CONTENTS

| ABSTRACT | vii |
| LIST OF FIGURES | xv |
| LIST OF TABLES | xvi |

CHAPTER 1 COMPUTING LIFE 1-13

1.1 Introduction 1
1.2 Bioinformatics: Problems and challenges 3
   1.2.1 Maintaining biological data bases and designing systems for data retrieval 3
   1.2.2 Algorithms and tools for sequence alignment 4
   1.2.3 Molecular phylogenetcs 5
   1.2.4 Gene Finding 5
   1.2.5 Evaluation of gene expressions 7
   1.2.6 Protein Structure prediction, Visualization and modelling 8
   1.2.7 Emerging areas and problems 9
1.3 Objectives of the present work 11
1.4 Organization of the thesis 13

CHAPTER 2 THE BIOLOGICAL BACKDROP 14-37

2.1 Introduction 14
2.2 Central dogma of molecular biology 14
2.3 Cells 15
2.4 DNA 18
2.5 Cell division and DNA replication 20
2.6 The process of protein making
  2.6.1 Transcription
  2.6.2 Translation
  2.7 Regulation of gene expression
  2.8 Proteins
    2.8.1 Classification of aminoacids
    2.8.2 Different levels of protein structure
    2.9 Human genome-the organization
    2.10 Conclusion
    2.11 A brief glossary of selected biological terms

CHAPTER 3 THE EXON FINDING PROBLEM

3.1 Introduction
3.2 Finding genes in prokaryotes
3.3 Structure of genes in eukaryotes
3.4 Features of genomes used for exon finding
    3.4.1 The signals
    3.4.2 Coding measures
3.5 Methods and programs
    3.5.1 Extrinsic methods
    3.5.2 Intrinsic methods
    3.5.3 Combinational methods
3.6 Measuring the performance of gene prediction programs
3.7 Conclusion

CHAPTER 4 GENOMIC SIGNAL PROCESSING AND NEW SCHEMES OF CODING MEASURES

4.1 Introduction
4.2 An overview of Digital Signal Processing
4.2.1 Digital filters 71
4.2.2 Finite Impulse Response (FIR) filters 72
4.2.3 Infinite Impulse Response (IIR) Filters 74

4.3 Genomic Signal Processing 75
4.3.1 Conversion of DNA sequences to Binary indicator sequences and a preliminary spectral measure for coding regions 75
4.3.2 A discussion on N/3 peaks 78
4.3.3 Mapping DNA sequences to complex number sequences 79
4.3.4 Mapping DNA sequences to Cumulative Categorical Periodogram (CCP) sequences 80

4.4 The proposed coding measure scheme using Electron Ion Interaction Pseudo potential (EIIP) indicator sequences. 80

4.5 Results of the experiments on N/3 periodicity as a coding measure using binary sequence indicators and the proposed EIIP indicator sequence 83

4.6 Existing methods of locating exons and gene finding using genomic signal processing 95
4.6.1 Finding exon locations using windowing technique as done in Genescan 95
4.6.2 Finding exon locations using optimization technique 96
4.6.3 Finding exon locations using position count function 99
4.6.4 Finding exon locations using digital filtering techniques 100

4.7 Proposed coding measure for improving the digital filtering technique by incorporating coding statistics 104
4.7.1 Using single nucleotide bias 105
4.7.2 Using dinucleotide bias 106
4.7.3 Using trinucleotide bias 107
4.7.4 Results of the experiments on incorporating coding statistics into existing digital filtering technique 108
4.7.5 A discussion on incorporating coding statistics into genomic signal processing 114

4.8 Proposed method for improving the filtering algorithm by replacing four binary indicator sequences by a single EIIP indicator sequence 117

  4.8.1 Results of the experiments using EIIP indicator sequence instead of binary indicator sequences in existing digital filtering technique 117

4.9 Comparison of the performance of sliding window techniques of locating exons using binary indicator sequences and EIIP indicator sequence 122

  4.9.1 Results of the comparison of the performance of sliding window techniques of locating exons using binary indicator sequences and EIIP indicator sequence 124

4.10 Locating single genes using genomic signal processing incorporating the proposed coding measures 130

  4.10.1 Results of employing different digital mappings to find single genes 131

4.11 A framework to find genes from eukaryotic genomes employing the proposed coding measures 133

4.12 Conclusion 137

CHAPTER 5 Conclusions and directions for further study 138-147

5.1 Introduction 138

5.2 Objectives revisited 138

5.3 Major findings and conclusions 139

5.4 Directions for further study 143

  5.4.1 Wavelets—the mathematical microscopes 143

  5.4.2 Combinational methods of gene identification—the integration of algorithms and procedures 144

  5.4.3 Experimenting with other parameters 145

xiii
5.4.4 Experimenting with sliding window size

5.4.5 Employing electron-ion interaction pseudopotential to other exon finding techniques.

5.4.6 Extending genomic signal processing techniques to other Bioinformatics and Computational Biology problems

5.4.7 Closing Remarks

REFERENCES