

References

1. Abrahams, P.W., Larson, B.R.: UNIX for the Impatient, 2nd edn. Addison-Wesley (1996)
2. Adjero, D., Bell, T., Mukherjee, A.: The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching. Springer (2008)
3. Aho, A., Corasick, M.: Efficient string matching: an aid to bibliographic search. *Commun. ACM* **18**, 333–340 (1975)
4. Aho, A.V., Kernighan, B.W., Weinberger, P.J.: The AWK Programming Language. Addison-Wesley, Reading, MA (1988)
5. Aken, B.L., Achuthan, P., Akanni, W., Amode, M.R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Gil, L., Gim, C.G., Gordon, L., Hourlier, T., Hunt, S.E., Janacek, S.H., Juettemann, T., Keenan, S., Laird, M.R., Lavidas, I., Maurel, T., McLaren, W., Moore, B., Murphy, D.N., Nag, R., Newman, V., Nuhn, M., Ong, C.K., Parker, A., Patricio, M., Riat, H.S., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Wilder, S.P., Zadissa, A., Kostadima, M., Martin, F.J., Muffato, M., Perry, E., Ruffier, M., Staines, D.M., Trevanion, S.J., Cunningham, F., Yates, A., Zerbino, D.R., Flicek, P.: Ensembl 2017. *Nucl. Acids Res.* **45**(D1), D635 (2017)
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990)
7. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.: Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402 (1997)
8. Baeza-Yates, R.A., Perleberg, C.H.: Fast and practical approximate string matching. *Proceedings 3rd Symposium on Combinatorial Pattern Matching*, Springer Lecture Notes in Computer Science, vol. 644, pp. 185–192. Springer, New York (1992)
9. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995)
10. Börsch-Haubold, A.G., Montero, I., Konrad, K., Haubold, B.: Genome-wide quantitative analysis of histone H3 lysine 4 trimethylation in wild house mouse liver: environmental change causes epigenetic plasticity. *PlosOne* **9**, e97,568 (2014)
11. Buck, L., Axel, R.: A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. *Cell* **65**, 175–187 (1991)
12. Burrows, M., Wheeler, D.J.: A block-sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation, Palo Alto, California (1994)
13. Chang, J.T.: Recent common ancestors of all present-day individuals. *Adv. Appl. Probab.* **31**, 1002–1026 (1999)
14. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C.: A model of evolutionary change in proteins. In: M.O. Dayhoff (ed.) *Atlas of Protein Sequence and Structure*, vol. 5/suppl.3, pp. 345–352. National Biomedical Research Foundation, Washington DC (1978)

15. Delcher, A.L., Kasti, S., Fleischmann, R.D., Peterson, J., White, O., Salzberg, S.L.: Alignment of whole genomes. *Nucl. Acids Res.* **27**, 2369–2376 (1999)
16. Farris, J.S.: Estimating phylogenetic trees from distance matrices. *Am. Nat.* **106**, 645–668 (1972)
17. Felsenstein, J.: *Inferring Phylogenies*. Sinauer, Sunderland (2004)
18. Gau, J., Watabe, H., Aotsuka, T., Pang, J., Zhang, Y.: Molecular phylogeny of the *Drosophila obscura* species group, with emphasis on the old world species. *BMC Molecular. Evol. Biol.* **7**, 87 (2007)
19. Gibbs, A.J., McIntyre, G.A.: The diagram, a method for comparing sequences; its use with amino acid and nucleotide sequences. *Eur. J. Biochem.* **16**, 1–11 (1970)
20. Gupta, S.K., Kececioglu, J.D., Schäffer, A.A.: Improving the practical space and time efficiency of the shortest-path approach to sum-of-pairs multiple sequence alignment. *J. Comput. Biol.* **2**, 459–472 (1995)
21. Gusfield, D.: *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge (1997)
22. Haig, D., Hurst, L.D.: A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* **33**, 412–417 (1991)
23. Haubold, B., Klötzl, F., Pfaffelhuber, P.: `and.i`: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics* **31**, 1169–75 (2015)
24. Haubold, B., Wiehe, T.: *Introduction to Computational Biology: An Evolutionary Approach*. Birkhäuser, Basel (2006)
25. Hudson, R.R.: Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002)
26. Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. In: Munro, H.N. (ed.) *Mammalian Protein Metabolism*, vol. 3, pp. 21–132. Academic Press, New York (1969)
27. Kasai, T., Lee, G., Arimura, H., Arikawa, S., Park, K.: Linear-time longest-common-prefix computation in suffix arrays and its applications. *LNCS* **2089**, 181–192 (2001)
28. Knuth, D.E.: *The TeXbook*. Addison-Wesley, Reading, Massachusetts (1994)
29. Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.: Versatile and open software for comparing large genomes. *Genome Biol.* **5**(2), R12 (2004)
30. Lamport, L.: *A Document Preparation System: L^AT_EX*, 2nd edn. Addison-Wesley, Boston (1994)
31. Lee, H., Popodi, E., Tang, H., Foster, P.L.: Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. USA* **109**, E2774–83 (2012)
32. Lempel, A., Ziv, J.: On the complexity of finite sequences. *IEEE Trans. Inf. Theory* **IT-22**, 75–81 (1976)
33. Li, H., Durbin, R.: Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009)
34. Lipman, D.J., Altschul, S.F., Kececioglu, J.D.: A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA* **86**, 4412–4415 (1989)
35. Lynch, M.: *The Origins of Genome Architecture*. Sinauer, Sunderland (2007)
36. Manber, U., Myers, E.W.: Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.* **22**, 935–948 (1993)
37. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970)
38. Ohlebusch, E.: *Bioinformatics Algorithms: Sequence Analysis, Genome Rearrangements, and Phylogenetic Reconstruction*. Enno Ohlebusch, Ulm (2013)
39. Plank, L.D., Harvey, J.D.: Generation time statistics of *Escherichia coli* B measured by synchronous culture techniques. *J. Gen. Microbiol.* **115**, 69–77 (1979)
40. Rice, W.R.: Analyzing tables of statistical tests. *Evolution* **43**, 223–225 (1989)
41. Rohde, D.L.T., Olson, S., Chang, J.T.: Modelling the recent common ancestry of all living humans. *Nature* **431**, 562–566 (2004)

42. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987)
43. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., Petersen, G.B.: Nucleotide sequence of bacteriophage λ DNA. *J. Mol. Biol.* **162**, 729–773 (1982)
44. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981)
45. Student: The probable error of a mean. *Biometrika* **6**, 1–25 (1908)
46. Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M.: Phylogenetic inference. In: Hillis, D.M., Craig, M., Marble, B.K. (eds.) *Molecular Systematics*, 2nd edn, pp. 407–514. Sinauer, Sunderland (1996)
47. The Gene Ontology consortium: Gene ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000)
48. Ukkonen, E.: On-line construction of suffix-trees. *Algorithmica* **14**, 249–260 (1995)
49. Wakeley, J.: *Coalescent Theory: An Introduction*. Roberts & Company, Colorado (2009)
50. Watterson, G.A.: On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975)
51. Wikipedia: Observable universe (2016). http://en.wikipedia.org/wiki/Observable_universe

Index

Symbols

<, 9, 295
>, 9, 90, 297
>&, 50
>>, 176, 295
*, 2
&, 10
|, 297

A

ABC transporters, 92–94, 244, 245
Adh/Adh-dup, 36, 37, 41, 43, 72, 76, 97, 172, 174, 175, 177, 181
 age of duplication, 44–45, 186
Alcohol dehydrogenase, 36, 60
Alignment, 24
 and dot plot, 38, 179
 compare score schemes, 75
 dynamic programming, 40, 214
 fast global, 83–85, 228–234
 fast local, 72–78, 210–221
 gap score, 25, 40, 41, 180
 global, 23, 38–41, 43, 179, 180
 global/local, 72
 glocal, 207
 heuristic, 74, 83, 95
 k-error, 69–72, 206–209
 local, 38, 42, 179, 181–183
 match score, 40
 mismatch score, 40
 multiple, *see* separate entry
 number of alignments, 32–34, 168–171
 optimal, 38–42, 74, 179–183
 overlap, 79, 80, 221, 222
 pairwise, 23
 random, 43, 183, 184

 score, 25, 31, 43, 183, 184
 score scheme, 74
 ungapped, 73, 210
Alleles, 23
 α -hemoglobin, 96
Amino acids, 25–32
 conservation of pairs, 26
 distances, 27, 29
 frequencies, 32
 mismatched, 31, 161, 164
 most frequent, 31
 polarity, 27, 28
 side chains, 27, 28
Amino acids conservation of pairs, 31
Ancestors, 113
*and*i, 106, 108
apt, 8, 9
Array, *see* *awk*
awk, 14, 17–21, 48, 49, 60, 151–155, 302–306
 action block, 18
 actions, 302
 arithmetic functions, 304
 array, 19
 BEGIN, 21
 built-in variables, 303
 END, 19
 for loop, 19
 formatting for *printf*, 303
 hash, 21
 input/output, 303
 operators, 304, 305
 patterns, 18, 302
 print column, 17
 string manipulation, 304

B

bash, 2, 6, 14, 211, 297
 as calculator, 6
 autocompletion, 2, 3
 .bashrc, 11
 cursor moves, 5
 loops, 14, 15, 74, 297
 scripts, 14, 297
 text replacement, 298
bc, 235
Benjamini–Hochberg correction, 130
 β -hemoglobin, 96
Big-O notation, 47
Binary search, 75, 214, 215
BLAST, 72, 88, 90
 algorithm, 73
 all-against-all, 89, 238
 blastn mode, 75, 78, 215
 blastn-short mode, 86, 235
 database, 78
E-value, 77, 78, 220
 extension strategy, 214
 megablast mode, 75, 78, 215
 position-specific iterated, 93, 245–247
 reciprocal hits, 88
 run time, 235
 score, 76
 score distribution, 77, 218
 sensitivity, 75, 212, 213, 215, 216
 significance, 77, 220
 simple, 73–74
 tabular result, 239
 word list, 73
 word size, 75
blast2dot.awk, 89
blastn, 73, 75–77, 214, 216, 238
blastp, 89, 90, 248
Bonferroni correction, 130
Book website, vi
brew, 8, 10
Burrows–Wheeler transform, 86
bwa, 86, 87, 236, 237
BWT, 63, 66, 206
bwt, 63, 64, 206
gzip, 63, 67

C

cat, 8, 9, 31, 296
cchar, 35, 207
cd, 2, 3, 299
Chaperones, 239
chmod, 8, 13

clustal2, *see* **clustalw**
circo, 89, 92, 241, 243
clustalw, 95, 97, 98, 251
clustDist, 106, 107, 111
Coalescent, 121–126, 275–280
 algorithm, 124
 coalescence event, 121
 construction, 121, 123, 276
 mutations, 124, 277
 population size, 122
 print tree, 125
 sample size, 122
 segregating sites, 125
 simulations, 125, 126, 278
 time intervals, 122
 time to the most recent common ancestor,
 122, 276
coalescent.awk, 121, 125
Codd, E. F., 131
Coding sequence (CDS), 158, 178
Codon, 158
Command line, 2
Complexity, 66, 205
Compressibility, 63, 66
Contig, 81, 82, 224
cp, 296, 299
cut, 8, 9, 296
cutSeq, 38, 206, 209

D

Darwin, C., 101
Database client, 132
Database server, 132
Deletion, 24, 156
de novo sequencing, 86
D. guanche, 36
diff, 71, 149, 296
Directories, 7, 11, 290
Distance matrix, 96, 109, 261
Divergence time, 43, 45
D. melanogaster, 36, 71, 172
dnadist, 106
Dot plot, 34–38, 172–178
 and alignment, 38, 179
 duplication, 172
 repeat length, 36, 174
 size, 35
dotPlotFilter.awk, 35, 36
drawGenealogy, 113, 115
drawGenes, 8, 10
drawStreets, 84
drawWrightFisher, 113, 117

du, 63, 206
dvips, 53

E

echo, 6, 25
E. coli, 83, 85
 divergence time, 234
emacs, 8, 9, 293
Enhanced suffix array, 58
ENSEMBL, 131, 132, 137–138, 290–292
Entity-relation model, 132, 133
Exact matching problem, 47
Example data, 8

F

Factorial, 159
False discovery rate, 127, 130
False-negative rate, 127, 129, 130, 281, 282
False-positive rate, 127–129, 281
FASTA format, 21
fasta2tab.awk, 95, 99
File permissions, 7, 12
Files, 293
File system, 7, 299
find, 8, 11
Fisher, R. A., 115
fold, 48, 49, 188

G

gal, 23, 25, 41, 70, 72, 180, 206, 209
gd, 106, 107, 261
GenBank, 37
genCode, 26, 29, 160
Genealogy, 113, 115, 116, 270
 bi-parental, 113–115, 268–271
 uni-parental, 113, 115, 271
 universal ancestors, 115, 270
Gene duplication, 35, 88
Gene families, 88
Gene ontology, 127, 130, 131
Genetic code, 26–28
Genome assembly, 79, 81, 82, 84, 224, 226
genTree, 101, 104, 256
getSeq, 89, 96
gnuplot, 8, 12, 27, 34, 145, 211, 215
Gosset, W. S., 127
Graphical user interfaces, 1
Great apes, 107
grep, 8, 9, 16, 296, 300
gv, 48, 53
Gyrase, 17

gzip, 63, 66, 67

H

Hamlet, 64, 71
Hash, 14
head, 8, 9, 296
Hemoglobin, 99
histogram, 26, 27
history, 4
Hominidae, 107, 108, 260, 261, 263
Homology, 23, 36, 43, 72
Hyper-cube, 95

I

Insertion, 24, 156, 248
Interactive editor, *see* emacs
Inverse suffix array, 58

J

Java, 132, 136
java, 136
javac, 136
JDBC, 137
join, 296

K

kerror, 70–72, 74, 207, 209
k-error alignment, 70
keywordMatcher, 48, 51, 190–192, 206,
 207
Keyword tree, 48–53, 186–193
 failure links, 50–52, 192

L

lal, 42, 72, 74, 76, 207, 209, 211, 216
latex, 48, 53, 115
Lempel–Ziv decomposition, 63, 66
less, 38, 64, 71, 296
ln, 71
Longest repeat, 55, 56, 60, 84, 201
ls, 2, 3, 296
lscpu, 106, 108
lzd, 63

M

make, 8, 10
makeblastdb, 90
man, 2, 5

Match probability, 56
 Matrix multiplication, 26, 30
 Mean, 82, 221
 Median, 82, 221, 222
M. genitalium, 8, 9, 15, 19, 55, 79, 80, 82, 90, 145, 148, 151–153, 242
 Midpoint rooting, 109
 Mismatches per site, 44, 185
 Mitochondrial genomes, 108, 263
`mkdir`, 2, 3, 299
 Molecular clock, 44, 105, 260
 Monte Carlo test, 128, 280
 Mouse genome, 126, 279
 Mouse transcriptome, 127, 130, 132
 Move to front, *see* MTF
 mRNA, 158
 ms, 121, 125
 MTF, 63, 65, 66, 206
`mtf`, 63, 65
 Multi-threading, 108
 Multiple sequence alignment, 94–99, 247–253
 dynamic programming, 96
 gaps, 98, 99, 250
 guide tree, 97, 98, 249
 multidimensional matrices, 94
 optimal, 94, 95
 polymorphisms, 107, 261
 progressive, 95–99, 249–253
 query-anchored, 95–96, 247–249
 MUMmer, 83
`mummer`, 83–85, 230–231
 plot, 84, 85, 230, 231, 233
`mum2plot.awk`, 84, 225
 Mutation, 24, 156
`mutator`, 70, 73, 75, 206
 mv, 4, 296
`mysql`, 290

N

N_{50} , 82, 226, 227
`naiveMatcher`, 48, 50, 51, 188–191
 Naïve string matching, 47–49
`neato`, 89, 90, 92, 240, 241
`new2view`, 101, 102, 107, 253
 Non-synonymous mutation, 28, 31
`nucmer`, 84, 85
`numAl`, 34

O

`oal`, 79, 221
 Open reading frame, 27, 158

Orthology, 35, 36

P

`pamLog`, 26, 31
`pamNormalize`, 26, 31
`pamPower`, 26, 30, 31
 Paralogy, 35, 36, 43
 PATH, 7
 PATH, 11, 13
 Pattern preprocessing, 51
`percentDiff.awk`, 32
 PHYLLIP, 111, 267
 Phylogen
 midpoint rooting, 267
 Phylogeny, 101–111, 253–268
 branch lengths, 104, 105, 256
 four point criterion, 109, 264
 leaf labels, 103, 254
 midpoint rooting, 108, 266
 mutation rates, 109
 neighbor-joining, 108–111, 264–268
 Newick format, 101, 102, 111
 number of phylogenies, 105, 258, 259
 primates, 111, 267, 268
 radial layout, 255
 random, 104, 256
 root, 102, 254
 rooted, 106–108, 260–264
 three point criterion, 107, 109
 traversal, *see* tree
 ultrametric distances, 107–109, 260, 262
 unrooted, 103, 108–111, 264–268
 UPGMA, 106–108, 260–264
 Pipe (|), 5, 295
Plin5, 128
 Population size, 115
 POSIX standard, 2
 Prefix, 47, 52
 Primates, 108, 263
 Protein family, 88–90, 92, 240, 242, 245
 Protein sequences, 30, 31, 163
`psiblast`, 93, 94, 245, 247
`pwd`, 299

R

Random DNA sequence, 21, 80
 Random graph, 91
`randomizeSeq`, 35, 56, 67, 80, 173, 206, 223
 Random number generator, 105
`randot.awk`, 89, 91
`ranseq`, 52, 57, 80, 84, 223

- Read mapping, 86, 235
- Recursive function, 32–34, 168
 - bottom-up solution, 33
 - top-down solution, 33, 34
- Redirection, 297
- Regular expressions, 306, 307
- Relational databases, 131–138, 285–292
- repeater, 35, 57, 173, 196
- Re-sequencing, 86
- retree, 111, 267
- revComp, 48, 49, 188, 231
- rm, 2, 4, 296
- rmdir, 2–4, 296, 299
- rpois.awk, 121
- Run time, 48, 74

- S**
- SAM file, 87, 88
- Sampling without replacement, 123, 276
- samtools, 86–88
- Sanger, F., 78
- sblast, 73, 74, 76, 211, 213
- Scripts, 12
- sed, 14, 16, 17, 37, 175, 301–302
- seq, 14, 15, 297
- Sequence evolution, 23
- sequencer, 79, 81, 86, 224, 234
- Sequencing reads, 79
- Set matching, 48, 52
- Shell, *see* bash
- Shell script, *see* bash
- Shortest unique substring, 55, 61
- Shotgun sequencing, 78–82, 221–228
 - coverage, 80, 81, 223
 - error rate, 81, 225
 - paired-end, 81
 - unsequenced nucleotides, 81, 223
- show-snps, 84, 85
- shuffle.awk, 95
- shustring, 55, 61, 202
- simNorm, 127
- simOrf.awk, 27
- Single Nucleotide Polymorphisms (SNPs), 85, 126, 279
- Singleton proteins, 92, 243
- sort, 17, 149, 288
- source, 11, 146
- SQL, 132, 135–136, 286–289
 - avg, 135
 - count, 135, 287
 - delete, 134, 286
 - host language, 136
 - insert, 134, 135, 286
 - join, 135
 - limit, 135
 - max, 135
 - min, 135
 - select, 134, 286
- sqlite3, 134, 137, 285
- Statistical significance
 - effect size, and, 130
 - multiple experiments, 128–130, 280–285
 - sample size, 283
 - sample size, and, 130
 - single experiments, 128, 280
- Statistics
 - sample size, 283
- Stream editor, 14
- Student's t-test, 127
- Substitution matrices, 25–32, 158–167
 - BLOSUM62, 94
 - PAM, 30–32
- Substitution rate, 44, 185
- sudo, 10
- Suffix, 47
- Suffix array, 57–61, 64, 197–202
 - common prefix, 58, 197
 - enhanced, 59, 60
 - inverse, 60, 199
 - lcp array, 58, 60, 61, 198
 - lcp interval, 59, 60, 198
 - lcp interval tree, 58, 199
 - suffix tree, and, 59
- Suffix tree, 54–58, 60, 61, 83, 193–196, 198, 199
 - construction, 54, 57
 - generalized, 83, 84, 228
 - searching, 55
 - string depth, 84
- Symbolic links, 71
- Synonymous mutation, 28

- T**
- tabix, 126
- tac, 296
- tail, 8, 9, 296
- tar, 8
- testMeans, 127, 128
- Text compression, 62–67, 202–206
- The Origin of Species*, 101
- time, 38, 48, 50, 74, 108
- touch, 2, 3
- tr, 16, 300

traverseTree, 101, 104, 256

Tree

notation, 98, 102, 103

recursive structure, 101, 103

traversal, 101, 103, 104, 256

Tree:traversal, 104, 256

Type I error, 127, 129, 130, 281, 282

Type II error, 127, 130, 282, 283

U

Uniprot/Swissprot, 96, 99

uniq, 14, 20, 296

Universal ancestors, 115, 269

UNIX, 2, 7, 292–307

V

var, 20

Variance, 20

velveth/velvetg, 79, 81, 224

W

watterson, 121, 125

Watterson's equation, 121, 125

wc, 5, 160, 296

which, 8, 11

World population, 268

Wright–Fisher model, 113, 115–118, 120, 126, 271, 280

ancestor event, 119, 273

common ancestors, 117–119, 271–273

lineages per generation, 120, 273

most recent common ancestor, 119

simulations, 116, 117, 271, 272

test of, 126, 279

time to the most recent common ancestor, 120, 274, 275

Wright, S., 115