

Nucleic acid and protein sequence analysis a practical approach

Edited by

M J Bishop

University of Cambridge, Computer Laboratory,
Corn Exchange Street, Cambridge CB2 3QG, UK

C J Rawlings

Biomedical Computing Unit, Imperial Cancer Research Fund,
PO Box 123, Lincoln's Inn Fields, London WC2A 3PX, UK

FACHBEREICH BIOLOGIE (10)
der Technischen Hochschule Darmstadt

– Bibliothek –

D – 6100 Darmstadt / B. R. D.

Schnittsahnstraße

Inv.-Nr. 11034

 **IRL PRESS**

Oxford · Washington DC

Contents

INTRODUCTION	1
Sir Walter Bodmer	
1. INTRODUCTION TO COMPUTER HARDWARE AND SYSTEMS SOFTWARE	3
M.J.Bishop	
Introduction	3
The Computer	5
The Peripherals	7
Storage	7
Terminals	8
Printers	9
Digitizers	10
Speech synthesis	10
Computer Communications	10
Asynchronous communications	10
Local area networks	11
Wide area networks	12
Example Systems	13
IBM Personal Computer AT	13
MicroVAX II	15
Sun-3	16
References	17
2. DNA SEQUENCE ANALYSIS SOFTWARE	19
P.A.Stockwell	
Introduction	19
The Range of Different Programs	19
Sources of software	19
Quality of software	20
Sequence Files	21
<i>Sequential text files</i>	21
Direct access files	23
Sequence symbols	23
Sequence Editing	24
Line editors	25
Screen-orientated editors	25
Hybrid editors	27
Choosing and learning a sequence editor	28
Security against data loss with editors	29
Simulation of artificial sequence constructs	30

Sequence Manipulation	31
Transcription – RNA to DNA	31
Complementing sequences	31
Sequence translation	33
Finding open reading frames	33
Sequence Composition	33
Base composition	33
Dinucleotide frequency	34
Codon frequency	35
Mapping	35
Restriction endonuclease mapping	36
Mapping other sequence features	38
Sequence Comparison	40
DNA/DNA similarities	41
Protein/protein similarities	41
Repeats and palindromes	43
Sequence Conversion	43
Formatting	44
Simple formatting	44
Formatting programs	44
Acknowledgements	45
References	45

3. USE OF COMMERCIAL SOFTWARE ON IBM PERSONAL COMPUTERS	47
P.Hoyle	
Introduction	47
Applications of Computing to the Biological Sciences	47
A Comparison of Commercial Software for IBM Personal Computers	48
Hardware requirements	48
Customer support	51
The software	53
Gene Design with the DNASTAR System	64
The design objectives	64
The steps of the gene design	68
The expression vector	68
Sites to design into the gene	69
Designing a DNA sequence that codes for the protein	70
Designing the restriction sites into the sequence	72
Checking candidate sites for consistency with the protein	75
Insertion of the gene into the vector	79
Further investigations	81
Conclusion	81
Acknowledgements	82
References	82

4. MOLECULAR SEQUENCE DATABASES	83
M.J.Bishop, M.Ginsburg, C.J.Rawlings and R.Wakeford	
Introduction	83
Growth of sequence databases	83
Co-ordination of sequence data	83
Submission of sequence data to the banks	85
The storage and retrieval problem	86
Supplementary information	87
Sequence features	87
Sequence Symbols	87
Nucleic acid sequence symbols	87
Protein sequence symbols	88
Nucleic Acid Sequence Databases	88
The EMBL data library	91
The GenBank genetic sequence data bank	91
The CODATA recommendations	93
Line structure	93
Difficulties	101
Protein Sequence Databases	102
The NBRF-PIR database	102
The PSD-Kyoto database	103
The NEWAT database	103
The PGtrans and PseqIP collections	103
Other Databases Relevant to Molecular Biology	104
Macromolecular structure databases	104
Enzyme databases	104
Genetic map databases	105
Hybridoma databases	106
Cloning vector databases	106
Culture collection databases (including strains and cell lines)	106
Databases for taxonomy and identification	107
Molecular biology software databases	107
Using Molecular Sequence Databases	107
Organization of data	108
Retrieval of entries	109
Analysis programs	111
Molecular sequence databases as bibliography collections	112
References	112
5. ONLINE SERVICES	115
M.Ginsburg	
i.	
Introduction	115
Organized Computer Communications	116
Interactive computing and networks	118

System failures and networks	119
Establishing the Connection	119
Terminal emulation and file transfer	120
Computer facilities at ICRF Clare Hall laboratories	122
The Major Resources	123
BIONET	123
Cambridge	130
Edinburgh	134
GenBank	139
The Howard Hughes Medical Institute human gene mapping library	142
Protein identification resource	144
CITI2	144
References	145
Appendix	146
6. APPROACHES TO RESTRICTION MAP DETERMINATION	147
G.Zehetner, A.Frischauf and H.Lehrach	
Introduction	147
Mapping strategies	147
The Different Approaches to Restriction Map Determination	149
Indirect methods	149
Direct approach: partial mapping protocols	156
References	164
7. COMPUTER-AIDED ANALYSIS OF ONE-DIMENSIONAL RESTRICTION FRAGMENT GELS	165
J.K.Elder and E.M.Southern	
Introduction	165
Measurement of Electrophoretic Mobility	165
Sources of error	165
Traditional methods	165
Computer-based methods	166
Mobility measurement from digital profiles	167
Correction for inter-track variation	168
Relative accuracy of manual and computer-based methods for measuring mobility	168
Methods for Converting Mobility to Fragment Length	169
Manual and semi-logarithmic methods	169
Reciprocal method	169
Relative accuracy of semi-logarithmic and reciprocal methods	170
Effect of base composition on gel mobility	170
References	171
Appendix	171

8. COMPUTER HANDLING OF DNA SEQUENCING PROJECTS	173
R.Staden	
Introduction	173
The shotgun sequencing strategy	173
Definition of a contig	173
Symbols for uncertainty in gel readings	173
List of programs	174
Equipment	174
Introduction to the Computer Method	176
Objectives of a sequencing project	176
Required operations	177
Gel reading files and files of file names	178
Project databases	178
File and project names	180
How contigs are named	181
Searching for overlaps	181
Overview of computer processing	181
Entering Gel Readings into the Computer	182
Typing gel readings into the computer	182
Entering gel readings using a digitizer	183
Screening Gel Readings Against Vector Sequences	185
Format of vector and consensus sequence files	185
Running the vector screening program	186
Screening Gel Readings Against Restriction Enzyme Sites	186
Format of restriction enzyme recognition sequence files	186
Running the restriction enzyme screening program	186
Starting a New Project Database	187
Automatic Assembly of the Gel Readings	188
Running the automatic assembly program	191
Command Procedures for Screening and Assembling Gel Readings	192
Interactive Operations on a Project Database	194
Opening a project database	195
Displaying the aligned sequences in a contig	196
Examining the relational information	196
Editing gel readings in the database	198
Complementing contigs	200
Entering new gel readings into the database	201
Joining contigs	203
Calculating a consensus	204
Copying the database	206
Finding gel readings by name	206
Printing aligned sequences or relational information	207
Examining the quality of a contig	207
Checking the logical consistency of a database	209
Altering the relationships	210

Safeguarding databases	212
Highlighting Disagreements in Contigs	212
Preparing for the highlighting program	212
Using the highlighting program	213
Screen Editing of Contigs	213
Rules for screen editing	214
Using the screen editing procedure	215
Searching for Missed Overlaps	216
Steps used to search for possible missing overlaps	216
References	217
9. AUTOMATIC READING OF DNA SEQUENCING GEL AUTORADIOGRAPHS	219
J.K.Elder and E.M.Southern	
Introduction	219
Scanning and Digitization	220
Scan dimensions	220
Spatial resolution, amount of data and storage	220
Optical density range and resolution	220
Scanners	220
Analysis	221
Track boundary detection	221
Track straightening	224
Background subtraction	224
Estimation of band shapes and generation of track density profiles	224
Profile registration	226
Band detection	227
Reading the sequence	228
Time and cost	228
Acknowledgements	229
References	229
10. IDENTIFYING CODING SEQUENCES	231
G.D.Stormo	
Introduction	231
Search by Signal	232
Consensus sequences	232
Matrix evaluation of sequences	233
Alternative matrices	234
Search by Content	234
Open reading frames	235
Base/position preferences	235
Codon bias	237

Prokaryotic Coding Regions	238
Search by signal	238
Search by content	242
Combined methods	244
Eukaryotic Coding Regions	246
Search by signal	246
Search by content	251
Combined methods	255
Evaluation	256
Availability of the Programs Used	256
Conclusions	256
References	257
11. SECONDARY STRUCTURE PREDICTION OF RNA	259
M.Gouy	
Introduction	259
The Basics of Secondary Structure Modelling and Stability	
Computation	259
Elementary motifs of secondary structure models	259
Energy models for the computation of folding stabilities	260
Algorithms for the Prediction of Secondary Structure in RNA	
Molecules	266
Combinatorial algorithms	267
Recursive algorithms	269
Differences between combinatorial and recursive methods	271
Overview of two heuristic methods	272
Use of the Zuker–Stiegler Secondary Structure Prediction Program	272
Program distribution	273
Options	273
Example	274
Use of the Ninio Secondary Structure Prediction Program	274
Program distribution	274
Options	276
Example	277
Detection of Locally Stable Secondary Structures	277
Computer-aided Secondary Structure Model Building	278
Secondary Structure Plots and Drawings	281
Stüber's secondary structure drawing program	281
Shapiro's displays of secondary structures	281
An unconventional representation of secondary structures	282
Concluding Remarks	282
Acknowledgements	283
References	283
Note Added in Proof	284

12. PROTEIN STRUCTURE PREDICTION	285
W.R.Taylor	
The Protein Folding Problem	285
Theoretical importance	285
Practical importance	285
Protein Structure	286
Structure representations	287
Structural hierarchy	287
Protein topology	289
Domains	290
A Mechanistic Approach to Folding	291
Energy minimization	291
Future use of energy minimization	293
Empirical Methods of Structure Prediction	293
Secondary structure prediction	294
Hydrophobic plots	298
Pattern recognition methods	299
Accuracy of secondary structure prediction	302
Prediction of structural classes	305
Empirical Tertiary Structure Prediction	309
Secondary structure docking	309
Finger-print templates	310
Structure prediction by sequence homology	311
Structure building by direct sequence homology	316
Conclusions	317
A practical approach	317
The Availability of Programs	321
References	321
13. MOLECULAR SEQUENCE COMPARISON AND ALIGNMENT	323
J.F.Collins and A.F.W.Coulson	
Introduction	323
The 'Dot-Plot'	326
The simple dot-plot	326
Filtering by translation	329
Threshold filtering	329
Reduced alphabet	335
Similarity scoring	338
Exhaustive Alignment Algorithms	343
Introduction and definitions	343
Scoring schemes	345
The total alignment algorithm (Type I problem)	346

The best location algorithm (Type II problem)	346
The best local similarity algorithm (Type III problem)	348
Conclusions	349
Similarity Building Algorithms	350
Introduction	350
Practical considerations	352
Database Searching	353
Introduction	353
High speed filters	355
Application of the exhaustive algorithms	356
References	358
14. INFERENCE OF EVOLUTIONARY RELATIONSHIPS	359
M.J.Bishop, A.E.Friday and E.A.Thompson	
Introduction	359
Assessing sequence homology	359
Inferring an evolutionary relationship	360
Inferring the nature of an evolutionary process	360
Inferring the function constraints	360
The status of evolutionary inferences	360
Pathways of Genetic Transmission	361
The evolutionary tree model	361
More complex topologies	363
Processes of Genetic Sequence Change	363
Processes of genetic sequence change as observables	363
The need to consider generation time	365
Processes of genetic sequence change inferred from comparative studies	365
Methods of Inference for Evolutionary Problems	365
Methods employing heuristic approaches	365
Methods employing probabilistic models of genetic sequence change	366
A method of statistical inference	366
A Simple Model of DNA Sequence Divergence	367
The exponential failure model	368
Incorporating substitution, deletion and insertion	369
Pairwise Estimates of Evolutionary Parameters Under a Simple Model	371
Computing the likelihood	371
Worked example	373
Illustration of the method	375
Limitations of the model	375
In the absence of deletion/insertion an analytical solution is obtained	377
Pairwise divergence times lead to a heuristic phylogenetic tree	378

Joint Estimates of Phylogeny Under a Simple Model	379
Availability of Programs	383
References	383

GLOSSARY	387
-----------------	------------

INDEX	397
--------------	------------