

## CONTENTS

Contents .....	v
Preface.....	x
Acknowledgement .....	xiv
Chapter 1 BLAST and FASTA.....	1
1. Introduction.....	1
2. Mathematics of string matching.....	4
2.1 Basic concepts .....	4
3. String-matching algorithms in FASTA and BLAST.....	11
3.1 FASTA .....	11
3.2 BLAST .....	16
4. Homology search and sequence annotation .....	20
5. Postscript.....	21
Chapter 2 Sequence alignment.....	23
1. Introduction.....	23
2. Pairwise alignment.....	24
2.1 Pairwise alignment with constant gap penalty.....	25
2.1.1 Global alignment .....	25
2.1.2 Local alignment .....	29
2.1.3 The simple scoring scheme needs extension .....	30
2.2 Pairwise alignment with a similarity matrix.....	30
2.2.1 DNA matrices .....	30
2.2.2 Protein matrix .....	32
2.3 Pairwise alignment with gap penalty specified by the affine function .....	34
3. Multiple sequence alignment .....	38
3.1 Profile alignment .....	38
3.2 Multiple alignment with a guide tree.....	41
4. Sequence alignment with secondary structure .....	43

5. Align nucleotide sequences against amino acid sequences.....	44
6. Postscript.....	47
Chapter 3 Contig assembly .....	49
1. Introduction.....	49
2. Skeletal output of contig assembly .....	51
3. string matching of two sequence ends .....	56
4. New development in contig assembly.....	59
5. Postscript.....	60
Chapter 4 DNA replication and viral evolution .....	62
1. Introduction.....	62
2. Fundamentals of viruses.....	63
2.1 The virion and the viral genome.....	64
2.2 Variation in viral genome size can be explained by variation in mutation rate.....	65
2.3 A representative virus: Phage $\lambda$ .....	66
3. Fundamentals of bacterial species.....	67
4. genomic AT% of bacterial species is indicative of cellular AT availability.....	68
5. Formulating the hypothesis and predictions.....	70
6. Are our predictions supported?.....	71
7. A short play featuring phages and bacteria .....	77
Chapter 5 Gene and motif prediction .....	78
1. Introduction.....	78
2. Bayes' theorem and odds ratios .....	79
3. Characterizing features of Signal sensor.....	83
3.1 Position weight matrix.....	83
3.2 Perceptron.....	93
4. Characterizing features of Content sensors.....	100
4.1 Indices of content sensors related to DNA methylation and spontaneous deamination.....	101
4.2 Are these indices useful in discriminating between coding and non-coding sequences? .....	104
Chapter 6 Hidden Markov Models.....	109
1. Introduction.....	109
2. Markov models .....	110
3. Hidden Markov Models .....	115
3.1 The Essential Elements in a Hidden Markov Model .....	115
3.2 Training HMM .....	117
3.3 The Viterbi algorithm .....	120
3.4 Forward algorithm.....	127
3.5 HMM and gene prediction.....	130
4. Postscript.....	131

Chapter 7 Gibbs Sampler .....	133
1. Introduction.....	133
2. A numerical illustration of the computational details of Gibbs sampler.....	135
2.1 Initialization.....	136
2.2 Predictive update .....	137
3. Motif sampler.....	146
Chapter 8. Bioinformatics and vertebrate mitochondria .....	148
1. Introduction.....	148
1.1 Mitochondria and mitochondrial genomes .....	149
1.2 DNA-Replication and Strand-biased mutation spectrum .....	150
1.3 The effect of strand-biased mutation on codon usage .....	152
2. Three hypotheses on tRNA anticodon .....	156
2.1 The mutation hypothesis.....	157
2.2 The codon-anticodon adaptation hypothesis .....	157
2.3 The wobble versatility hypothesis .....	158
3. Empirical evaluation of the three alternative hypotheses.....	160
3.1 Evaluation of the mutation hypothesis against the two selectionist hypotheses .....	160
3.2 Evaluating the two selectionist hypotheses .....	161
4. Integrating the codon-anticodon adaptation hypothesis (CAAH) and the wobble versatility hypothesis (WVH).....	162
4.1 Four-fold NNN codons.....	163
4.2 Two-fold NNY codon families.....	164
4.3 Two-fold NNR codon families .....	165
5. Conflict between translation initiation and elongation .....	165
6. PostScript .....	171
Chapter 9. Characterizing translation efficiency.....	173
1. Introduction.....	173
2. RSCU (Relative synonymous codon usage) .....	175
3. CAI (Codon adaptation index).....	176
3.1 Computation and basic properties of CAI .....	176
3.2 Problems with CAI and its current implementation .....	178
3.2.1 Problem when $w = 0$ .....	178
3.2.2 Problems with codon families containing a single codon 179	
3.2.3 Problems with amino acids coded by two different codon families .....	180
3.2.4 Problems with initiation and termination codons .....	181
3.2.5 The problem with the compilation of the reference set of genes.....	181
4. Indices of codon-anticodon adaptation .....	182

4.1	CAI with a tRNA anticodon-derived codon usage table .....	185
4.2	Codon-anticodon adaptation index (CAAI).....	187
5.	Why CAI or CAAI should not be taken as a measure of gene expression?.....	192
6.	Will AT-rich mRNA be translated inefficiently?.....	200
7.	Codon adaptation index and proteomics: clarification of some misunderstandings.....	201
Chapter 10	Protein isoelectric point.....	207
1.	Introduction.....	207
2.	Amino acid and protein isoelectric point .....	209
3.	Genomic profiling of protein isoelectric point: a case study with <i>Helicobacter pylori</i> .....	213
4.	An alternative explanation of <i>H. pylori</i> data .....	217
Chapter 11	Bioinformatics and Two-Dimensional Protein Separation.....	220
1.	Introduction.....	220
2.	Scientific rationale behind the 2D-SDS-PAGE .....	221
3.	Expected separation pattern of 2D-SDS-PAGE for the genome-derived proteome .....	222
4.	posttranslational modification.....	227
4.1	Importance in studying posttranslational modification .....	227
4.2	Posttranslational modification changes the migration pattern of proteins on 2D-SDS-PAGE.....	227
Chapter 12	Self-Organizing Map and other clustering Algorithms .....	231
1.	Introduction.....	231
1.1	Classification and clustering.....	231
1.2	Clustering and gene expression .....	233
1.3	Similarity and distance indices .....	235
2.	UPGMA .....	239
3.	Self-organizing map (SOM).....	243
3.1	The SOM algorithm.....	244
3.2	Variations of the basic SOM algorithm .....	249
Chapter 13	Molecular Phylogenetics .....	251
1.	Introduction.....	251
2.	Biodiversity, historical information, and phylogenetics .....	252
3.	Substitution models.....	253
3.1	Nucleotide-based substitution models and genetic distances .....	254
3.2	Amino acid-based and codon-based substitution models.....	264
4.	Tree-building methods .....	266
4.1	Distance-based methods .....	266

4.2	Maximum parsimony methods .....	272
4.2.1	The Fitch algorithm .....	272
4.2.2	The uphill search and branch-and-bound search algorithms .....	275
4.2.3	The long-branch attraction problem .....	277
4.3	Maximum likelihood methods .....	279
4.4	Bayesian inference .....	283
4.4.1	Bayes theorem for a continuous variable .....	283
4.4.2	Alternative computational approaches in Bayesian inference .....	287
Chapter 14 Fundamentals of Proteomics .....		293
1.	Introduction .....	293
2.	Protein Mass Spectrometry .....	294
3.	Charge deconvolution .....	295
4.	Peptide mass fingerprinting .....	301
4.1	Peptide digestion .....	301
4.2	MS determination of peptide mass .....	304
4.3	Protein database and <i>in silico</i> digestion .....	305
4.4	Protein identification .....	305
References .....		309
Postscript .....		342
Index .....		344

Bioinformatics and the Cell  
Modern Computational Approaches in Genomics,  
Proteomics and Transcriptomics

Xia, X.

2007, XVI, 350 p., Hardcover

ISBN: 978-0-387-71336-6