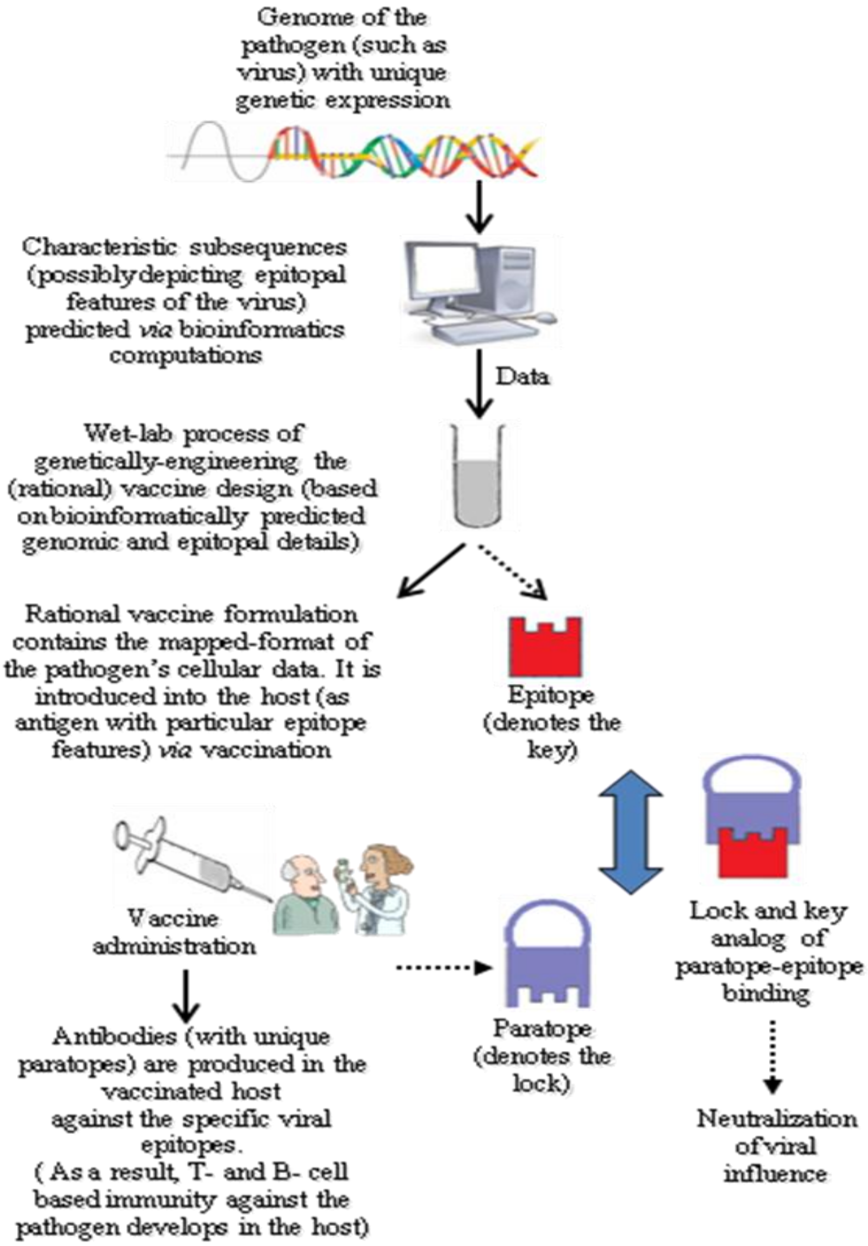

Genome of the pathogen (such as virus) with unique genetic expression
Characteristic subsequences (possibly depicting epitopal features of the virus) predicted via bioinformatics computations
Data
Wet-lab process of genetically-engineering the (rational) vaccine design (based on bioinformatically predicted genomic and epitopal details)
Rational vaccine formulation contains the mapped-format of the pathogen's cellular data. It is introduced into the host (as antigen with particular epitope features) via vaccination
Epitope (denotes the key)
Lock and key analog of paratope-epitope binding
Vaccine administration
Paratope (denotes the lock)
Antibodies (with unique paratopes) are produced in the vaccinated host against the specific viral epitopes. (As a result, T- and B- cell based immunity against the pathogen develops in the host)
Neutralization of viral influence

Figure 8.4: Concept of rational design of vaccine

Various computational approaches and algorithms have been proposed and used for recognition/design of possible epitopal candidates and reverse vaccinology. Epitope identification is a kind of pattern recognition and the computational methods used for pattern recognition are applicable [8.27]. Other methods to determine antigens use techniques such as neural network [8.28], support vector machines [8.29], Hidden Markov models (HMM) [8.30] and many more. A good review of various methods and the immunological entity that can be determined through these algorithms and tools is in literature [8.31].

8.5 Rational Design of Vaccine for Dengue Virus

As already stated earlier, exposure to one serotype of dengue virus usually results into mild symptoms which goes away easily and the patient is immune for life to a second infection of the same serotype. However, subsequent exposure to a second dengue serotype increases the chance of the illness progressing to the severe and sometimes fatal dengue hemorrhagic fever. Although many different approaches have been tried to develop vaccine effective against all the four serotypes of dengue virus, but, none have proved effective. Cardosa has provided a good review of the issues and challenges in designing a vaccine against dengue virus [8.32]. Individual vaccines though are effective against a particular serotype, but, combining them together effectively to form a reliable vaccine against all four serotypes has been unsuccessful. Various researchers have made attempts to develop vaccine against all the four serotypes using many innovative methods and substantial progress has been made towards finding a vaccine, yet, each comes with its own unique challenges and benefits and an effective tetravalent vaccine has not been developed till date. A comprehensive review on this topic is by Durbin and Whitehead [8.33]. The most recent clinical trial of the vaccine developed against dengue by Sanofi Pasteur researchers showed promising results against dengue virus of serotypes DEN1, DEN3 and DEN4, but, proved ineffective against DEN2 [8.34].

The goal of this research is to design or suggest a minimal set of immunogens capable of inducing a robust and sustained immune response through computational approach. The potential of this new strategy is illustrated in this thesis for the development of a vaccine against all the four serotypes of dengue virus. Especially, for a particular pathogen (such as virus), which may prevail in different forms of serotypes, (as in the case of dengue 1 to 4 viral strains) with genomic variability as well as common genomic features, finding such common genomic information of a serogroup may be useful in knowing epitopal details for designing unique vaccines for the immunity across multiple serotypes.

8.6 Limitations of Rational Vaccine Design

- A limitation of rational vaccine design is that it models only short biological sequences, which often does not work so effectively as vaccine prepared from an live pathogen
- Under certain conditions immunogenic epitopes can do more harm than good and might therefore be considered pathogenic. the specific removal of such pathogenic epitopes from vaccines might increase their prophylactic potential, while minimizing the risk of side-effects from vaccine use. [8.35]

8.7 Closure

Despite the overall success of vaccination efforts in this modern era, there is still a great need for completely new and improved old vaccines. The predictive accuracies of the computational methods increase as more data and knowledge about the mechanisms of immune response become available. Therefore, rational vaccine design can be expected to become a common approach in immunology laboratories in the future. However, the predictions of rational vaccine design should always be confirmed by biological experiments before conclusions are drawn.

CHAPTER IX

INFERENTIAL CONCLUSIONS AND OPEN-QUESTIONS FOR FUTURE RESEARCH

9.1 General

Summarized in this chapter are essential details on the perceived objectives of the research and the results obtained. Relevantly, chapter-by-chapter efforts are briefly revisited and corresponding outcomes of the research are enumerated. Discussions on the efforts indicated across various chapters are highlighted and salient conclusions are listed. Further, the host of possible research that can be undertaken as future efforts are identified along with motivating considerations.

9.2 An Overview on the Research

In Chapter I, relevant to the topic of research, namely 'Bioinformatic Analysis of Viral Genomic Sequences: Application to Rational Vaccine Design', introductory details on the analysis of viral genomic sequences are presented along with the scope and objectives, motivated essentially by the considerations on elucidating subtle features in test sequences of certain viruses and their serotypes. Indicated thereof is the cohesive use of bioinformatic approaches involving entropy, energetics and spectral-domain methods to determine the underlying common and/or differential features across the test viral sequences; and, the health-science related objective is indicated to identify the salient signatures that can be seen in viral sequences, which are useful in synthesizing appropriate vaccines.

To support the conceived research on the aforesaid genomic sequence analyses, the existing details on biological sequence information, entropic aspects of genomic structures, and energetic profiles of genomes and spatial spectral-domain considerations (in representing the genomic features) are elaborated in Chapter II.

9.3 Viral Genome Considerations

Pursuant to the generic details on genomic sequence analyses reviewed in Chapters I and II, a directed effort exclusively to address the viral genomic features is reviewed in Chapter III. Basic definitions, classifications and structural aspects of viral genomes are presented; and, the intricate details on the double-stranded (ds-) and single-stranded (ss-) DNA genomes are described. Specifically, particulars concerning ssDNA of Parvovirus B19V and ssRNA structure of dengue viral serotypes (DEN1, DEN2, DEN3 and DEN4) are considered. The existence of multiple strains or serotypes implicating anti-viral vaccine design is outlined.

For the purpose of research pursued, the aforementioned Parvovirus B19V and the serotypes of dengue virus are regarded as test species throughout the study.

9.4 Viral Genome Analysis: Entropic Considerations

Consistent with the scope, objectives and the suite of test-viral genomes mentioned above, the first research task undertaken is described in Chapter IV. It refers to applying entropy considerations (in Shannon’s sense) to extract the genetic information that prevails in the test sequences. The methods pursued and results obtained in Chapter IV thereof, are concerned with the following:

- Applying various statistical distance/divergence methods (such as Kullback-Leibler (KL) measure) and Shannon's information redundancy (R) considerations to the test genomic sequences

- Revisiting various entropy concepts and statistical ordering of residue structures in their sequences, so as to identify both parametric and non-parametric divergence matrices compatible for: (i) discriminating informative and non-informative sub-segments in a sequence; and (ii) elucidating similar/dissimilar features across a set of sequences

- Finding splice-junctions between codon-noncodon segments in line with earlier studies due to Arredondo and Neelakanta [9.1]: Specifically, a new approach is indicated to describe the fuzzy transitions at the splice-junctions in terms of a spatial jitter algorithm

- Predicting splice-junctions (canonical and cryptic versions) in viral sequences: The possibility of aberrant splice-junctions appearing in viral sequences as a result of mutation is indicated with reference to DEN1 virus taken as a case-study example. Hence, the complete locations of splice-junctions in DEN1 viral sequence are identified and compared with NCBI GenBank data along with the presence of cryptic sites depicting fuzzy subspaces

- Extending the concept of information-theoretic description of genomic statistics in terms of Shannon's information redundancy factor (R): Relevant results on DEN1-DEN4 viral sequences are presented in Chapter VII

- Deducing the efficacy of information-theoretic approach to identify CpG islands: Relevant computational methodology and results are presented with respect to Parvovirus B19V ssDNA, invoking mutual entropy measures such as Jensen-Shannon measure

9.5 Viral Genome Analysis: Energetics Considerations

The devoted effort in Chapter V is concerned with a method to deduce genetic statistical features using the underlying parametric details in genome sequences. Specifically, the parametric profile considered refers to the inherent minimum potential-specific energetics that can be seen across the sequence, wherein the positions of the entities in the set {A, T, G, C} present themselves to attain the minimum potential energy states, so as to guarantee stability of the system. For example, considering an ssDNA, wherein no complementary strand is present (to facilitate the Watson-Crick (WC) pairing and hence, the associated energy minimization), such ssDNA structures "bend" themselves into different forms (like hairpin bends, loops, bulges etc.), so that the WC matching and related palindromic signatures are eventually sustained.

In order to apply such free-energy dynamics of nearest-neighbor nucleotides, the Parvovirus B19V ssDNA is considered; and, the statistics of nearest-neighbor (NN) energy profile is invoked to locate the sites of loop, hairpin and bulge formations along the sequence. Normally, either dot-plot or Nussinov algorithms are used to determine the secondary structural features of ssRNA/ssDNA. In context, the present study uses a new method based on the statistics of NN-E energy and statistical divergence methods to determine the hairpin bend/bulges in the test sequence. This method is relatively simple in its computational procedure. The results agree with available data [], demonstrating the efficacy of the method pursued.

9.6 Viral Genome Analysis: Fourier Spectral Characteristics

Another method of making use of parametric value representation of a genomic sequence is concerned with deducing the Fourier transform of the parametric values (numerical) assigned to the residues in the sequence; and, extracting the spectral features is based on short-term Fourier transform (STFT) method. Hence, considering the DEN1 genome sequence, appropriate spectral-domain results are obtained as discussed in Chapter VI. The parametric/numerical data adopted thereof conform to the so-called *electron-ion interaction potential* (EIIP) values as advocated in [9.2]. The methodology of such (parametric domain)-to-(Fourier domain) conversion and relevant computations forms the gist of efforts addressed in Chapter VI.

9.7 Cohesive Analysis of Genomic Sequences

Considering cohesively all the three methods of genomic sequence analysis using entropy, energetics and Fourier transform technique, a comprehensive approach is detailed in Chapter VII to compare the four test sequences of DEN1-DEN4 viruses. Hence, the obtained portal of data mining has enabled finding distinguishable details buried across them. Using the results so obtained, the common proteins across all the four viral strains are deduced *via* the motifs specified by the BLAST tool.

In the existing literature, rigorous computational search on such data-mining is not available to obtain probabilistically the most common motif structures across a set of genomes like dengue serotypes. Hence, the approach of deducing common proteins across the four test genomes as identified in this study is novel and useful as a viable framework in designing a single vaccine for all the four serotypes, coping with the diversity seen across a single virus.

9.8 Vaccine Design Considerations

The principle of vaccine design, related existing methods and futuristic considerations are reviewed and described in Chapter VIII. It is of a comprehensive review on the related topics of interest. The question of rational vaccine design is also considered *vis-à-vis* the proposed efforts of the present study.

9.9 Scope for Future Studies

Motivated by the needs to address DNA/RNA analysis specific to viral species, the present study is developed with the objectives and the scope outlined earlier. Concurrently, researched in this study are the prospects of developing vaccines from the data-mined details and the associated meta-learning frameworks of viral sequence analysis in the contexts of vaccine designs.

The possible topics that can be considered as viable for future research can be identified in terms of the associated open-questions. Relevantly, the following research topics are indicated for future studies:

- Though the entropy method is adopted in the present study on viral sequences, yet, another similar parallel analysis can be undertaken by applying Fisher linear-discriminant method. Relevant approach is indicated by Arredondo in [9.3] to identify the CDS in ssDNA. More elaborate pursuits and fuzzy considerations pertinent to cryptic splice-junctions, untranslated region (UTR) etc. need to be investigated further

- There are a host of viruses for which vaccine designs are imminent. For example, the rationale for epitope-based vaccine design against sapovirus is predicted in [9.4], where, T- and B- cell epitopes are predicted on the capsid protein of sapovirus by immuno-informatics approach. Relevant proteomic level studies for example, homology modeling of the 3D-structure of sapovirus capsid proteins and the related similar folding-patterns exhibited by norovirus and vesivirus can be attempted as adjunct bioinformatic study on viruses

- In order to predict epitopes from a set of proteins (for example HBx) for vaccine designs, computer-aided approach as in [9.5] can be considered and applied across various viral species. Essentially, the maximum expression of sequence can be ascertained *via* codon and CpG optimization efforts; and relevant modeling plus pattern searching. It enables the identification of helix, sheets and turns that carry the predicted epitopes

- As indicated in the present study, given the results on a sequence obtained by different methods (such as entropy, energetic and Fourier transform), such results can be assessed by logistically regressing them using a logit function. Applying logistic regression on multiple data mined *via* different techniques on a specific framework (like a biological sequence) has to be studied further. Specifically, obtaining an upper and lower bound on the combined data as considered in the present study needs more efforts

- For local analysis of short-segments, use of wavelet transform (discrete or continuous) can be considered. For example, the wavelet transform of "DNA walk" constructed from a genomic sequence [9.6], wavelet-based fractal analysis of DNA sequences [9.7], signal representation of DNA sequences using Haar representation [9.8], wavelet analysis of DNA sequences using integer representation [9.9] etc. can be applied to the analyses of sequences in virology context

- More ways of numerical sequence construction using nucleotide and/or amino-acid related physic-chemical parameters can be attempted (for analysis using Fourier transform etc.). For example, relative mutability of amino acids can be a new candidate for the said analysis

9.10 Closure

This study is mainly concerned with bioinformatic analyses exclusively on the genetic features of viruses. Though not addressed, corresponding proteomic efforts may also yield fortifying results. Since this research is of debut effort, more research needs to be undertaken on the related issues. Nevertheless, the studies addressed here can be considered as seed efforts for futuristic research.

CHAPTER X

EXECUTIVE SUMMARY

Avenues of using traditional bioinformatic algorithm to explore their applications exclusively for genomic sequences of viruses form the thematic scope of this research. Hence, known concepts of entropy methods, energetic concepts and Fourier transform techniques are invoked and applied to a set of test viral genome sequences. Considered thereof, are ssDNA sequence pertinent to Parvovirus B19V and ssRNA sequences of dengue viral serovar. The motivating aspect of this study is concerned with eventual identification of structural features in such sequences that correspond to viable epitopes; and, these epitopes could be considered as promising signatures for vaccine designs.

10.1 An Overview on the Research

In Chapter I, relevant to the topic of the research, namely ‘Bioinformatic Analysis of Viral Genomic Sequences: Application to Rational Vaccine Design’, introductory details on the analysis of viral genomic sequences are presented along with the scope and objectives motivated essentially by considerations on elucidating subtle features in test sequences of certain viruses and their serotypes. Hence, indicated thereof is the cohesive use of bioinformatic approaches involving entropy, energetics and spectral-domain methods to determine the underlying common and/or differential features across the test viral sequences; and, the health-science related objective is posed to identify the salient signatures that can be seen in the viral sequences, which can be useful in synthesizing appropriate vaccines.

Commensurate with the scope and motivations of the research outlined above, the specific bioinformatic exercises carried out are as follows:

- A comprehensive survey on genomic sequence details pertinent to viral species
- An outline description of the sequence features of the test entities
- Formulating and applying entropy-based statistical divergence techniques for test sequence analyses
- Identifying unique features such as bends, bulges. loops etc. in a test sequence using the entropy concept
- Specifying the residues of the test sequences in terms of numerical characteristics such as *nearest-neighbor interaction potential (NN-E)* and analyzing the new format of sequences in terms of the NN-E profile
- Representing the residues in terms of *electron-ion interaction potential* (EIIP) values and analyzing the resulting numerical sequence in the Fourier transform domain so as to elucidate the underlying characteristics
- Combining all the aforesaid methods (based on entropy, energetics and Fourier transform) and cohesively obtain common signature details of the test sequences. For example, such a cohesive analysis leads to finding the most probable common

attributes among the diverse serovar sequences of dengue virus. The common features so deduced are applied to data-mining of the associated protein information useful in conceiving epitope format for vaccine design. The methodology used to combine the results of three different methods as above relies on the classical technique of logistic regression. However, more considerations are emphasized in specifying an upper bound and lower bound on the logistically regressed data using Langevi-Bernoulli expression as the logit function

- Lastly, a review on rational vaccine design is presented consistent with the general theme and results of the present theme

The outcome of this research study is deliberated in the following publications referenced as [10.1 – 10.3]:

- Fuzzy Splicing in Precursor-mRNA Sequences: Prediction of Aberrant Splice-junctions in Viral DNA Context, *Journal of Biomedical Science and Engineering (JBiSE)*, vol. 4, no. 4, April 2011

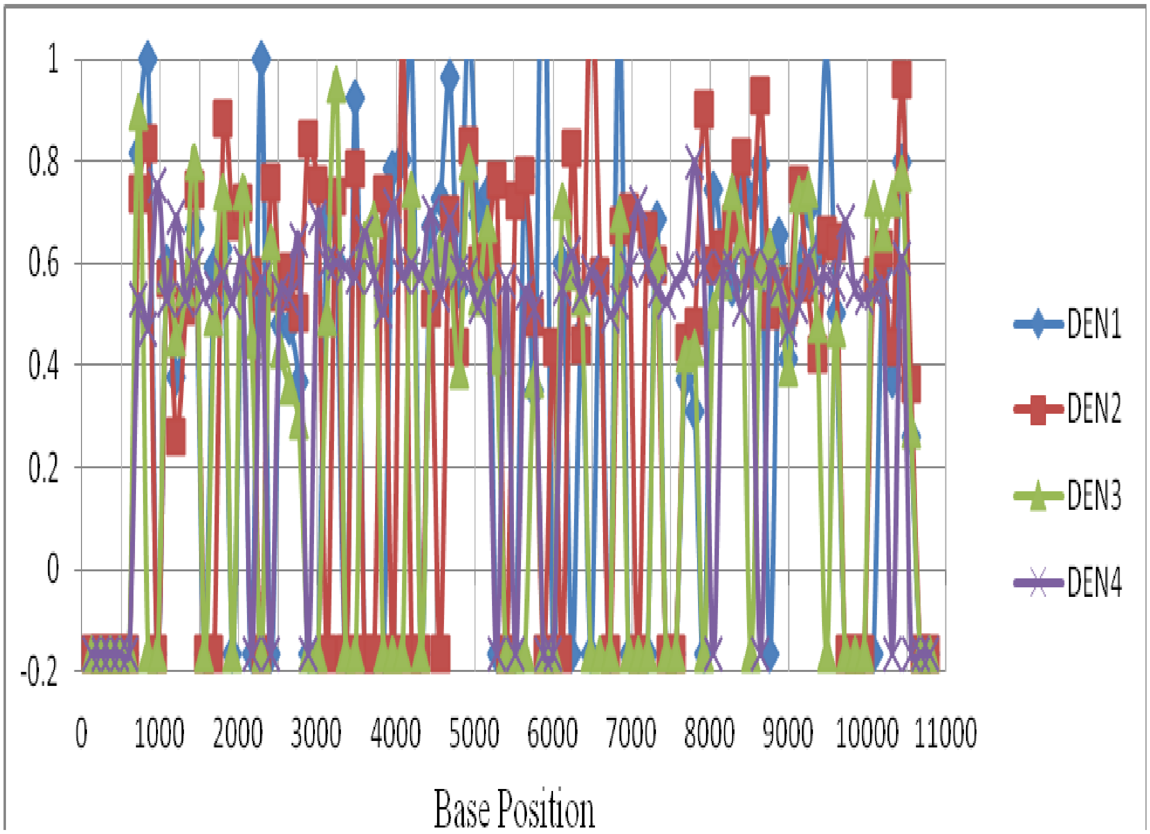
- Information-theoretic Algorithms in Bioinformatics and Bio- /Medical-imaging: A Review, *Proceedings of the IEEE International Conference on Recent Trends in Information Technology (IEEE ICRTIT)*, Chennai, India, pp. 183-188, June 2011

- Computation of Entropy and Energetics Profiles of a Single-stranded Viral DNA, *International Journal of Bioinformatics Research and Applications (IJBRA)*, vol. 7, no. 3, pp. 239-261, August 2011

- A Cohesive Analysis of DNA/RNA Sequences *via* Entropy, Energetics and Spectral-domain Methods to Assess Genomic Features Across Single Viral Diversity, *International Journal of Bioinformatics Research and Applications (IJBRA)*, (Under Review)

The snippets of the code for the entire dissertation are available at:
https://sites.google.com/site/bioinformaticanalysis/

APPENDIX


1
0.8
0.6
0.4
0.2
0
-0.2
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
Base Position
DEN1
DEN2
DEN3
DEN4

REFERENCES

[1.1] "Human Genome Project Information", [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[1.2] "GenBank Overview", [Online], (Accessed on Sep 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/genbank/

[1.3] "Entrez, The Life Sciences Search Engine", [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/sites/gquery

[1.4] "ExPASy Bioinformatics Resource Portal", [Online], (Accessed on Sept. 15, 2012), Available at: http://expasy.org/

[1.5] H. M. Berman, A. Gelbin, L. Clowney, S. Hsieh, C. Zardecki and J. Westbrook, " The Nucleic Acid Database: Present and Future", *Journal of Research of the National Institutes of Standards and Technology*, vol. 101, no. 3, pp. 243-257, May-June 1996

[1.6] "DBD, Database of Biological Database", [Online], (Accessed on Sept. 15, 2012), Available at: http://www.biodbs.info/BiologicalDatabase.html

[1.7] A. M. Campbell and L. J. Heyer, *Discovering Genomics, Proteomics and Bioinformatics*, Pearson-Benjamin Cummings, San Francisco, CA: 2002

[1.8] J. T. Edsall and H. Gutfreund, *Biothermodynamics: The Study of Biochemical Processes at Equilibrium.* John Wiley & Sons, Inc., West Sussex, UK: 1983

[1.9] M. L. Johnson, J. Holtz and G. K. Ackers, *Biothermodynamics, Part 1,* Academic Press, New York, NY:2009 [Available online as e-book]

[1.10] M. L. Johnson, J. Holtz and G. K. Ackers, *Biothermodynamics, Part 2*, Academic Press, New York, NY:2009

[1.11] T. V. Arredondo, P.S. Neelakanta and D. DeGroff, "Fuzzy Attributes of a DNA: A Fuzzy Inference Engine for Codon -"Junk" Codon Delineation", *Artificial Intelligence in Medicine*, vol. 35, pp. 87-105, Oct. 2005

[1.12] K. Deergha Rao and M.N.S. Swamy, "Analysis of Genomics and Proteomics Using DSP Techniques", *IEEE Transactions on Circuits and Systems I (regular papers)*, vol. 55, no. 1, pp. 370-378, Feb 2008

[1.13] D. Anastassiou, "Genomic Signal Processing", *IEEE Signal Processing*, vol. 18, no. 4, pp. 8-20, 2001

[1.14] D. Anastassiou, H. Liu and V. Varadan, "Variable Window Binding for Mutually Exclusive Alternative Splicing", *Genome Biology*, vol. 7, pp., Jan 2006

[1.15] P. S. Neelakanta, T. V. Arredondo and D. DeGroff, "Redundancy Attributes of a Complex System: Application to Bioinformatics", *Complex Systems*, vol. 14, pp. 215-233, 2003

[1.16] L. Florea, "Bioinformatics of Alternative Splicing and its Regulation", *Briefing in Bioinformatics*, vol. 7, no. 1, pp. 55-69, 2006

[1.17] Perambur S. Neelakanta, S. Chatterjee, M. Pavlovic, A. Pandya and D. DeGroff, "Fuzzy Splicing in Precursor-mRNA Sequences: Prediction of Aberrant Splice-junctions in Viral DNA Context", *Journal of Biomedical Science and Engineering*, vol. 4, no. 4, pp. 272-281, 2011

[1.18] P. S. Neelakanta, S. Chatterjee and G. A. Thengum-Pallil, "Computation of Entropy and Energetics Profiles of a Single-stranded Viral DNA", *International Journal of Bioinformatics Research and Applications*, vol. 7, no. 3, pp. 239-261, 2011

[1.19] "GenBank, Dengue virus type 1: Complete genome. NCBS reference Sequence NC_001477.1.", [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/NC_001477

[1.20] "Gen Bank, Dengue virus type 2, complete genome NCBI Reference Sequence: NC_001474.2", [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/NC_001474

[1.21] "GenBank, Dengue virus type 3, complete genome NCBI Reference Sequence: NC_001475.2", [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/NC_001475

[1.22] "GenBank, Dengue virus type 4, complete genome NCBI Reference Sequence: NC_002640.1", [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/NC_002640

[1.23] A. A. T. Bui and R. K. Taira (eds.) '*Medical Imaging Informatics*', Chapter 10, Springer-Verlag LLC, New York, NY: 2010

[1.24] M. Pavlovic, M. Cavallo, A. Kats, A. Kotlarchyk, H. Zhuang, Y. Shoenfels, "From Pauling's Abzyme Concept to the New era of Hydrolytic Anti-DNA Autoantobodies: A link to rational vaccine design? – A review", *International Journal of Bioinformatics Research and Applications*, vol. 7, no. 3, pp. 220-238, 2011

[1.25] Information-theoretic Algorithms in Bioinformatics and Bio-/Medical-imaging: A Review, *Proceedings of the IEEE International Conference on Recent Trends in Information Technology (IEEE ICRTIT)*, pp. 183-188, 2011

[2.1] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK: 1998

[2.2] P. A. Pevzner, Computational Molecular Biology: An Algorithmic Approach, The MIT Press, Cambridge, MA: 2002

[2.3] C. E. Shannon and W. W. Weaver, The Mathematical Theory of Communication, Urbana, USA, University of Illinois Press, 1949

[2.4] T. V. Arredondo, P.S. Neelakanta and D. DeGroff, “Fuzzy Attributes of a DNA Complex: Development of a Fuzzy Inference Engine for Codon -“Junk” Codon Delineation, ” Artificial Intelligence in Medicine, vol. 35, pp. 87-105, October 2005

[2.5] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán and H. E. Stanley, “Finding Borders Between Coding and Noncoding DNA Regions by Entropic Segmentation Method”, Physical Review Letters, vol.85, pp.1342-1345, 2000

[2.6] P. S. Neelakanta, T. V. Arredondo and D. DeGroff, “Redundancy Attributes of a Complex System: Application to Bioinformatics,” Complex Systems, vol. 14, pp. 215-233, 2003

[2.7] P. S. Neelakanta, S. Pandya and T. V. Arredondo, “Binary Representation of DNA Sequences Towards Developing Useful Algorithms in Bioinformatics”, The 7th World Multi Conference on Systemics, Cybernetics and Informatics (SCI 2003), (July 27-30, 2003, Orlando, FL, USA), Vol. VIII, 195-197, 2003

[2.8] P. S. Neelakanta, Information-Theoretic Aspects of Neural Networks, CRC-Press, Boca Raton, FL: 1999

[2.9] P. S. Neelakanta, S. Chatterjee, M. Pavlovic, A. Pandya and D. DeGroff, “Fuzzy Splicing in Precursor-mRNA Sequences: Prediction of Aberrant Splice-junctions in Viral DNA Context”, Journal of Biomedical Science and Engineering, vol. 4, no. 4, pp. 272-281, 2011

[2.10] X. Tianbing, J. McDowell and D. H. Turner, “Thermodynamics of Nonsymmetric Tandem Mismatches Adjacent to G & C Base Pairs in RNA”, Biochemistry, vol. 36, pp. 12486-12497, 1997

[2.11] X. Tianbing, J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, J. Xiaoqi, C. Cox and D. H. Turner, “Thermodynamic Parameters for an Expanded Nearest-neighbor Model for Formation of RNA Duplexes with Watson-Crick base pairs”, Biochemistry, vol. 37, pp. 14719-14735, 1998

[2.12] T. Xia, J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox and D. H. Turner, “Thermodynamic Parameters for an Expanded Nearest-neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs”, Biochemistry, vol. 37, no.42, pp. 4719-4735, 1998

[2.13] P. S. Neelakanta, S. Chatterjee and G. A. Thengum-Pallil, “Computation of Entropy and Energetics Profiles of a Single-stranded Viral DNA”, International Journal of Bioinformatics Research and Applications, vol. 7, no. 3, pp. 239-261, August 2011

[2.14] M. Pavlovic, M. Cavallo, A. Kats, A. Kotlarchyk, H. Zhuang and Y. Shoenfeld, “From Pauling’s Abyzyme Concept to the New Era of Hydrolytic Anti-DNA Autoantibodies: A Link to Rational Vaccine Design? – A Review”, International Journal of Bioinformatics Research and Applications, vol. 7, no. 3, pp. 220-238, August 2011

[2.15] D. C. Benson, “Digital Signal Processing Methods for Biosequence Comparison”, Nucleic Acids Research, vol. 18, no. 10, pp. 3001-3006, May 1990

[2.16] D. C. Benson, “Fourier Methods for Biosequence Analysis”, Nucleic Acids Research, vol. 18, no. 21, pp. 6305-6310, November 1990

[2.17] E. A. Cheever, G. C. Overton and D. B. Searls, “Fast Fourier Transform-based Correlation of DNA Sequences using Complex Plane Encoding”, Computer Applications in the Biosciences, vol. 7, no. 2, pp. 143-154, 1991

[2.18] Y. Zhou, L. Zhou, Z. Yu and V. Anh, “Distinguish Coding and Noncoding Sequences in a Complete Genome Using Fourier Transform”, Third International Conference on Natural Computation (ICNC 2007), pp. 295-299, 2007

[2.19] D. Anastassiou, “Genetic Data Processing”, IEEE Signal Processing, vol. 18, no. 4, pp. 8-20, July 2001

[2.20] D. Rao and M. N. S. Swamy, “Analysis of Genomics and Proteomics Using DSP Techniques”, IEEE Transactions on Circuits and Systems I, vol. 55, no. 1, pp. 370-378, February 2008

[3.1] P. Forterre, “Defining Life: The Virus Viewpoint”, *Origins of Life and Evolution of the Biosphere*, vol. 40, no. 2, pp. 151-160, April 2010

[3.2] H. R. Gelderblom, “Structure and Classification of Viruses”, In: Baron, S., ed., Chapter 41, 4th edition, *Medical Microbiology,* University of Texas Medical Branch at Galveston, Galveston, TX: 1996

[3.3] C. P. McKay, “What Is Life-and How Do We Search for It in Other Worlds?”, *Public Library of Science Biology,* vol.2, no. 9, e203, September 2004

[3.4] K. H. Nealson, P. G. Conrad, “Life: Past, Present and Future”, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences,* vol. 354, pp.1923–1939, December 1999

[3.5] C. I. Bândea, “A New Theory on the Origin and the Nature of Viruses”, *Journal of Theoretical Biology,* vol. 105, no. 4, pp. 591-602, December 1983

[3.6] B. Roizman, “Multiplication”, In: Baron S, ed., *Medical Microbiology,* 4th edition, Chapter 42, University of Texas Medical Branch at Galveston, Galveston, TX: 1996

[3.7] F. Holmes, “Problems of Viral Nomenclature and Classification”, *Annals of the New York Academy of Sciences*, vol. 56, pp. 414-421, March 1953

[3.8] A. Lwoff, R. Horne and P. Tournier, “A System of Viruses”, *Cold Spring Harbor Symposia on Quantitative Biology,* vol. 27, pp. 51-55, 1962

[3.9] “International Committee on Taxonomy of Viruses”, [Online], (Accessed on Sept. 15, 2012), Available at: http://ictvonline.org/virusTaxonomy.asp?version=2009

[3.10] D. Baltimore, “Expression Of Animal Virus Genomes”, *Bacteriological Reviews*, vol. 35, no. 3, pp. 235-241, September 1971

[3.11] G. Sumbali and R. Mehrotra, *Principles of Microbiology*, Tata-McGraw Hill Education Pvt. Limited, 7 West Patel Nagar, New Delhi, India, pp. 158, 2009

[3.12] J. T. Patton, *Segmented Double-stranded RNA Viruses: Structure and Molecular Biology*, Caister Academic Press, Norflok, UK: January 2008

[3.13] J. R. Stephenson, "Rational Design of Vaccines Against Enveloped RNA Viruses", *Vaccine*, vol. 3, no. 1, pp. 69-72, 1985

[3.14] J. Holland, K. Spindler, F. Horodyski, E. Grabau, S. Nichol and S. VandePol, "Rapid Evolution of RNA Genomes", *Science*, vol. 215, March 1982

[3.15] D. White and F. Fenner, *Medical Virology*, Academic Press, San Diego, CA: 1994

[3.16] H. M. Blau and M. Springer, "Gene Therapy - A Novel Form of Drug Delivery", *New England Journal of Medicine*, vol. 333, pp. 1204-1207, 1995

[3.17] R. Kurth and N. Bannert, (eds.), *Retroviruses: Molecular Biology, Genomics and Pathogenesis*, Caister Academic Press, Norfolk, UK: 2010

[3.18] C. Seeger and W. S. Mason, (eds.), *Hepadnaviruses: Molecular Biology and Pathogenesis*, Springer-Verlag, Berlin, Germany: 1991

[3.19] J. Summers and W. S. Mason, "Replication of the Genome of a Hepatitis B-like Virus by Reverse Transcription of an RNA Intermediate", *Cell*, vol. 29, pp. 403-415, 1982

[3.20] J. Sohn, S. Litwin and C. Seeger, "Mechanism for CCC DNA Synthesis in Hepadnaviruses", *PLoS ONE*, vol. 4, no. 11, e8093 (6 pages), November 2008

[3.21] C. W. Hilbers, H. A. Heus, M. J. Van Dongen and S. S. Wijmenga, "The Hairpin Elements of Nucleic Acid Structure: DNA and RNA Folding", In: F. Eckstein and D. M. J. Lilley, (eds.), *Nucleic Acids and Molecular Biology*, Springer-Verlag, Berlin, Germany, pp. 56-104, 1994

[3.22] K. C. Chen, J. J. Tyson, M. Lederman, E. R. Stout, and R. C. Bates, "A Kinetic Hairpin Transfer Model for Parvoviral DNA Replication", *Journal of Molecular Biology*, vol. 208, no.2, pp. 283-296, 1989

[3.23] E. Costello, R. Sahli, H. Bernhard and P. Beard, “The Mismatched Nucleotides in the 59-Terminal Hairpin of Minute Virus of Mice Are Required for Efficient Viral DNA Replication”, *Journal of Virology,* vol. 69, no. 12, pp. 7489-7496, 1995

[3.24] K. Chin, F. Chen and S. Chou, “Solution Structure of the ActD–5prime-CCGTT3GTGG-3prime Complex: Drug Interaction with Tandem G•T Mismatches and Hairpin Loop Backbone”, *Nucleic Acids Research*, vol. 31, no. 10, pp. 2622-2629, 2003

[3.25] “Human parvovirus B19, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/356457872

[3.26] “Dengue virus 1, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/9626685

[3.27] “Dengue virus 2, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/158976983

[3.28] “Dengue virus 3, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/163644368

[3.29] “Dengue virus 4, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/12084822

[3.30] A. M. Khan, A. T. Heiny, K. Lee, K. N. Srinivasan, T. Tan, J. August and B. Vladimir, “Large-scale Analysis of Antigenic Diversity of T-cell Epitopes in Dengue Virus”, *BMC Bioinformatics,* vol. 7, no. 5, S4, 2006

[3.31] D. J. Gubler, “Dengue”, In: T. P. Monath, *The Arboviruses: Epidemiology and Ecology. Volume II,* CRC Press, Inc., Boca Raton, FL, pp. 223-260, 1988

[3.32] D. E. Alvarez, M. F. Lodeiro, S. J. Luduena, L. I. Pietrasanta and A. V. Gamarnik, “Long-range RNA-RNA Interactions Circularize the Dengue Virus Genome”, *Journal of Virology*, vol. 79, no. 11, pp. 6631-6643, 2005

[3.33] K. Clyde and E. Harris, "RNA Secondary Structure in the Coding Region of Dengue Virus Type 2 Directs Translation Start Codon Selection and is Required for Viral Replication", *Journal of Virology*, vol. 80, no.5, pp. 2170-2182, 2006

[3.34] N. G. Iglesias, and A. Gamarnik, "Dynamic RNA Structures in the Dengue Virus Genome", *RNA Biology*, vol. 8, no. 2, pp. 249-257, March/April 2011

[3.35] R. Curtiss, III, "The Impact of Vaccines and Vaccinations: Challenges and Opportunities for Modelers", *Mathematical Biosciences And Engineering*, vol. 8, no. 1, pp. 77-93, January 2011

[3.36] A. R. Gould, "Virus Evolution: Disease Emergence and Spread", *Australian Journal of Experimental Agriculture*, vol. 44, no. 11, pp. 1085-1094

[3.37] J. A. Mumford, "Vaccines and Viral Antigenic Diversity", *Scientific and Technical Review of the Office International des Epizooties,* (*Rev. sci. tech. Off. int. Epiz.*), vol. 26, no. 1, pp. 69-90, 2007

[4.1] P. S. Neelakanta, *Information-Theoretic Aspects of Neural Networks*, CRC-Press, Boca Raton, FL: 1999

[4.2] J. N. Kapur, "Measures of Uncertainty in Mathematical Programming and Physics", Journal of the Indian Society of Agricultural Statistics, vol. 24, pp. 47-66, 1972

[4.3] C. E. Shannon, "Transmission of Information", The Bell System Technical Journal, vol. 27, pp. 379- 423: 1948

[4.4] T. V. Arredondo, P.S. Neelakanta and D. DeGroff, "Fuzzy Attributes of a DNA a Fuzzy Inference Engine for Codon -"Junk" Codon Delineation", Artificial Intelligence in Medicine, vol. 35, pp. 87-105, Oct. 2005

[4.5] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán and H. E. Stanley, "Finding Borders Between Coding and Noncoding DNA Regions by Entropic Segmentation Method", Physical Review Letters, vol.85, pp.1342-1345, 2000

[4.6] G. N. Alekseev, Energy and Entropy, Moscow, Mir Publishers, USSR: 1986

[4.7] R. A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems”, Annals of Eugenics, vol. 7, pp. 179-188, 1936

[4.8] C. E. Shannon and W. W. Weaver, The Mathematical Theory of Communication, Urbana, USA, University of Illinois Press: 1949

[4.9] P. S. Neelakanta, T. V. Arredondo and D. De Groff, “Redundancy Attributes of a Complex System: Application to Bioinformatics”, Complex Systems, vol. 14, pp.215-233, 2003

[4.10] R. Román-Roldán, P. Bernaola-Galván and J. L. Oliver, “Application of Information Theory to DNA Sequence Analysis: A Review,” Pattern Recognition, vol. 29, pp.1187-1194, 1996

[4.11] M. B. Shapiro and P. Senapathy, “RNA Splice Junctions of Different Classes of Eukaryotes: Sequence Statistics and Functional Implications in Gene Expression”, Nucleic Acid Research, vol.15, no. 17, pp. 7155-7174, September 1987

[4.12] M. Krawczak, J. Reiss and D. N. Cooper, “The Mutational Spectrum of Single Base-pair Substitutions in mRNA Splice Junctions of Human Genes: Causes and Consequences”, Human Genetics, vol. 90, no. 1-2, pp. 41-54, October 1992

[4.13] J.-S. R. Jang, C. -T. Sun and E. Mizutani, Neuro-fuzzy and Soft Computing, Prentice Hall, New Jersey, NJ: 1997

[4.14] P. S. Neelakanta, S. T. Abusalah, D. De Groff, and J. C. Park, “Fuzzy Nonlinear Activity and Dynamics of Fuzzy Uncertainty in the Neural Complex”, Neurocomputing, vol. 20, pp. 123-153, 1998

[4.15] P. S. Neelakanta, J. C. Park and D. Degroff, “Complexity Parameter vis-à-vis Interaction Systems: Application to Neurocybernetics”, Cybernetica, vol. XL, no.4, pp.243-253, 1997

[4.16] P. S. Neelakanta and D. De Groff, Neural Network Modeling: Statistical Mechanics and Cybernetic Perspectives, CRC Press, Boca Raton, FL: 1994

[4.17] P. S. Neelakanta and W. Deecharoenkul, “A Complex System Characterization of Modern Telecommunication Services”, Complex Systems, vol. 12, pp. 31-69, 2000

[4.18] M. Pavlovic, M. Cavallo, A. Kats, A. Kotlarchyk, H. Zhuang, Y. Shoenfels, “From Pauling’s Abzyme Concept to the New Era of Hydrolytic Anti-DNA Autoantobodies: A link to Rational Vaccine Design? – A review”, International Journal of Bioinformatics Research and Applications, vol. 7, no. 3, pp. 220-238, 2011

[4.19] A. Krishnamachari, V. M. Mandal and Karmeshu, “Study of Binding Sites Using Renyi Parametric Entropy Measure”, Journal of Theoretical Biology, vol. 227, pp. 429-436, 2004

[4.20] L. Florea, “Bioinformatics of Alternative Splicing and its Regulation”, Briefing in Bioinformatics, vol. 7, no. 1, pp. 55-69, 2006

[4.21] R. M. Stephens and T. D. Schneider, “Features of Spliceosome Evolution and Function Inferred from an Analysis of the Information at Human Splice Sites”, Journal of Molecular Biology, vol. 228, pp. 1124-1136, 1992

[4.22] Gardiner-Garden and M. Frommer, “CpG Islands in Vertebrate Genomes”, Journal of Molecular Biology, vol. 196, no.2, pp. 261-282, 1987

[4.23] D. Takai and P. A. Jones, “Comprehensive Analysis of CpG Islands in Human Chromosomes 21 and 22”, Proceedings of National Academy of Science (USA), vol. 99, no.6, pp. 3740-3745, 2002

[4.24] S. Karlin, W. Doerfler and L. R. Cardon, “Why is CpG Suppressed in the Genomes of Virtually All Small Eukaryotic Viruses But Not in Those of Large Eukaryotic Viruses?”, Journal of Virology, vol. 68, no.5, pp. 2889-2897, 1994

[5.1] H. D. Nguyen and C. L. Brooks, III, “Generalized Structural Polymorphism in Self-assembled Viral Particles”, *Nano Letters,* vol. 8, no.12, pp. 4574–4581, 2008

[5.2] E. Costello, R. Sahli, H. Bernhard and P. Beard, “The Mismatched Nucleotides in the 5-terminal Hairpin of Minute Virus of Mice are Required for Efficient Viral DNA Replication”, *Journal of Virology*, vol. 69, no.12, pp. 7489–7496, 1995

[5.3] K. Chin, F. Chen and S. Chou, “Solution Structure of the ActD–5prime-CCGTT3GTGG-3prime Complex: Drug Interaction with Tandem G·T

Mismatches and Hairpin Loop Backbone", *Nucleic Acids Research*, vol. 31, no. 10, pp. 2622-2629, 2003

[5.4] C. W. Hilbers, H. A. Heus, M. J. van Dongen and S. S. Wijmenga, "The Hairpin Elements of Nucleic Acid Structure: DNA and RNA Folding". In F. Eckstein, F. and D. M. J. Lilley, (eds), *Nucleic Acids and Molecular Biology*, pp.56-104, Springer-Verlag, Berlin, Germany: 1994

[5.5] V. A. Belyi and M. Muthukuma, "Electrostatic Origin of the Genome Packing in Viruses", *Proceedings of the National Academy of Sciences*, vol. 103, no. 46, pp. 17174-17178, 2006

[5.6] C. L. Ting, J. Wu and Z.-G. Wang, "Thermodynamic Basis for the Genome to Capsid Charge Relationship in Viral Encapsidation", *Proceedings of the National Academy of Sciences*, vol. 108, no. 41, pp. 16986-16991, 2011

[5.7] V. P. Antao and I. Tinoco, Jr., "Thermodynamic Parameters for Loop Formation in RNA and DNA Hairpin Tetraloops", *Nucleic Acids Research*, vol. 20, no.4, pp. 819-824, 1992

[5.8] P. Guillaume, H. Santini, C. Pakleza and J. A. H. Cognet, "DNA tri- and tetra-Loops and RNA Tetra-loops Hairpins Fold as Elastic Biopolymer Chains in Agreement with PDB Coordinates", *Nucleic Acids Research*, vol. 31, no. 3, pp. 1086-1096, 2003

[5.9] E. M. Moody, J. C. Feerrar and P. C. Bevilacqua, "Evidence that Folding of an RNA Tetraloop Hairpin is Less Cooperative than Its DNA Counterpart", *Biochemistry*, vol. 43, no. 25, pp. 7992–7998, 2004

[5.10] P. C. Bevilacqua and J. M. Blose, "Structures, Kinetics, Thermodynamics, and Biological Functions of RNA Hairpins", *Annual Review of Physical Chemistry*, vol. 59, pp. 79-103, 2008

[5.11] T. Xia, J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox and D. H. Turner, "Thermodynamic Parameters for an Expanded Nearest-neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs", *Biochemistry*, vol. 37, no.42.1, pp. 4719-4735, 1998

[5.12] A. D. Baxevanis and B. F. Oullette, *Bioinformatics: A Practical Guide to the Analyses of Genes and Proteins*, pp. 146-147, John Wiley and Sons, Hoboken, NJ: 2005

[5.13] D. H. Mathews, J. Sabina, M. Zuker and D. H. Turner, "Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure", *Journal of Molecular Biology*, vol. 288, pp. 911-940, 1999

[5.14] X. Tianbing, J. A. McDowell and D. H. Turner, "Thermodynamics of Nonsymmetric Tandem Mismatches Adjacent to G & C Base Pairs in RNA", *Biochemistry*, vol. 36, pp. 12486-12497, 1997

[5.15] X. Tianbing, J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, J. Xiaoqi, C. Cox and D. H. Turner, "Thermodynamic Parameters for an Expanded Nearest-neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs", *Biochemistry***,** vol. 37, pp. 14719-14735, 1998

[5.16] S. M. Ali and S. M. Silvey, "A General Class of Coefficients of Divergence of One Distribution from Another", *Journal of the Royal Statistical Society*, *Series B* (*Methodological*), vol. 28, no. 1, pp. 131-142, 1966

[5.17] P. S. Neelakanta, T. V. Arredondo and D. De Groff, D., "Redundancy Attributes of a Complex System: Application to Bioinformatics", *Complex Systems*, vol. 14, pp. 215-233, 2003

[5.18] "Human parvovirus B19, complete genome" [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/356457872

[5.19] T. V. Arredondo, P. S. Neelakanta and D. DeGroff, "Fuzzy Attributes of a DNA Complex: Development of a Fuzzy Inference Engine for Codon -''Junk'' Codon Delineation", *Artificial Intelligence in Medicine*, vol. 35, pp. 87-105, 2005

[5.20] H. D. Nguyen and C. L. Brooks, III, "Generalized Structural Polymorphism in Self-assembled Viral Particles", *Nano Letters*, vol. 8, no.12, pp. 4574–4581, 2008

[5.21] E. Costello, R. Sahli, H. Bernhard and P. Beard, “The Mismatched Nucleotides in the 5-terminal Hairpin of Minute Virus of Mice are Required for Efficient Viral DNA Replication”, *Journal of Virology*, vol. 69, no.12, pp. 7489–7496, 1995

[5.22] K. Chin, F. Chen and S. Chou, “Solution Structure of the ActD–5prime-CCGTT3GTGG-3prime Complex: Drug Interaction with Tandem G·T Mismatches and Hairpin Loop Backbone”, *Nucleic Acids Research*, vol. 31, no. 10, pp. 2622-2629, 2003

[5.23] C. W. Hilbers, H. A. Heus, M. J. van Dongen and S. S. Wijmenga, “The Hairpin Elements of Nucleic Acid Structure: DNA and RNA Folding”. In F. Eckstein, F. and D. M. J. Lilley, (eds), *Nucleic Acids and Molecular Biology*, pp.56-104, Springer-Verlag, Berlin, Germany: 1994

[5.24] V. A. Belyi and M. Muthukuma, “Electrostatic Origin of the Genome Packing in Viruses”, *Proceedings of the National Academy of Sciences*, vol. 103, no. 46, pp. 17174-17178, 2006

[5.25] C. L. Ting, J. Wu and Z.-G. Wang, “Thermodynamic Basis for the Genome to Capsid Charge Relationship in Viral Encapsidation”, *Proceedings of the National Academy of Sciences*, vol. 108, no. 41, pp. 16986-16991, 2011

[5.26] V. P. Antao and I. Tinoco, Jr., “Thermodynamic Parameters for Loop Formation in RNA and DNA Hairpin Tetraloops”, *Nucleic Acids Research*, vol. 20, no.4, pp. 819-824, 1992

[5.27] P. Guillaume, H. Santini, C. Pakleza and J. A. H. Cognet, “DNA tri- and tetra-Loops and RNA Tetra-loops Hairpins Fold as Elastic Biopolymer Chains in Agreement with PDB Coordinates”, *Nucleic Acids Research*, vol. 31, no. 3, pp. 1086-1096, 2003

[5.28] E. M. Moody, J. C. Feerrar and P. C. Bevilacqua, “Evidence that Folding of an RNA Tetraloop Hairpin is Less Cooperative than Its DNA Counterpart”, *Biochemistry*, vol. 43, no. 25, pp. 7992–7998, 2004

[5.29] P. C. Bevilacqua and J. M. Blose, “Structures, Kinetics, Thermodynamics, and Biological Functions of RNA Hairpins”, *Annual Review of Physical Chemistry*, vol. 59, pp. 79-103, 2008

[5.30] T. Xia, J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox and D. H. Turner, “Thermodynamic Parameters for an Expanded Nearest-neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs”, *Biochemistry*, vol. 37, no.42.1, pp. 4719-4735, 1998

[5.31] A. D. Baxevanis and B. F. Oullette, *Bioinformatics: A Practical Guide to the Analyses of Genes and Proteins*, pp. 146-147, John Wiley and Sons, Hoboken, NJ: 2005

[5.32] D. H. Mathews, J. Sabina, M. Zuker and D. H. Turner, “Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure”, *Journal of Molecular Biology*, vol. 288, pp. 911-940, 1999

[5.33] X. Tianbing, J. A. McDowell and D. H. Turner, “Thermodynamics of Nonsymmetric Tandem Mismatches Adjacent to G & C Base Pairs in RNA”, *Biochemistry*, vol. 36, pp. 12486-12497, 1997

[5.34] X. Tianbing, J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, J. Xiaoqi, C. Cox and D. H. Turner, “Thermodynamic Parameters for an Expanded Nearest-neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs”, *Biochemistry***,** vol. 37, pp. 14719-14735, 1998

[5.35] S. M. Ali and S. M. Silvey, “A General Class of Coefficients of Divergence of One Distribution from Another”, *Journal of the Royal Statistical Society, Series B* (*Methodological*), vol. 28, no. 1, pp. 131-142, 1966

[5.36] P. S. Neelakanta, T. V. Arredondo and D. De Groff, D., “Redundancy Attributes of a Complex System: Application to Bioinformatics”, *Complex Systems*, vol. 14, pp. 215-233, 2003

[5.37] “Human parvovirus B19, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/356457872

[5.38] T. V. Arredondo, P. S. Neelakanta and D. DeGroff, “Fuzzy Attributes of a DNA Complex: Development of a Fuzzy Inference Engine for Codon -‘‘Junk’’ Codon Delineation”, *Artificial Intelligence in Medicine,* vol. 35, pp. 87-105, 2005

[6.1] H. C. G. Leitão, L. S. Pessôa, and J. Stolfi, “Mutual Information Content of Homologous DNA Sequences”, *Genetics and Molecular Research,* vol. 4, no. 3, pp. 553-562, 2005

[6.2] “Dengue virus 1, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/9626685

[6.3] “Dengue virus 2, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/158976983

[6.4] “Dengue virus 3, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/163644368

[6.5] “Dengue virus 4, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/12084822

[6.6] D. C. Benson, “Digital Signal Processing Methods for Biosequence Comparison”, *Nucleic Acids Research,* vol. 18, no. 10, pp. 3001-3006, May 25, 1990

[6.7] D. C. Benson, “Fourier Methods for Biosequence Analysis”, *Nucleic Acids Research,* vol. 18, no. 21, pp. 6305-6310, Nov 21, 1990

[6.8] E. A. Cheever, G. C. Overton, and D. B. Searls, “Fast Fourier Transform-based Correlation of DNA Sequences Using Complex Plane Encoding”, *Computer Applications in the Biosciences,* vol. 7, no. 2, pp. 143-154, 1991

[6.9] Y. Zhou, L. Zhou, Z. Yu, and V. Anh, “Distinguish Coding and Noncoding Sequences in a Complete Genome Using Fourier Transform”, *Third International Conference on Natural Computation (ICNC 2007),* pp. 295-299, 2007

[6.10] D. Anastassiou, “Genetic Data Processing”, *IEEE Signal Processing,* vol.18, no. 4, pp. 8-20, July 2001

[6.11] D. Anastassiou, H. Liu and V. Varadan, “Variable Window Binding for Mutually Exclusive Alternative Splicing”, *Genome Biology,* vol. 7, pp., Jan 2006

[6.12] K. Deergha Rao and M.N.S. Swamy, “Analysis of Genomics and Proteomics Using DSP Techniques”, *IEEE Transactions on Circuits and Systems I (regular papers),* vol. 55, no. 1, pp. 370-378, Feb 2008

[6.13] C. Jeng, I. Yang, and K. Hsieh, “Bacteria Classification on Power Spectrums of Complete DNA Sequences by Self-organizing Map”, *Neural Information Processing,* vol.9, no. 3, Letters and Reviews, Dec 2005

[6.14] M. De Sousa Vieira, “Statistics of DNA Sequences: A Low-frequency Analysis”, *Physical Review E,* vol. 60, no. 5, pp. 5932-5937, 1999

[6.15] Y. H. Chen, S. L. Nyeo, and J. P. Yu, “Power-laws in the Complete Sequences of Human Genome”, *Journal of Biological Systems,* vol. 13, no. 2, pp. 105-115, 2005

[6.16] Y. Isohata, and M. Hayashi, “Analyses of DNA Base Sequences for Eukaryotes in Terms of Power Spectrum Method”, *Japanese Journal of Applied Physics,* vol. 44, no. 2, pp. 1143-1146, 2005

[6.17] H. Herzel and I. Grosse, “Measuring Correlations in Symbolic Sequences,” *Physica A*, vol. 216, pp. 518-542, 1995

[6.18] H. Herzel, O. Weiss, and E. N. Trifonov, “10-11 bp Periodicities in Complete Genome Reflect Protein Structure and DNA Folding”, *Bioinformatics,* vol. 15, no. 3, pp. 187-193, 1999

[6.19] W. J. Lee and L. F. Luo, “Periodicity of Base Correlation in Nucleotide Sequence”, *Physical Review E,* vol. 56, no. 1, pp. 848-851, 1997

[6.20] A. Fukushima, T. Ikemura, M. Kinouchi, T. Oshima, Y. Kudo, H. Mori and S. Kanaya, “Periodicity in Prokaryotic and Eukaryotic Genomes Identified by Power Spectrum Analysis”, *Gene,* vol. 300, no. 1-2, pp. 203-211, 2002

[6.21] S. L. Nyeo, I. C. Yang and C. H. Wu, “Spectral Classification of Archaeal and Bacterial Genomes”, *Journal of Biological Systems,* vol. 10, no. 3, pp. 233-241, 2002

[6.22] R. W. Hanson, “Fast Fourier Transform Analysis of DNA Sequences”, *BA Thesis, The Division of Mathematics and Natural Sciences, Reed College,* May 2003

[6.23] B. D. Silverman and R. Linsker, “A Measure of DNA Periodicity”, *Journal of Theoretical Biology*, vol. 118, pp. 295-300, 1986

[6.24] S. Tiwari, S. Ramachandran, S. Bhattacharya and R. Ramaswamy, “Prediction of Probable Genes by Fourier Analysis of Genomic Sequences”, *Computer Applications in the Biosciences*, vol. 113, pp. 263-270, 1997

[6.25] J. W. Fickett, “Recognition of Protein Coding Regions in DNA Sequences”, *Nucleic Acid Research*, vol. 10, pp. 5303-5318, 1982

[6.26] J. W. Fickett and C. S. Tung, “Assessment of Protein Coding Measures”, *Nucleic Acid Research*, vol. 20, pp. 6441-6450, 1992

[7.1] Wright, S., *Evolution and the Genetics of Populations: Genetics and Biometric Foundations v. 4 (Variability within and Among Natural Populations); New Edition*. University of Chicago Press, 1984

[7.2] Wright, S., *Evolution and the Genetics of Populations: Genetics and Biometric Foundations v. 3 (Experimental Results and Evolutionary Deductions); New Edition*. University of Chicago Press, 1984

[7.3] “Dengue virus type 1: Complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/NC_001477

[7.4] “Dengue virus 2, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/158976983

[7.5] “Dengue virus 3, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/163644368

[7.6] “Dengue virus 4, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/12084822

[7.7] T. V. Arredondo, P. S. Neelakanta and D. DeGroff, “Fuzzy Attributes of a DNA a Fuzzy Inference Engine for Codon -“Junk” Codon Delineation”, *Artificial Intelligence in Medicine,* vol. 35, pp. 87-105, Oct. 2005

[7.8] Tianbing, X., McDowell, J. A., and Turner, D. H., “Thermodynamics of Nonsymmetric Tandem Mismatches Adjacent to G & C Base Pairs in RNA’, *Biochemistry,* Vol. 36, pp. 12486-12497.1997

[7.9] K. Deergha Rao and M.N.S. Swamy, “Analysis of Genomics and Proteomics Using DSP Techniques”, *IEEE Transactions on Circuits and Systems I (regular papers),* vol. 55, no. 1, pp. 370-378, Feb 2008

[7.10] P. S. Neelakanta, *Information-Theoretic Aspects of Neural Networks,* CRC-Press, Boca Raton, FL: 1999

[7.11] D. G. Kleinbaum and M. Klein, *Logistic Regression: A Self Learning Text, 3rd edition,* Springer, New York, NY: 2010

[8.1] B. Bottazzi, A. Doni, C. Garlanda and A. Mantovani, “An Integrated View of Humoral Innate Immunity: Pentraxins as a Paradigm”, *Annual Review of Immunolgy*, vol. 28, pp. 157-183, 2010

[8.2] K. P. Murphy, P. Travers, M. Walport and C. Janeway, *Janeway's Immunobiology,* 7th edition, Garland Science, New York, NY: 2008

[8.3] P. Parham and C. Janeway, *The Immune System, 3rd ed.*, Garland Science, New York, NY: 2009

[8.4] S. Akira, S. Uematsu and O. Takeuchi, “Pathogen Recognition and Innate Immunity”, *Cell,* vol. 124, pp. 783-801, 2006

[8.5] D. Masopust, V. Vezys, E. J. Wherry and R. Ahmed, “A Brief History of CD8 T Cells”, *European Journal of Immunology*, vol. 37, Suppl. 1S, pp. 103-110, 2007

[8.6] J. R. Lees. and D. L. Farber, “Generation, Persistence and Plasticity of CD4 T-cell Memories”, *Immunology*, vol. 130, pp. 463-470, 2010

[8.7] K. K. McKinstry, T. M. Strutt and S. L. Swain, “The Potential of CD4 T-cell Memory”, *Immunology*, vol. 130, pp. 1-9, 2010

[8.8] M. Zanetti, P. Castiglioni and E. Ingulli, “Principles of Memory CD8 T-cells Generation in Relation to Protective Immunity”, *Advances in Experimental Medicine and Biology*, vol. 684, pp. 108-125, 2010

[8.9] I. Kumagai and K. Tsumoto, “Antigen–Antibody Binding”, *Encyclopedia of Life Sciences*, pp. 1-7, 2010

[8.10] R. Goldsby, T. J. Kindt, B. A. Osborne and J. Kuby, “Antigens”, Chapter 3, In *Immunology* , 4th edition, W. H. Freeman and Company, New York, NY: 2003

[8.11] P. Delves, S. Martin, D. Burton and I. Roitt, *Essential Immunology*, 4th edition, Wiley-Blackwell, Oxford, UK: 2006

[8.12] M. Pavlovic, A. Kats, M. Cavallo, R. Chen, J. X. Hartmann and Y. Shoenfeld, “Pathogenic and Epiphenomenal Anti-DNA Antibodies in SLE”, *Autoimmune Dis*ease, vol. 2010, 2010

[8.13] M. Pavlovic, M. Cavallo, A. Kats, A. Kotlarchyk, H. Zhuang and Y. Shoenfeld, “From Pauling's Abzyme Concept to the New Era of Hydrolytic Anti-DNA Autoantibodies: A Link to Rational Vaccine Design? - A Review”, *International Journal of Bioinformatic Research and Application*, vol. 7, no. 3, pp. 220-38, 2011

[8.14] E. Jenner, “An Inquiry into the Causes and Effects of the Variolae Vaccine: A Disease Discovered in Some of the Western Counties of England, Particularly Gloucestershire, and Known by the Name Cow Pox”, In *Classics of Medicine Library*, Birmingham, AL: 1978

[8.15] S. Y. Tan, "Edward Jenner (1749-1823): conqueror of smallpox.", *Singapore Medical Journal*, vol. 45, no. 11, pp. 507–508, 2004

[8.16] P. Debré , “Louis Pasteur”, In *The Johns Hopkins University Press*, Baltimore, MD: 1994

[8.17] A. W. Artenstein and G. A. Poland, “Vaccine history: The Past as Prelude to the Future”, *Vaccine*, vol. 30, no. 36, pp. 5299–5301, 2012

[8.18] R. Rappuoli, "From Pasteur to Genomics: Progress and Challenges in Infectious Diseases", *Nature Medicine*, vol. 10, no. 11, pp. 1177-1185, 2004

[8.19] R. Rappuoli and F. Bagnoli, *Vaccine Design: Innovative Approaches and Novel Strategies*, Caister Academic Press, Norfolk, UK: 2011

[8.20] M. Mora, D. Veggi, L. Santini, M. Pizza and R. Rappuoli, "Reverse Vaccinology", *Drug Discovery Today*, vol. 8, no. 10, pp. 459-464, 2003

[8.21] A. Henke, "DNA Immunization--A New Chance in Vaccine Research?", *Medical Microbiology and Immunology*, vol. 191, no. 3-4, pp. 187-90, 2002

[8.22] S. A. Abdulhaqq and D. B. Weiner, "DNA Vaccines: Developing New Strategies to Enhance Immune Responses", *Immunologic Research*, vol. 42, no. 1-3, pp. 219-232, 2008

[8.23] V. W. Bramwell and Y. Perrie, "The Rational Design of Vaccines", *Drug Discovery Today*, vol. 10, no. 22, pp. 1527-1534, 2005

[8.24] R. Rappuoli, "Reverse Vaccinology, A Genome-based Approach to Vaccine Development", *Vaccine*, vol. 19, no. 17-19, pp. 2688-2691, 2001

[8.25] A. S. DeGroot, H. Sbai, C. Saint Aubin, J. McMurry and W. Martin, "Immuno-informatics: Mining Genomes for Vaccine Components", *Immunology and Cell Biology*, vol. 80**,** pp. 255–269, 2002

[8.26] M. Hagmann. "Computers Aid Vaccine Design", *Science*, vol. 290, no. 5489, pp.80-82, 2000

[8.27] M. Ardito, J. Fueyo, R. Tassone, F. Terry, K. DaSilva, S. Zhang, W. Martin, A. De Groot, S. Moss and L. Moise, "An Integrated Genomic and Immunoinformatic Approach to H. Pylori Vaccine Design" , *Immunome Research*, vol. 7, no. 2, article 1, 12 pages, 2011, (Open Access online)

[8.28] F. R. Burden and D. A. Winkler , "Predictive Bayesian Neural Network Models of MHC Class II Peptide Binding", *Journal of Molecular Graphics and Modelling*, vol. 23, pp. 481–489, 2005

[8.29] P. Donnes and A. Elofsson, “Prediction of MHC Class I Binding Peptides, Using SVMHC”, *BMC Bioinformatics,* vol. 3, p. 25, 2002, (Open Access online)

[8.30] H. Noguchi, R. Kato, T. Hanai, Y. Matsubara, H. Honda, V. Brusic and T. Kobayashi, “Hidden Markov Model-based Prediction of Antigenic Peptides that Interact with MHC Class II Molecules”, *Journal of Bioscience and Bioengineering,* vol. 94, no. 3, pp. 264-270, 2002

[8.31] M. N. Davies and D. R. Flower, “Harnessing Bioinformatics to Discover New Vaccines”, *Drug Discovery Today,* vol. 12, no. 9–10, pp. 389–395, 2007

[8.32] M. J. Cardosa, “Dengue Vaccine Design: Issues and Challenges”, *British Medical Bulletin,* vol. 54, no. 2, pp. 395-405, 1998

[8.33] A. P. Durbin and S. S. Whitehead, “Next-generation Dengue Vaccines: Novel Strategies Currently Under Development”, *Viruses*, vol. 3, pp. 1800-1814, 2011

[8.34] D. Normile, “Mixed Results for Dengue Vaccine Trial”, *Science,* Sept. 2012, Available online: http://news.sciencemag.org/sciencenow/2012/09/mixed-results-for-dengue-vaccine.html

[8.35] R. M. Welsh & R. S. Fujinami, “Pathogenic epitopes, heterologous immunity and vaccine design”, *Nature Reviews Microbiology*, vol. 5, pp. 555-563, July 2007

[9.1] T. V. Arredondo, P.S. Neelakanta and D. DeGroff, “Fuzzy Attributes of a DNA a Fuzzy Inference Engine for Codon -“Junk” Codon Delineation”, *Artificial Intelligence in Medicine,* vol. 35, pp. 87-105, Oct. 2005

[9.2] “Human parvovirus B19, complete genome” [Online], (Accessed on Sept. 15, 2012), Available at: http://www.ncbi.nlm.nih.gov/nuccore/356457872

[9.3] K. Deergha Rao and M.N.S. Swamy, “Analysis of Genomics and Proteomics Using DSP Techniques”, *IEEE Transactions on Circuits and Systems I (regular papers),* vol. 55, no. 1, pp. 370-378, Feb 2008

[9.4] T. V. Arredondo, “CDS Identification in a Viral single-strand DNA (ssDNA) Using Fisher Linear Discriminant”, *International Journal of Bioinformatic Research and Application*, vol. 7, no. 3, pp. 262-272, 2011

[9.5] M. Ruhul Amin, M. S. Siddiqui, D. Ahmed, F. Ahmed and A. Hossain, "B- and T-Cell Epitope Mapping of Human Sapovirus Capsid Protein: An Immunomics Approach", *International Journal of Bioinformatic Research and Application*, vol. 7, no. 3, pp. 287-298, 2011

[9.6] M. Sagar and A. K. Yadav, "Computer-aided Vaccine Design for Liver Cancer Using Epitopes of HBx Protein Isolates from HBV Substrains", *International Journal of Bioinformatic Research and Application*, vol. 7, no. 3, pp. 299-316, 2011

[9.7] A.D. Haimovich, B. Byrne, R. Ramaswamy and W. J. Welsh, "Wavelet Analysis of DNA Walks", *Journal of Computational Biology*, vol. 13, no. 7, pp. 1289-1298, 2006

[9.8] A. Arneodo, Y. d'Aubenton-Carafa, E. Bacry, P. V. Graves, J. F. Muzy, C. Thermes, "Wavelet Based Fractal Analysis of DNA Sequences", *Physica D: Nonlinear Phenomena*, vol. 96, no. 1-4, pp. 291-320, 1996

[9.9] M. El-Zanaty, M. Saeb, A. B. Mohammed, S. K. Guirguis, "Haar Wavelet Transform of the Signal Representation of DNA Sequences", *International Journal of Computer Science and Communication Security (IJCSCS)*, vol. 1, pp. 56-62, July 2011

[9.10] M. Oyapero, "Wavelet Analysis of DNA Sequences Using Integer Representation", *Dissertation, Department of Mathematics, Central Michigan University*, December 2011

[10.1] P. S. Neelakanta, S. Chatterjee, M. Pavlovic, A. Pandya and D. DeGroff, "Fuzzy Splicing in Precursor-mRNA Sequences: Prediction of Aberrant Splice-junctions in Viral DNA Context", *Journal of Biomedical Science and Engineering (JBiSE)*, vol. 4, no. 4, pp. 272-281, April 2011

[10.2] P. S. Neelakanta, S. Chatterjee, D. Pappusetty, M. Pavlovic and A. Pandya, "Information-theoretic Algorithms in Bioinformatics and Bio-/Medical-imaging: A Review", *Proceedings of the IEEE International Conference on Recent Trends in Information Technology (IEEE ICRTIT)*, Chennai, India, pp. 183-188, June 2011

[10.3] P. S. Neelakanta, S. Chatterjee and G. A. Thengum-Pallil, “Computation of Entropy and Energetics Profiles of a Single-stranded Viral DNA”, *International Journal of Bioinformatics Research and Applications (IJBRA),* vol. 7, no. 3, pp. 239-261, August 2011