

Contents

1	The UNIX Command Line	1
1.1	Getting Started	2
1.2	Files	7
1.3	Scripts	13
1.3.1	Bash	14
1.3.2	Sed	16
1.3.3	AWK	17
2	Constructing and Applying Optimal Alignments	23
2.1	Sequence Evolution and Alignment	23
2.2	Amino Acid Substitution Matrices	25
2.2.1	Genetic Code	26
2.2.2	PAM Matrices	30
2.3	The Number of Possible Alignments	32
2.4	Dot Plots	34
2.5	Optimal Alignment	37
2.5.1	From Dot Plot to Alignment	38
2.5.2	Global Alignment	39
2.5.3	Local Alignment	42
2.6	Applications of Optimal Alignment	42
2.6.1	Homology Detection	43
2.6.2	Dating the Duplication of Adh	44
3	Exact Matching	47
3.1	Keyword Trees	47
3.2	Suffix Trees	54
3.3	Suffix Arrays	57
3.4	Text Compression	62
3.4.1	Move to Front (MTF)	65
3.4.2	Measuring Compressibility: The Lempel–Ziv Decomposition	65

4 Fast Alignment	69
4.1 Alignment with k Errors	69
4.2 Fast Local Alignment	72
4.2.1 Simple BLAST	73
4.2.2 Modern BLAST	75
4.3 Shotgun Sequencing	78
4.4 Fast Global Alignment	82
4.5 Read Mapping	86
4.6 Clustering Protein Sequences	88
4.7 Position-Specific Iterated BLAST	92
4.8 Multiple Sequence Alignment	94
4.8.1 Query-Anchored Alignment	96
4.8.2 Progressive Alignment	96
5 Evolution Between Species: Phylogeny	101
5.1 Trees of Life	101
5.2 Rooted Phylogeny	106
5.3 Unrooted Phylogeny	108
6 Evolution Within Populations	113
6.1 Descent from One or Two Parents	113
6.1.1 Bi-Parental Genealogy	113
6.1.2 Uni-Parental Genealogy	115
6.2 The Coalescent	120
7 Additional Topics	127
7.1 Statistics	127
7.1.1 The Significance of Single Experiments	128
7.1.2 The Significance of Multiple Experiments	128
7.1.3 Mouse Transcriptome Data	130
7.2 Relational Databases	131
7.2.1 Mouse Expression Data	132
7.2.2 SQL Queries	135
7.2.3 Java	136
7.2.4 ENSEMBL	137
Answers	139
Appendix: UNIX Guide	293
References	309
Index	313