

Contents

List of Figures	I
List of Tables	II
1 Introduction	1
1.1 Objective and Motivation	5
1.2 Organization of the Dissertation	6
2 Literature Review	7
2.1 Bacterial Genomics	8
2.1.1 Genomic data as a cornerstone for Bio-sciences	9
2.1.2 The Birth of Bacterial Genome Sequencing	9
2.1.3 G+C Content Variation	10
2.2 Genomic Features	11
2.2.1 Biomolecular Sequences	11
2.3 Sequential Data	12
2.3.1 Genome Modelling Tools	12
2.3.2 Application	13
2.4 Computational Sequence Analysis	13
2.4.2 Escherichia Coli and Related Species	14
2.5 Markovian Analysis	16
2.5.1 Markov Process	16
2.5.2 Markov Chain	17
2.5.3 Markov Model	19

2.5.4	Hidden Markov Model	20
2.5.4.1	Introduction	20
2.5.4.2	Example of a HMM	21
3	Proposed Method	25
3.1	Introduction	26
3.2	Proposed Methodology	27
3.2.1	Assumptions and Terminologies used	31
3.2.2	Proposed Algorithm	32
3.2.3	Complexity Analysis	33
4	Results and Discussion	34
4.1	Experimental Results	35
4.2	Discussion	38
5	Conclusion and Future Work	39
5.1	Conclusion	40
5.2	Future Work	41
	Bibliography	43
	Appendices	47

List of Figures

1	Sample intergenic DNA from Genomic Sequence	7
2.2.1	Example of a Typical DNA Sequence	11
2.5.2	A Markov chain for modeling a DNA sequence	18
2.5.4.3	HMM for CpG islands and non-CpG islands	23
3.2	Proposed Markov Chain	30
3.2	Proposed Hidden Markov Model	31
4.1	Plot of First four Altered GC Regions by our method	39
5.2	HMM combined with selection, mutation, crossover	44
	Appendices	49

List of Tables

2.5.2	Transition probabilities inside/outside a CpG island	19
2.5.4.3	Transition Probability matrix for our HMM	24
3.2	Transition table for Proposed Markov chain	30
3.2	Intra State Transition table for average GC region, high GC region, and low GC region	30
3.2	Inter State Transition table for High state to Average state (T1), Average state to Low state (T2)	31
4.1	Existing altered-GC regions (High/Low) in E. coli CFT073 strain	38
4.1	Potential Altered GC regions (High/Low)	38