Contents

Acknowledgments vii
Abstract ix
Resumo xi
Preface xv

1 Introduction 1
1.1 Bioinformatics in the post-genomic era 2
1.2 Molecular biology and genetics 3
1.2.1 Biological sequences 3
1.2.2 Synthesis of macromolecules and the central dogma . . . 8
1.3 Sequence analysis and comparison 11
1.3.1 Alignment methods for sequence comparison 11
1.3.2 Vector maps for sequence comparison 14
1.4 Iterative Function Systems 15
1.4.1 Definitions 15
1.4.2 Chaos game representation (CGR) 17
1.5 Entropy and information theory 22
1.5.1 Shannon's entropy function 23
1.5.2 Rényi's entropy generalization 25
1.6 Thesis outline 27
1.7 References 29

2 Alignment-free sequence comparison – a review 33
2.1 Introduction 34
2.2 Background 35
2.2.1 Words in sequences 36
2.2.2 Distance between sequences 36
2.2.3 Word statistics 37
2.2.4 Information theory 37
2.3 Alignment-free sequence comparison 38
2.3.1 Methods based on word frequencies 39
2.3.2 Resolution free methods 45
2.4 Algorithm implementation –NASC-Toolbox 47
2.4.1 Matlab functions 48
2.4.2 DNA 49
2.4.3 Proteins 49
2.4.4 Natural languages 54

2.5 Conclusions 58
2.6 Acknowledgements 58
2.7 References 58

3 Universal sequence map (USM) 65
3.1 Background 66
3.2 Results 66
3.2.1 Conceptual foundations 67
3.2.2 Implementation of USM algorithm 67
3.3 Discussion 75
3.4 Conclusions 79
3.5 Methods 79
3.5.1 Computation 79
3.5.2 Source code 79
3.5.3 Test data 79
3.6 Acknowledgements 80
3.7 References 80

4 Comparative evaluation of word composition distances 83
4.1 Introduction 84
4.2 Systems and Methods 85
4.2.1 Word statistics 85
4.2.2 W-metric definition 86
4.2.3 ROC curve definition 87
4.2.4 Protein test datasets – SCOP/ASTRAL classification . . 88
4.2.5 Protocol for comparative assessment 89
4.2.6 Computation 90
4.3 Results and Discussion 90
4.3.1 Complete dataset 90
4.3.2 Stratified analysis by class 95
4.4 Conclusion 97
4.5 Acknowledgements 98
4.6 References 98

5 Rényi continuous entropy of DNA sequences 101
5.1 Introduction 101
5.2 System and methods 104
5.2.1 CGR/USM representation of a sequence 104
5.2.2 Rényi continuous entropy definition 106
5.2.3 Parzen window density estimation 107
5.2.4 Simplification of Rényi entropy calculation for USM maps 107
5.2.5 Asymptotic properties of H_2 and random sequence simulation 108
5.2.6 DNA sequence dataset description 109
5.3 Results and Discussion 110
5.3.1 Rényi continuous quadratic entropies H_2 111
5.3.2 Equivalent sequence length N_{eq} 114

5.3.3 Comparison between continuous and discrete measures of entropy . . . 115
5.3.4 Algorithm implementation – Rényi-Toolbox . . . 116
5.4 Conclusions . . . 116
5.5 Acknowledgements . . . 117
5.6 Appendix . . . 117
A. Gaussian or Normal distribution function definition . . . 117
B. Rényi quadratic entropy simplification . . . 118
C. Asymptote calculation . . . 118
D. Convolution of Normal distribution functions . . . 120
5.7 References . . . 123

6 Final discussion 127
6.1 References . . . 130

Index 133