

Contents

Preface	xii
Overview	xv
1 Introduction	xvi
2 Information networks	xvii
3 Protein information resources	xvii
4 Genome information resources	xvii
5 DNA sequence analysis	xviii
6 Pairwise alignment techniques	xviii
7 Multiple sequence alignment	xviii
8 Secondary database searching	xviii
9 Building a sequence search protocol	xix
10 Analysis packages	xix
1 Introduction	1
1.1 Introduction	1
1.2 The dawn of sequencing	1
1.3 What is bioinformatics?	2
1.4 The biological sequence/structure deficit	3
1.5 Genome projects	4
1.6 Status of the human genome project	6
1.7 Why is bioinformatics important?	6
1.8 Pattern recognition and prediction	7
1.9 The folding problem	9
1.10 The role of chaperones	10
1.11 Sequence analysis	10
1.12 Homology and analogy	12
1.13 The devil is in the detail	13
1.14 Summary	16
1.15 Further reading	16

2 Information networks	19
2.1 Introduction	19
2.2 What is the Internet?	19
2.3 How do computers find each other?	19
2.4 Facilities used on the Internet	20
2.5 What is the World Wide Web?	21
2.6 Web browsers	21
2.7 HTTP, HTML and URLs	24
2.8 The European Molecular Biology network – EMBnet	24
2.9 The National Center for Biotechnology Information– NCBI	30
2.10 Virtual tourism	31
2.11 Summary	33
2.12 Further reading	33
2.12 Web addresses	34
3 Protein information resources	35
3.1 Introduction	35
3.2 Biological databases	35
3.3 Primary sequence databases	36
3.4 Composite protein sequence databases	43
3.5 Secondary databases	45
3.6 Composite protein pattern databases	62
3.7 Structure classification databases	62
3.8 Summary	65
3.9 Further reading	66
3.10 Web addresses	67
4 Genome information resources	69
4.1 Introduction	69
4.2 DNA sequence databases	69
4.3 Specialised genomic resources	75
4.4 Summary	78
4.5 Further reading	78
4.6 Web addresses	80
5 DNA sequence analysis	81
5.1 Introduction	81
5.2 Why analyse DNA?	81
5.3 Gene structure and DNA sequences	82
5.4 Features of DNA sequence analysis	85
5.5 Issues in the interpretation of EST searches	89
5.6 Two approaches to gene hunting	93
5.7 The expression profile of a cell	94
5.8 cDNA libraries and ESTs	95
5.9 Different approaches to EST analysis	96
5.10 Effects of EST data on DNA databases	100

5.11 A practical example of EST analysis	103
5.12 Summary	106
5.13 Further reading	107
6 Pairwise alignment techniques	108
6.1 Introduction	108
6.2 Database searching	108
6.3 Alphabets and complexity	109
6.4 Algorithms and programs	109
6.5 Comparing two sequences – a simple case	109
6.6 Sub-sequences	110
6.7 Identity and similarity	111
6.8 The Dotplot	116
6.9 Local and global similarity	119
6.10 Global alignment: the Needleman and Wunsch algorithm	119
6.11 Local alignment: the Smith–Waterman algorithm	122
6.12 Dynamic programming	125
6.13 Pairwise database searching	125
6.14 Summary	130
6.15 Further reading	131
7 Multiple sequence alignment	132
7.1 Introduction	132
7.2 The goal of multiple sequence alignment	132
7.3 Multiple sequence alignment: a definition	133
7.4 The consensus	134
7.5 Computational complexity	134
7.6 Manual methods	135
7.7 Simultaneous methods	136
7.8 Progressive methods	136
7.9 Databases of multiple alignments	137
7.10 Searching databases with multiple alignments	141
7.11 Summary	142
7.12 Further reading	143
8 Secondary database searching	145
8.1 Introduction	145
8.2 Why bother with secondary database searches?	145
8.3 What's in a secondary database?	146
8.4 Summary	161
8.5 Further reading	162
9 Building a sequence search protocol	163
9.1 Introduction	163
9.2 A practical approach	163

9.3	When to believe a result	178
9.4	Structural and functional interpretation	179
9.5	Summary	184
9.6	Further reading	184

10	Analysis packages	185
10.1	Introduction	185
10.2	What's in an analysis package?	185
10.3	Commercial databases	188
10.4	Commercial software	188
10.5	Comprehensive packages	189
10.6	Packages specialising in DNA analysis	192
10.7	Intranet packages	193
10.8	Internet packages	194
10.9	Summary	197
10.10	Further reading	198
10.11	Web addresses	198
	Glossary	199
	Index	211

Further information for use with this book can be accessed via a link in the catalogue entry for *Introduction to Bioinformatics* on the publisher's Web site at <http://www.awl-he.com/biology>. The aim of this facility is to provide updates and news of major advances in bioinformatics since publication. It also includes study guides for chapters, and cites additional Web site addresses which provide information of interest to biologists. We would encourage readers to use the Web site regularly.