

Contents

PREFACE

CHAPTER 1 SOME BIOLOGICAL CONCEPTS	1
1.1 Cell	1
1.2 Genetic Material: DNA, Gene and RNA.....	1
1.2.1 DNA	1
1.2.2 Gene.....	3
1.2.3 RNA.....	5
1.3 Protein and Amino Acids.....	7
1.4 Chromosome.....	9
1.5 Omics.....	10
1.5.1 Genomics.....	11
1.5.2 Microarray	11
1.5.3 Proteomics	11
1.5.4 Lipidomics.....	12
REFERENCES	13
CHAPTER 2 GRAPHICAL REPRESENTATIONS OF DNA SEQUENCE.....	14
2.1 Three-Dimension (3-D) Graphical Representation.....	14
2.2 2-D Graphical Representation	15
2.3 2-D Graphical Representations Without Degeneracy	17
2.4 Used a 1-D Numerical Representation of four Nucleotides to Construct a 2-D Graphical Representation of the DNA Sequence	22
REFERENCES	22
CHAPTER 3 NUMERICAL REPRESENTATIONS OF DNA SEQUENCE.....	24
3.1 4-D and 3-D Numerical Representations of a DNA Sequence	24
3.2 2-D Numerical Representations of a DNA Sequence.....	25
3.3 The Complex Numerical Representation.....	26
3.4 1-D Numerical Representations of four Nucleotides and 2-D Graphical Representation of a DNA Sequence	27
3.5 The Representations of Feature Vector, Genome Space and Matrix Representation of DNA Sequence	27

3.6	The Numerical Representation Based on Physical, Chemical and Structural Properties of DNA Sequence	29
3.6.1	The numerical representations based on some attribute equivalences of nucleotides	29
3.6.2	The representation of DNA by the inspiration from codon and the idea of three attribute equivalences	31
3.6.3	EIIP numerical representation for nucleotides	31
	REFERENCES	32
CHAPTER 4 NUMERICAL REPRESENTATIONS OF PROTEIN.....		33
4.1	1-D Numerical and Graphical Representations of the Amino Acid Sequence	33
4.2	2-D Numerical and Graphical Representations of the Amino Acid Sequence	34
4.3	A 2-D Graphical Representation and Moment Vector Representation of Protein	41
4.4	3-D Numerical Representation of Protein	44
4.5	The 10-D Representation of an Amino Acid	45
4.6	The Vector and Matrix Representations of Protein Sequence and Protein Space	46
4.7	Other Schemes of the Representation for Protein	46
	REFERENCES	47
CHAPTER 5 PRACTICAL ORTHOGONAL TRANSFORM		49
5.1	Some Features and Algorithms for the Discrete Fourier Transform	49
5.1.1	Fourier transforms of the original sequence and its subsequence	49
5.1.2	The independency of the Fourier transforms at several frequencies	52
5.1.3	The Fourier transform of symbolic sequence	53
5.1.4	Fourier transform of binary sequence	56
5.1.5	Several algorithms of Fourier transform	57
5.1.6	The properties of Fourier transform of real sequence	59
5.2	Wavelet Analysis	64
5.2.1	Introduction	64
5.2.2	Multiresolution analysis of a function by Haar scaling and wavelet function	65
5.2.3	Construction of wavelet systems	73
5.2.4	Mallet transform	78
	REFERENCES	82
CHAPTER 6 IDENTIFYING PROTEIN-CODING REGIONS (EXONS) BY NUCLEOTIDE DISTRIBUTIONS		83

6.1	Protein Coding Regions Finding in DNA Sequence.....	83
6.1.1	Introduction	83
6.1.2	The stochastic simulation and several computing formulae	84
6.1.3	FEND algorithm, predicting protein coding regions from nucleotide distributions on the three positions of a DNA sequence	97
6.1.4	Performance evaluation of FEND algorithm	104
6.2	The Experiment for Distinguishing Exon and Intron Sequences by a Threshold	105
6.2.1	Motivation	105
6.2.2	Idea of distinguishing exon and intron sequences	106
6.2.3	Results and discussion	108
	REFERENCES	109
CHAPTER 7 PROTEIN COMPARISON BY ORTHOGONAL TRANSFORMS		
		111
7.1	Protein Comparison by Discrete Fourier Transformation (DFT)	111
7.1.1	EIIP representation of protein sequence	111
7.1.2	Symmetry of discrete Fourier transform of real sequence	112
7.1.3	Cross-spectral function	112
7.2	Protein Comparison by Discrete Wavelet Transformation	115
7.2.1	Several techniques needed for DWT method	115
7.2.2	The performance of the DWT method	120
	REFERENCES	124
CHAPTER 8 THE APPLICATION OF VECTOR REPRESENTATIONS TO BIOLOGICAL MOLECULE ANALYSIS.....		125
8.1	Use Feature Vector to Analyze DNA Sequences	125
8.1.1	Feature vector representation of DNA sequence	125
8.1.2	Comparing DNA sequences	126
8.2	A Protein Map and its Applications	129
8.2.1	Recalling a 2-D graphical representation and moment vector representation of protein	129
8.2.2	Protein map and cluster analysis	129
8.3	An Appendix: Introduction to Cluster Analysis	133
	REFERENCES	137
CHAPTER 9 THE STATISTICS ANALYSIS OF LARGE AMOUNT OF EXPERIMENTAL DATA.....		138

9.1	A Way to Process Microarray Data.....	138
9.1.1	Data form.....	138
9.1.2	Microarray data set.....	140
9.1.3	Preliminary filtering.....	140
9.1.4	Assessing normalization.....	141
9.1.5	Hypothesis test.....	144
9.1.6	Conclusion.....	146
9.2	The Statistical Analysis of a Set of Lipidomics Data	146
9.2.1	Introduction	146
9.2.2	Statistical techniques of initial data processing	148
9.2.3	Initial data arrangement.....	150
9.2.4	Hypothesis testing analysis.....	154
	REFERENCES.....	155
CHAPTER 10 APPLY SINGULAR VALUE DECOMPOSITION TO MICROARRAY ANALYSIS		156
10.1	SVD, PCA and GSVD.....	156
10.1.1	Singular value decomposition	156
10.1.2	Principal component analysis.....	157
10.1.3	Generalized singular value decomposition.....	159
10.2	Apply SVD/PCA to Microarray Analysis.....	161
10.3	GSVD Analyzes the Microarray Data	165
	REFERENCES.....	169
CHAPTER 11 DYNAMICAL ANALYSIS MODELS OF GENE EXPRESSION		170
11.1	Differential Equations Model of Gene Expression.....	170
11.1.1	Transcription model.....	170
11.1.2	Nonlinear dynamic equations	171
11.1.3	Linearization of the nonlinear transcription model.....	172
11.1.4	Approximating coefficient matrix M by Fourier series	173
11.1.5	Solution to transcription matrix C and V	175
11.2	Modified Linear Differential Equations Model	176
11.3	Dynamical Model Based on Singular Value Decomposition.....	178
11.3.1	Introduction.....	178
11.3.2	Reducing gene's number	179
11.3.3	The approach based on singular value decomposition (SVD)	179

11.3.4 The methods of solving dynamical models	182
REFERENCES	183
CHAPTER 12 MISSING MICROARRAY DATA INPUTTING	184
12.1 The Ad Hoc Methods.....	184
12.2 Missing Data Inputting Based on SVD	186
12.2.1 A new way for missing data inputting	186
12.2.2 Other method based on SVD	188
12.3 Weighted K-Nearest Neighbors, KNN, Impute Algorithm.....	189
12.4 Estimation of Missing Values in Microarray Data Based on the Least Square Principle.....	190
12.4.1 Least squares estimate of the unknown variable.....	190
12.4.2 The least square estimation of missing data based on genes	191
12.4.3 The least square estimation of missing data based on arrays.....	192
12.4.4 Combining the gene and array based estimates	192
12.5 Local Least Square Inputting (LLSinpute)	193
12.5.1 Selecting genes	194
12.5.2 Gene-wise formulation of local least squares imputation.....	195
12.6 The Comparison of the Methods of Missing Data Inputting	197
REFERENCES	199
PLATE	

1.2 Genetic Material: DNA, Gene and RNA

1.2.1 DNA

DNA is the basic information macromolecule of life. It consists of a polymer of nucleotides, in which each nucleotide is composed of a standard deoxyribose sugar and phosphate group unit, connected to a nitrogenous base of one of four types: adenine, guanine, cytosine, or thymine, abbreviated as A, G, C, and T, respectively.

It is easy to appreciate the chemical structure of these nitrogenous bases, adenine