

Contents

Part I Whetting Your Appetite

1	Introduction	3
1.1	A Very Short History of Bioinformatics and Computational Biology	3
1.2	Contemporary Biology	5
1.3	Computational Biology or Bioinformatics?	6
1.4	Computers in Biology	7
1.5	The Future: Digital Lab Benches and Designing Organisms	8
2	Content of This Book	11
2.1	The Main Chapters	11
2.1.1	Linux	11
2.1.2	Shell Programming	12
2.1.3	Sed	12
2.1.4	AWK	13
2.1.5	Perl	13
2.1.6	MySQL	14
2.1.7	R	14
2.1.8	Worked Examples	15
2.2	Prerequisites	15
2.3	Conventions	16
2.4	Additional Resources	17

Part II Computer and Operating Systems

3	Unix/Linux	21
3.1	What is a Computer?	21
3.2	Some History	22

3.2.1	Versions of Unix.	23
3.2.2	The Rise of Linux.	24
3.2.3	Why a Penguin?	26
3.2.4	Linux Distributions	26
3.3	X-Windows	26
3.3.1	How Does it Work?	27
3.4	Linux Architecture	28
3.5	What is the Difference Between Unix and Linux?	29
3.6	What is the Difference Between Linux and Windows?	29
3.7	What is the Difference Between Linux and Mac OS X?	29
3.8	One Computer, Two Operating Systems	30
3.8.1	VMware.	30
3.8.2	CygWin.	30
3.8.3	Wine	31
3.8.4	Others	31
3.9	Knoppix.	31
3.9.1	Vigyaan Knoppix for Biosciences	32
3.10	Software Running Under Linux	33
3.10.1	Bioscience Software for Linux	33
3.10.2	Office Software and Co.	34
3.10.3	Graphical Desktops	35
3.10.4	Others	35

Part III Working with Linux

4	The First Touch	39
4.1	Just Before the First Touch	39
4.1.1	Running Linux from USB or CDROM	39
4.1.2	Running Linux as a Virtual Machine.	40
4.2	Login	43
4.2.1	Working with CygWin or Mac OS X	44
4.2.2	Working Directly on a Linux Computer.	44
4.2.3	Working on a Remote Linux Computer	44
4.3	Using Commands	46
4.3.1	History and Autocompletion.	47
4.3.2	Syntax of Commands	47
4.3.3	Editing the Command Line	48
4.3.4	Change Password	49
4.3.5	Help: I'm Stuck	49
4.3.6	How to Find out More.	49
4.4	Logout.	50

5	Working with Files	53
5.1	Browsing Files and Directories	53
5.2	Moving, Copying and Renaming Files and Directories	56
5.3	File Attributes	56
5.4	Special Files: <code>.</code> and <code>..</code>	57
5.5	Protecting Files and Directories	59
5.5.1	Directories	59
5.5.2	Files	59
5.5.3	Examples	59
5.5.4	Changing File Attributes	60
5.5.5	Extended File Attributes	62
5.6	File Archives	62
5.7	File Compression	65
5.8	Searching for Files	66
5.8.1	Selecting Files for Backups	68
5.9	Display Disk Usage	68
6	Remote Connections	71
6.1	Downloading Data from the Web	71
6.1.1	<code>wget</code>	71
6.1.2	<code>curl</code>	72
6.2	Secure Copy: <code>scp</code>	72
6.3	Secure Shell: <code>ssh</code>	73
6.4	Backups with <code>rsync</code>	73
6.5	<code>ssh</code> , <code>scp</code> & <code>rsync</code> without Password	74
7	Playing with Text and Data Files	77
7.1	Viewing and Analyzing Files	78
7.1.1	A Quick Start: <code>cat</code>	78
7.1.2	Text Sorting	79
7.1.3	Extract Unique Lines	81
7.1.4	Viewing File Beginning or End	81
7.1.5	Scrolling Through Files	81
7.1.6	Character, Word, and Line Counting	82
7.1.7	Splitting Files into Pieces	82
7.1.8	Cut and Paste Columns	83
7.1.9	Finding Text: <code>grep</code>	84
7.1.10	Text File Comparisons	85
7.2	Editing Text Files	86
7.2.1	The Sparse Editor <code>Pico</code>	87
7.2.2	The Rich Editor <code>Vim</code>	88
7.2.3	Installing <code>Vim</code>	89
7.2.4	Immediate Takeoff	89
7.2.5	Starting <code>Vim</code>	89

7.2.6	Modes	90
7.2.7	Moving the Cursor	91
7.2.8	Doing Corrections	92
7.2.9	Save and Quit	93
7.2.10	Copy and Paste	93
7.2.11	Search and Replace	94
7.3	Text File Conversion (Unix ↔ DOS)	95
7.3.1	Batch Editing	95
8	Using the Shell	97
8.1	What is the Shell?	97
8.2	Different Shells	98
8.3	Setting the Default Shell	99
8.4	Useful Shortcuts	99
8.5	Redirections	101
8.6	Pipes	103
8.7	Lists of Commands	104
8.8	Aliases	104
8.9	Pimping the Prompt	105
8.10	Batch Jobs	106
8.11	Scheduling Commands	107
8.12	Wildcards	108
8.13	Processes	109
8.13.1	Checking Processes	110
8.13.2	Realtime Overview	111
8.13.3	Background Processes	111
8.13.4	Killing Processes	113
9	Installing BLAST and ClustalW	115
9.1	Downloading the Programs via FTP	115
9.1.1	Downloading BLAST	116
9.1.2	Downloading ClustalW	117
9.2	Installing BLAST	118
9.3	Running BLAST	119
9.4	Installing ClustalW	120
9.5	Running ClustalW	121
9.6	A Wrapper for ClustalW	122
10	Shell Programming	125
10.1	Script Files	125
10.2	Modifying the Path	127
10.3	Variables	128

10.4	Input and Output	130
10.4.1	echo	130
10.4.2	Here Documents: <<.....	131
10.4.3	read and line.....	132
10.4.4	Script Parameters	133
10.5	Substitutions and Expansions	134
10.5.1	Variable Substitution	134
10.5.2	Command Expansion.....	136
10.6	Quoting	136
10.6.1	Escape Character	137
10.6.2	Single Quotes	137
10.6.3	Double Quotes	137
10.7	Decisions: Flow Control	138
10.7.1	if...then...elif...else...fi.....	138
10.7.2	test	139
10.7.3	while...do...done	141
10.7.4	until...do...done	143
10.7.5	for...in...do...done	144
10.7.6	case...in...esac	145
10.7.7	select...in...do	146
10.8	Desktop Notifications	147
10.8.1	MacOSX	147
10.8.2	Linux.....	148
10.8.3	Windows	148
10.9	Debugging	148
10.9.1	bash -xv.....	148
10.9.2	trap	150
10.10	Remote Control Interactive Programs	150
10.11	Examples	152
10.11.1	Check for DNA as File Content	152
10.11.2	Time Signal	153
10.11.3	Select Files to Archive	153
10.11.4	Remove Spaces.....	155
11	Regular Expressions	157
11.1	Regular Expression and Neurons	157
11.2	Get Started	158
11.3	Using Regular Expressions.....	160
11.4	Search Pattern and Examples	161
11.4.1	Single-Character Meta Characters	162
11.4.2	Quantifiers	164
11.4.3	Grouping	165
11.4.4	Anchors.....	166
11.4.5	Escape Sequences	167

11.4.6	Alternation	167
11.4.7	Back References	168
11.4.8	Character Classes	169
11.4.9	Priorities	169
11.4.10	egrep Options	170
11.5	RE Summary	170
11.6	Regular Expression Training	172
11.6.1	Regular Expressions and vim.	172
11.7	Regular Expression in Genome Research.	172
12	Sed	175
12.1	When to Use Sed?	176
12.2	Getting Started	176
12.3	How Sed Works	178
12.3.1	Pattern Space	178
12.3.2	Hold Space.	179
12.4	Sed Syntax.	179
12.4.1	Addresses.	180
12.4.2	Sed and Regular Expressions	181
12.5	Commands.	181
12.5.1	Substitutions.	182
12.5.2	Transliterations	184
12.5.3	Deletions	185
12.5.4	Insertions and Changes	186
12.5.5	Sed Script Files	188
12.5.6	Printing	189
12.5.7	Reading and Writing Files	189
12.5.8	Advanced Sed.	190
12.6	Examples.	190
12.6.1	Gene Tree	191
12.6.2	File Tree	191
12.6.3	Reversing Line Order	192

Part IV Programming

13	AWK.	197
13.1	Getting Started	198
13.2	AWK's Syntax	199
13.3	Example File	200
13.4	Patterns	201
13.4.1	Regular Expressions	201
13.4.2	Pattern-Matching Expressions.	202
13.4.3	Relational Character Expressions	203

13.4.4	Relational Number Expressions	205
13.4.5	Mixing and Conversion of Numbers and Characters	206
13.4.6	Ranges	206
13.4.7	BEGIN and END	207
13.5	Variables	208
13.5.1	Assignment Operators	208
13.5.2	Increment and Decrement	209
13.5.3	Predefined Variables	210
13.5.4	Arrays	215
13.5.5	Shell Versus AWK Variables	219
13.6	Scripts and Executables	219
13.7	Decisions: Flow Control	220
13.7.1	if...else	220
13.7.2	while	221
13.7.3	do...while	222
13.7.4	for	223
13.7.5	Leaving Loops	224
13.8	Actions	225
13.8.1	Printing	225
13.8.2	Numerical Calculations	228
13.8.3	String Manipulation	229
13.8.4	System Commands	234
13.8.5	User-Defined Functions	234
13.9	Input with <code>getline</code>	238
13.9.1	Reading from a File	239
13.9.2	Reading from the Shell	241
13.10	Produce Nice Output with <code>L^AT_EX</code>	242
13.11	Examples	243
13.11.1	Sort an Array by Indices	243
13.11.2	Sum up Atom Positions	244
13.11.3	Convert FASTA ↔ Table Formats	244
13.11.4	Mutate DNA Sequences	245
13.11.5	Translate DNA to Protein	246
13.11.6	Calculate Atomic Composition of Proteins	247
13.11.7	Dynamic Programming: Levenshtein Distance of Sequences	250
14	Perl	255
14.1	Intention of This Chapter	255
14.2	Running Perl Scripts	256
14.3	Variables	256
14.3.1	Scalars	257
14.3.2	Arrays	258

14.3.3	Hashes.....	262
14.3.4	Built-in Variables.....	266
14.4	Decisions: Flow Control.....	266
14.4.1	if...elsif...else.....	267
14.4.2	unless, die and warn.....	267
14.4.3	while... ..	268
14.4.4	do...while... ..	268
14.4.5	until... ..	268
14.4.6	do...until... ..	268
14.4.7	for... ..	269
14.4.8	foreach... ..	269
14.4.9	Controlling Loops.....	269
14.5	Data Input.....	271
14.5.1	Command Line.....	271
14.5.2	Internal.....	272
14.5.3	Files.....	272
14.6	Data Output.....	274
14.6.1	print, printf and sprintf.....	274
14.6.2	Here Documents:<<.....	274
14.6.3	Files.....	275
14.7	Hash Databases.....	276
14.8	Regular Expressions.....	277
14.8.1	Special Escape Sequences.....	277
14.8.2	Matching: m/.../.....	278
14.9	String Manipulations.....	279
14.9.1	Substitute: s/.../.../.....	279
14.9.2	Transliterate: tr/.../.../.....	279
14.9.3	Common Commands.....	280
14.10	Calculations.....	281
14.11	Subroutines.....	282
14.12	Packages and Modules.....	283
14.13	Bioperl.....	285
14.14	You Want More?.....	286
14.15	Examples.....	286
14.15.1	Reverse Complement DNA.....	286
14.15.2	Calculate GC Content.....	287
14.15.3	Restriction Enzyme Digestion.....	288
15	Other Programming Languages.....	291

Part V Advanced Data Analysis

16	Relational Databases with MySQL	295
16.1	What is MySQL	295
16.1.1	Relational Databases	296
16.2	Administration	297
16.2.1	Get in Touch with the MySQL Server	297
16.2.2	Starting and Stopping the Server	298
16.2.3	Setting a Root Password	299
16.2.4	Set Up an User Account and Database	301
16.3	Utilization	303
16.3.1	Remote Access	303
16.3.2	MySQL Syntax	303
16.3.3	Creating Tables	304
16.3.4	Filling and Editing Tables	306
16.3.5	Querying Tables	309
16.3.6	Joining Tables	312
16.3.7	Access from the Shell	313
16.4	Backups and Transfers	314
16.5	How to Move on?	315
17	The Statistics Suite R	317
17.1	Getting Started	317
17.1.1	Getting Help	320
17.2	Reading and Writing Data Files	320
17.3	Data Structures	322
17.3.1	Vectors	322
17.3.2	Arrays and Matrices	322
17.3.3	Lists and Data Frames	324
17.4	Programming Structures	325
17.5	Data Exploration	327
17.5.1	Saving Graphic	330
17.5.2	Regression	332
17.5.3	t-Test	336
17.5.4	Chi-Square Test	337
17.6	R Scripts	338
17.7	Extensions: Installing and Using Packages	339
17.7.1	Connect to MySQL	339
17.7.2	Life-Science Packages	341

Part VI Worked Examples

18	Genomic Analysis of the Pathogenicity Factors from <i>E. coli</i> Strain O157:H7 and EHEC Strain O104:H4	345
18.1	The Project	345
18.2	Bioinformatic Tools and Resources	346
18.2.1	BLAST+	347
18.2.2	Genome Databases	347
18.3	Detailed Instructions	348
18.3.1	Downloading and Installing BLAST+	349
18.3.2	Downloading the Proteomes	351
18.3.3	Comparing the Genomes	353
18.3.4	Processing the BLAST+ Result File	357
18.3.5	Playing with the E-Value	360
18.3.6	Organize Results with MySQL	361
18.3.7	Extract Unique ORFs	366
18.3.8	Visualize and Analyze Results with R	367
19	Limits of BLAST and Homology Modeling	375
19.1	The Project	375
19.2	Bioinformatic Tools	376
19.2.1	genomicBLAST	376
19.2.2	ClustalW	376
19.2.3	Jpred	377
19.2.4	SWISS-MODEL	378
19.2.5	Jmol	378
19.3	Detailed Instructions	378
19.3.1	Download <i>E. coli</i> HybD Sequences	378
19.3.2	Cyanobacterial BLAST I	379
19.3.3	Cyanobacterial BLAST II	380
19.3.4	Compute Alignment and Tree Diagram	382
19.3.5	Secondary Structure Examination	384
19.3.6	Tertiary Structure Alignment (Homology Modeling)	386
19.3.7	Download and Visualize <i>E. coli</i> HybD Structure	387
19.3.8	View and Compare Cyanobacterial Structures	389
20	Virtual Sequencing of pUC18c	393
20.1	The Project	393
20.2	Bioinformatic Tools	393
20.2.1	Sequencer	393
20.2.2	Assembler	393
20.2.3	Dotter	394
20.2.4	GenBank	394

20.3	Detailed Instructions	394
20.3.1	Download the pUC18c Sequence	395
20.3.2	Download and Setup the Virtual Sequencer	395
20.3.3	Download and Install the TIGR Assembler	396
20.3.4	Download and Setup Dotter	397
20.3.5	Virtual Sequencing	399
20.3.6	Sequence Assembly	402
20.3.7	Testing Different Parameters	403
20.3.8	Visualizing Results with R	406
21	Querying for Potential Redox-Regulated Enzymes	409
21.1	The Project	409
21.2	Bioinformatic Tools and Resources	409
21.2.1	RCSB PDB	409
21.2.2	Jmol	410
21.2.3	Surface Racer	410
21.3	Detailed Instructions	411
21.3.1	Download Crystal Structure Files	411
21.3.2	Analyze Structures Visually	412
21.3.3	Analyze Structures Computationally	413
21.3.4	Expand the Analysis to Many Proteins	417
21.3.5	Automatically Test for Solvent Accessibility	419
	Appendix A: Supplementary Information	425
	References	433
	Solutions	437
	Index	443