Bibliography

[1] A. Aho, J. Hopcroft, and J. Ullman. *Data Structures and Algorithms*. Addison-Wesley, Boston, 1983.

[2] A.V. Aho and M.J. Corasick. Efficient string matching: an aid to bibliographic search. *Communication of ACM*, 18:333–340, 1975.

[3] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson. *Molecular Biology of the Cell*. Garland Publishing, New York, 1994.

[4] S. Altschul, W. Gish, W. Miller, E. Myers, and J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[5] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and Psi-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[6] V.L. Arlazarov, E. A. Dinic, M. A. Kronrod, and I. A. Faradzev. On economical construction of the transitive closure of an oriented graph. *Soviet Math. Dokl.*, 11:1209–1210, 1970.

[7] P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, MA, 1997.

[8] P. Baldi, Y. Chauvin, T. Hunkapiller, and M. McClure. Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America*, 91:1059–1063, 1994.

[9] L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains. *Annals of Mathematical Satistics*, 41:164–171, 1970.

[10] A. Baxevanis and B.F. Ouellette. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Wiley-Interscience, Hoboken, NJ, 1998.

[11] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6:281–297, 1999.

[12] S.M. Berget, C. Moore, and P.A. Sharp. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74:3171–3175, 1977.

[13] P. Berman, S. Hannenhalli, and M. Karpinski. 1.375-approximation algorithm for sorting by reversals. In *European Symposium on Algorithms*, volume 2461 of *Lecture Notes in Computer Science*, pages 200–210, Rome, Italy, 2002. Springer-Verlag.

[14] M. Borodovsky and J. McIninch. Recognition of genes in DNA sequences with ambiguities. *BioSystems*, 30:161–171, 1993.

[15] P. Bourne and H. Weissig (eds). *Structural Bioinformatics*. Wiley–Liss, Hoboken, NJ, 2002.

[16] R.S. Boyer and J.S. Moore. A fast string searching algorithm. *Communication of ACM*, 20:762–772, 1977.

[17] T. Brown. *Genomes*. John Wiley and Sons, New York, 2002.

[18] J. Buhler and M. Tompa. Finding motifs using random projections. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB-01)*, pages 69–76, Montreal, Canada, April, 2001.

[19] P. Buneman. *The Recovery of Trees from Measures of Dissimilarity*. Edinburgh University Press, Edinburgh, 1971.

[20] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94, 1997.

[21] L.T. Chow, R.E. Gelinas, T.R. Broker, and R.J. Roberts. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell*, 12:1–8, 1977.

[22] G. Churchill. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, 51:79–94, 1989.

[23] S. A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the 3rd annual ACM Symposium on Theory of Computing*, pages 151–158, Shaker Heights, OH, 1971. ACM Press.

[24] T. H. Cormen, C. L. Leieserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, Cambridge MA, 2001.

[25] V. Dancik, T. A. Addona, K. R. Clauser, J. E. Vath, and P. A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6:327–342, 1999.

[26] K. J. Danna, G. H. Sack Jr., and D. Nathans. Studies of simian virus 40 DNA. VII. A cleavage map of the SV40 genome. *Journal of Molecular Biology*, 78:363–376, 1973.

[27] A. Dembo and S. Karlin. Strong limit theorem of empirical functions for large exceedances of partial sums of i.i.d. variables. *Annals of Probability*, 19:1737–1755, 1991.

[28] R. F. Doolittle, M. W. Hunkapiller, L. E. Hood, S. G. Devare, K. C. Robbins, S. A. Aaronson, and H. N. Antoniades. Simian sarcoma virus oncogene, ν-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science*, 221:275–277, 1983.

[29] R. Drmanac, I. Labat, I. Brukner, and R. Crkvenjakov. Sequencing of megabase plus DNA by hybridization: Theory of the method. *Genomics*, 4:114–128, 1989.

[30] A. Duarat, Y. Gerard, and M. Nivat. The chords problem. *Theoretical Computer Science*, 282:319–336, 2002.

[31] R. Durbin, S. Eddy, A. Krogh, and G. Mitchinson. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, England, 1998.

[32] D. Dussoix and W. Arber. Host specificity of infectious DNA from bacteriophage lambda. *Journal of Molecular Biology*, 11:238–246, 1965.

[33] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95:14863–14868, 1998.

[34] J. K. Eng, A. L. McCormack, and J. R. Yates (III). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am. Soc. Mass Spectrom.*, 5:976–989, 1995.

[35] E. Eskin and P. A. Pevzner. Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 18:S354–363, 2002.

[36] D. Feng and R. F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, 60:351–360, 1987.

[37] W. M. Fitch. Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20:406–416, 1971.

[38] S.P.A. Fodor, J.L. Read, M.S. Pirrung, L. Stryer, A.T. Lu, and D. Solas. Light-directed spatially addressable parallel chemical synthesis. *Science*, 251:767–773, 1991.

[39] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. W. H. Freeman and Co., 1979.

[40] W. Gates and C. Papadimitriou. Bounds for sorting by prefix reversals. *Discrete Mathematics*, 27:45–57, 1979.

[41] M.S. Gelfand, A.A. Mironov, and P.A. Pevzner. Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 93:9061–9066, 1996.

[42] O. Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982.

[43] M. Gribskov, M. McLachlan, and D. Eisenberg. Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 84:4355–4358, 1987.

[44] D. Gusfield. *Algorithms on Strings, Trees, and Sequences. Computer Science and Computational Biology*. Cambridge University Press, Cambridge, England, 1997.

[45] D. Haussler, A. Krogh, I. S. Mian, and K. Sjölander. Protein modeling using hidden Markov models: Analysis of globins. In *Proceedings of the Hawaii International Conference on System Sciences*, pages 792–802, Los Alamitos, CA, 1993. IEEE Computer Society Press.

[46] G. Z. Hertz, G. W. Hartzell 3rd, and G. D. Stormo. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Computer Applications in Bioscience*, 6:81–92, 1990.

[47] G.Z. Hertz and G.D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563–577, 1999.

[48] D.G. Higgins, J.D. Thompson, and T.J. Gibson. Using CLUSTAL for multiple sequence alignments. *Methods in Enzymology*, 266:383–402, 1996.

[49] D. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Communication of ACM*, 18:341–343, 1975.

[50] C. A. R. Hoare. Quicksort. *Computer Journal*, 5:10–15, 1962.

[51] S. Hopper, R. S. Johnson, J. E. Vath, and K. Biemann. Glutaredoxin from rabbit bone marrow. purification, characterization, and amino acid sequence determined by tandem mass spectrometry. *Journal of Biological Chemistry*, 264:20438–20447, 1989.

[52] S. Karlin and S.F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the United States of America*, 87:2264–2268, 1990.

[53] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103, Yorktown Heights, NY, 1972. Plenum Press.

[54] R.M. Karp and M.O. Rabin. Efficient randomized pattern-matching algorithms. *IBM Journal of Research and Development*, 31:249–260, 1987.

[55] J. Kececioglu and E.W. Myers. Combinatorial algorithms for DNA sequence assembly. *Algorithmica*, 13:7–51, 1995.

[56] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two permutation. *Algorithmica*, 13:180–210, 1995.

[57] D. E. Knuth. *The Art of Computer Programming*. Addison-Wesley, 1998.

[58] D.E. Knuth, J.H. Morris, and V.R. Pratt. Fast pattern matching in strings. *SIAM Journal on Computing*, 6:323–350, 1977.

[59] A. Krogh, M. Brown, I.S. Mian, K. Sjölander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.

[60] S. Kurtz, J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. Reputer: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, 29:4633–4642, 2001.

[61] G.M. Landau and U. Vishkin. Efficient string matching in the presence of errors. In *26th Annual Symposium on Foundations of Computer Science*, pages 126–136, Los Angeles, October 1985.

[62] E. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[63] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, and J.C. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[64] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10:707–710, 1966.

[65] L. Levin. Universal sorting problems. *Problems of Information Transmission*, 9:265–266, 1973.

[66] B. Lewin. *Genes VII*. Oxford University Press, Oxford, UK, 1999.

[67] D.J. Lipman and W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.

[68] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[69] Y. Lysov, V. Florent'ev, A. Khorlin, K. Khrapko, V. Shik, and A. Mirzabekov. DNA sequencing by hybridization with oligonucleotides. *Doklady Academy Nauk USSR*, 303:1508–1511, 1988.

[70] J. MacQueen. On convergence of k-means and partitions with minimum average variance. *Annals of Mathematical Statistics*, 36:1084, 1965.

[71] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66:4390–4399, 1994.

[72] L. Marsan and M. F. Sagot. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *Journal of Computational Biology*, 7:345–362, 2000.

[73] W. J. Masek and M. S. Paterson. A faster algorithm computing string edit distances. *Journal of Computational System Science*, 20:18–31, 1980.

[74] A. M. Maxam and W. Gilbert. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74:560–564, 1977.

[75] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[76] D. Mount. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Press, Cold Spring Harbor, NY, 2001.

[77] E.W. Myers and W. Miller. Optimal alignments in linear space. *Computer Applications in Biosciences*, 4:11–17, 1988.

[78] J. Nadeau and B. Taylor. Lengths of chromosome segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of America*, 81:814–818, 1984.

[79] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[80] S. J. O'Brien. *Tears of the Cheetah*. Thomas Dunne Books, New York, 2003.

[81] S. J. O'Brien, W. G. Nash, D. E. Wildt, M. E. Bush, and R. E. Benveniste. A molecular solution to the riddle of the giant panda's phylogeny. *Nature*, 317:140–144, 1985.

[82] H. Peltola, H. Soderlund, and E. Ukkonen. SEQAID: A DNA sequence assembling program based on a mathematical model. *Nucleic Acids Research*, 12:307–321, 1984.

[83] P.A. Pevzner. *l*-Tuple DNA sequencing: computer analysis. *Journal of Biomolecular Structure and Dynamics*, 7:63–73, 1989.

[84] P.A. Pevzner, V. Dancik, and C.L. Tang. Mutation-tolerant protein identification by mass-spectrometry. *Journal of Computational Biology*, 7:777–787, 2000.

[85] P.A. Pevzner and G. Tesler. Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes. *Genome Research*, 13:37–45, 2003.

[86] P.A. Pevzner and M.S. Waterman. Multiple filtration and approximate pattern matching. *Algorithmica*, 13:135–154, 1995.

[87] J.C. Roach, C. Boysen, K. Wang, and L. Hood. Pairwise end sequencing: A unified approach to genomic mapping and sequencing. *Genomics*, 26:345–353, 1995.

[88] D. F. Robinson. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory (B)*, 11:105–119, 1971.

[89] J. R. Sadler, M. S. Waterman, and T. F. Smith. Regulatory pattern identification in nucleic acid sequences. *Nucleic Acids Research*, 11:2221–2231, 1983.

[90] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biological Evolution*, 4:406–425, 1987.

[91] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74:5463–5467, 1977.

[92] D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics*, 28:35–42, 1975.

[93] D. Sankoff. Edit distances for genome comparisons based on non-local operations. In *Third Annual Symposium on Combinatorial Pattern Matching*, volume 644 of *Lecture Notes in Computer Science*, pages 121–135, Tucson, AZ, 1992. Springer-Verlag.

[94] S.S. Skiena, W.D. Smith, and P. Lemke. Reconstructing sets from interpoint distances. In *Proceedings of Sixth Annual Symposium on Computational Geometry*, pages 332–339, Berkeley, CA, June, 1990.

[95] H.O. Smith and K.W. Wilcox. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *Journal of Molecular Biology*, 51:379–391, 1970.

[96] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[97] E.E. Snyder and G.D. Stormo. Identification of protein coding regions in genomic DNA. *Journal of Molecular Biology*, 248:1–18, 1995.

[98] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.

[99] E.L. Sonnhammer, S.R. Eddy, and R. Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28:405–420, 1997.

[100] E. Southern. United Kingdom patent application GB8810400. 1988.

[101] G. Stormo, T. Schneider, L. Gold, and A. Ehrenfeucht. Use of the perceptron algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*, 10:2997–3011, 1982.

[102] A. H. Sturtevant and T. Dobzhansky. Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species. *Proceedings of the National Academy of Sciences of the United States of America*, 22:448–450, 1936.

[103] A. R. Templeton. Out of Africa again and again. *Nature*, 416:45–51, 2002.

[104] J.C. Venter et al. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[105] T.K. Vintsyuk. Speech discrimination by dynamic programming. *Kibernetika*, 4:52–57, 1968.

[106] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, 1967.

[107] M. S. Waterman. *Skiing the Sun*. (unpublished manuscript), 2004.

[108] M.S. Waterman. *Introduction to Computational Biology*. Chapman Hall, New York, 1995.

[109] M.S. Waterman and M. Eggert. A new algorithm for best subsequence alignments with application to tRNA–rRNA comparisons. *Journal of Molecular Biology*, 197:723–728, 1987.

[110] J. Weber and G. Myers. Whole genome shotgun sequencing. *Genome Research*, 7:401–409, 1997.

[111] P. Weiner. Linear pattern matching algorithms. In *Proceedings of the 14th IEEE Symposium on Switching and Automata Theory*, pages 1–11, University of Iowa, October 1973.

[112] J. R. Yates III, J. K. Eng, A. L. McCormack, and D. Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Analytical Chemistry*, 67:1426–1436, 1995.

[113] K. Zaretskii. Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspekhi Mat. Nauk.*, 20:90–92, 1965.

[114] Z. Zhang. An exponential example for a partial digest mapping algorithm. *Journal of Computational Biology*, 1:235–239, 1994.