# References


1. G. R. Abecasis, L. R. Cardon, and W. O. Cookson. A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics*, 66:279–292, 2000.
2. F. Achard, G. Vaysseix, and E. Barillot. XML, bioinformatics and data integration. *Bioinformatics*, 17:115–125, 2001.
3. R. M. Adams, B. Stancampiano, M. McKenna, and D. Small. Case study: a virtual environment for genomic data visualization. In *Proceedings of IEEE Visualization Conference*, pages 513–516, Boston, MA, 2002.
4. D. A. Agard. Optical sectioning microscopy: cellular architecture in three dimensions. *Annu. Rev. Biophys. Bioeng.*, 13:191–219, 1984.
5. R. Agrawal, H. Mannila, R. Srikant, H. T. T. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), *Advances in Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, pages 307–328, 1996.
6. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of 1994 Int. Conf. Very Large Data Bases*, pages 487–499, Santiago, Chile, 1994.
7. A. V. Aho, Y. Sagiv, T. G. Szymanski, and J. D. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing*, 10:405–421, 1981.
8. R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ, 1993.
9. T. Akutsu, S. Miyano, and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16(8):727–734, 2000.
10. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, 2002.
11. D. O. V. Alonso, E. Alm, and V. Daggett. The unfolding pathway of the cell cycle protein p13suc1: implications for domain swapping. *Structure*, 8(1):101–110, 2000.
12. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
13. S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
14. A. An and Y. Wang. Comparisons of classification methods for screening potential compounds. In *Proceedings of IEEE International Conference on Data Mining*, 2001.
15. T. A. Andrea and H. Kalayeh. Applications of neural networks in quantitative structure-activity relationships of dihydrofolate reductase inhibitors. *Journal of Medicinal Chemistry*, 34:2824–2836, 1991.

16. C. Anfinsen and H. Scheraga. Experimental and theoretical aspects of protein folding. *Advances in Protein Chemistry*, 29:205–300, 1975.
17. M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. In *Proceedings of ACM SIGMOD Int. Conf. Management of Data*, pages 49–60, Philadelphia, PA, 1999.
18. E. Arjas. Survival models and martingale dynamics. *Scandinavian Journal of Statistics*, 16:177–225, 1989.
19. A. Arkin, J. Ross, and H. A. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage-lambda infected *Escherichia coli* cells. *Genetics*, 149:1633–1648, 1998.
20. K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3D point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):698–700, 1987.
21. M. M. Astrahan, M. W. Blasgen, D. D. Chamberlin, K. P. Eswaran, J. Gray, P. P. Griffiths, W. F. King III, R. A. Lorie, P. R. McJones, J. W. Mehl, G. R. Putzolu, I. L. Traiger, B. W. Wade, and V. Watson. System R: relational approach to database management. *ACM Transactions on Database Systems*, 1(2):97–137, 1976.
22. J. C. Avise and G. C. Johns. Proposal for a standardized temporal scheme of biological nomenclature. *Proceedings of the National Academy of Sciences of the USA*, 96:7358–7363, 1999.
23. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, Reading, MA, 1999.
24. V. Bafna, S. Muthukrishnan, and R. Ravi. Comparing similarity between RNA strings. In *Proceedings of Combinatorial Pattern Matching Conference*, pages 1–14, 1995.
25. S. Bain, J. Todd, and A. Barnett. The British Diabetic Association—Warren Repository. *Autoimmunity*, 7:83–85, 1990.
26. D. Baker. A surprising simplicity to protein folding. *Nature*, 405:39–42, 2000.
27. P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach* (2nd ed.). MIT Press, Cambridge, MA, 2001.
28. P. Baldi and G. W. Hatfield. *DNA Microarrays and Gene Expression*. Cambridge University Press, New York, 2002.
29. R. Ballew, T. Duncan, and M. Blasingame. Relational data structures for implementing thesauri. Museum Informatics Project, Information Systems and Technology, University of California, Berkeley, 1999.
30. B. R. Baum. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 41:3–10, 1992.
31. S. Batzoglou, L. Pachter, J. P. Mesirov, B. Berger, and E. S. Lander. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research*, 10:950–958, 2000.
32. A. Baxevanis and D. Davison, editors. *Current Protocols in Bioinformatics*. John Wiley, New York, 2002.
33. A. Baxevanis and B. F. F. Ouellette. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (2nd ed.). John Wiley, New York, 2001.
34. N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: an efficient and robust access method for points and rectangles. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 322–331, Atlantic City, NJ, 1990.
35. R. Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6):284, 1961.
36. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and B. A. Rapp. GenBank. *Nucleic Acids Research*, 28:15–18, 2000.

37. W. G. Berendsohn. The concept of "potential taxa" in databases. *Taxon*, 44:207–212, 1995.
38. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.
39. J. Bernardo and A. Smith. *Bayesian Theory*. John Wiley, New York, 1994.
40. M. R. Berthold and C. Borgelt. Mining molecular fragments: finding relevant substructures of molecules. In *Proceedings of IEEE International Conference on Data Mining*, 2002.
41. O. Bininda-Emonds, J. L. Gittleman, and M. A. Steel. The (super)tree of life. *Annual Review of Ecology and Systematics*, 33:265–289, 2002.
42. O. Bininda-Emonds and M. Sanderson. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology*, 50:565–579, 2001.
43. M. V. Boland, M. K. Markey, and R. F. Murphy. Classification of protein localization patterns obtained via fluorescence light microscopy. In *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 594–597, 1997.
44. M. V. Boland, M. K. Markey, and R. F. Murphy. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. *Cytometry*, 33:366–375, 1998.
45. M. V. Boland and R. F. Murphy. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells. *Bioinformatics*, 17:1213–1223, 2001.
46. N. B. Booth and A. F. M. Smith. A Bayesian approach to retrospective identification of change-points. *J. Econometr.*, 19:7–22, 1992.
47. M. Borodovsky and J. McIninch. GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem.*, 17:123–133, 1993.
48. J. Bower and H. Bolouri. *Computational Modeling of Genetics and Biochemical Networks*. MIT Press, Cambridge, MA, 2001.
49. N. Bray, I. Dubchak, and L. Pachter. AVID: a global alignment program. *Genome Research*, 13(1):97–102, 2003.
50. A. Brazma, et al. Minimum information about a microarray experiment (miame)—toward standards for microarray data. *Nature Genetics*, 29:365–371, 2001.
51. S. P. Brooks. Markov chain Monte Carlo and its applications. *Statistician*, 47:69–100, 1998.
52. S. P. Brooks and P. Giudici. Diagnosing convergence of reversible jump MCMC algorithms. In J. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith (eds.), *Bayesian Statistics* 6. Oxford University Press, Oxford, pages 733–742, 1999.
53. J. W. Brown. The ribonuclease P database. *Nucleic Acids Research*, 27:314, 1999.
54. M. Brudno, C. Do, G. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow, and S. Batzoglou. Lagan and multi-lagan: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research*, 13(4):721–731, 2003.
55. S. Bryant. Evaluation of threading specificity and accuracy. *Proteins*, 26(2):172–185, 1996.
56. D. Bryant and M. Steel. Extension operations on sets of leaf-labeled trees. *Advances in Applied Mathematics*, 16:425–453, 1995.
57. C. B. Burge and S. Karlin. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.*, 8:346–354, 1998.
58. C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–168, 1998.

59. S. Burkhardt, A. Crauser, P. Ferragina, H.-P. Lenhof, E. Rivals, and M. Vingron. q-gram based database searching using a suffix array (QUASAR). In *Proceedings of International Conference on Research in Computational Molecular Biology*, pages 77–83, Lyon, 1999.
60. A. Califano and I. Rigoutsos. FLASH: fast look-up algorithm for string homology. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, pages 56–64, 1993.
61. O. Camoglu, T. Kahveci, and A. K. Singh. Towards index-based similarity search for protein structure databases. In *Proceedings of IEEE Computer Society Bioinformatics Conference*, pages 148–158, 2003.
62. D. Cantone, G. Cincotti, A. Ferro, and A. Pulvirenti. An efficient algorithm for the 1-median problem. Technical Report, University of Catania, submitted 2003.
63. D. Cantone, A. Ferro, A. Pulvirenti, D. Reforgiato, and D. Shasha. Antipole indexing to support range search and $k$-nearest neighbor metric spaces. Technical Report, University of Catania, submitted 2003.
64. M. J. Carey, D. J. DeWitt, M. J. Franklin, N. E. Hall, M. L. McAuliffe, J. F. Naughton, D. T. Schuh, M. H. Solomon, C. K. Tan, O. G. Tsatalos, S. J. White, and M. J. Zwilling. Shoring up persistent applications. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, Minneapolis, MN, 1994.
65. H. Carrillo and D. Lipmann. The multiple sequence alignment problem in biology. *SIAM Journal on Applied Mathematics*, 48:1073–1082, 1988.
66. K. R. Castleman. *Digital Image Processing*. Prentice Hall, Upper Saddle River, NJ, 1996.
67. T. Cech and B. Bass. Biological catalysis by RNA. *Annual Review of Biochemistry*, 55:599–629, 1988.
68. J. Celko. *SQL for Smarties: Advanced SQL Programming*. Morgan Kaufmann, San Francisco, 1995.
69. E. Chavez, G. Navarro, R. Baeza-Yates, and J. L. Marroquin. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, 2001.
70. D. Chen, O. Eulenstein, D. Fernández-Baca, and M. Sanderson. Supertrees by flipping. Technical Report TR02-01, Department of Computer Science, Iowa State University, 2001.
71. S. Chen, Z. Wang, and K. Zhang. Pattern matching and local alignment for RNA structures. In *Proceedings of International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences*, pages 55–61, Las Vegas, Nevada, 2002.
72. X. Chen, M. Velliste, S. Weinstein, J. W. Jarvik, and R. F. Murphy. Location proteomics—building subcellular location trees from high resolution 3D fluorescence microscope images of randomly tagged proteins. In *Proceedings of the SPIE (International Society for Optical Engineering)*, pages 298–306, 2003.
73. Y. Cheng and G. Church. Biclustering of expression data. In *Proceedings of Int. Conf. on Intelligent Systems for Molecular Biology*, La Jolla, CA, 2000.
74. G. Chikenji and M. Kikuchi. What is the role of non-native intermediates of beta-lactoglobulin in protein folding? *Proceedings of the National Academy of Sciences of the USA*, 97:14273–14277, 2000.
75. K. Chou and Y. Cai. Using function domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry*, 277:45765–45769, 2002.

76. J. Clarke, E. Cota, S. B. Fowler, and S. J. Hamill. Folding studies of immunoglobulin-like $\beta$-sandwich proteins suggest that they share a common folding pathway. *Structure*, 7:1145–1153, 1999.
77. C. Clementi, P. A. Jennings, and J. N. Onuchic. How native-state topology affects the folding of dihydrofolate reductase and interleukin-1beta. *Proceedings of the National Academy of Sciences of the USA*, 97(11):5871–5876, 2000.
78. M. Cline, G. Liu, A. E. Loraine, J. Cheng, R. Shigeta, G. Mei, D. Kulp, and M. A. Siani-Rose. Structure-based comparison of four eukaryotic genomes. In *Proceedings of Pacific Symposium on Biocomputing*, pages 127–138, 2002.
79. G. Collins, S. Y. Le, and K. Zhang. A new method for computing similarity between RNA structures. In *Proceedings of the 2nd International workshop on Biomolecular Informatics*, pages 761–765, Atlantic City, NJ, 2000.
80. R. Collobert, S. Bengio, and J. Mariéthoz. Torch: a modular machine learning software library. In *IDIAP (Dalle Molle Institute for Perceptual Artificial Intelligence) Research Report*, pages 2–46, 2002.
81. W. Colon and H. Roder. Kinetic intermediates in the formation of the cytochrome c molten globule. *Nature Structural Biology*, 3(12):1019–1025, 1996.
82. L. Conte, S. Brenner, T. Hubbard, C. Chothia, and A. Murzin. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Research*, 30:264–267, 2002.
83. D. J. Cook and L. B. Holder. Graph-based data mining. *IEEE Intelligent Systems*, 15(2):32–41, 2000.
84. A. Cornish-Bowden. *Fundamentals of Enzyme Kinetics*. Portland Press, London, 1996.
85. F. Corpet and B. Michot. RNAlign program: alignment of RNA sequences using both primary and secondary structures. *Comput. Appl. Biosci.*, 10(4):389–399, 1995.
86. J. Cracraft. The seven great questions of systematic biology: an essential foundation for conservation and the sustainable use of biodiversity. *Annals of the Missouri Botanic Garden*, 89:127–144, 2002.
87. M. Crochemore, G. M. Landau, and M. Ziv-Ukelson. A sub-quadratic sequence alignment algorithm for unrestricted scoring matrices. In *Proceedings of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 679–688, 2002.
88. M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.
89. A. Danckaert, E. Gonzalez-Couto, L. Bollondi, N. Thompson, and B. Hayes. Automated recognition of intracellular organelles in confocal microscope images. *Traffic*, 3:66–73, 2002.
90. T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley, New York, 2003.
91. T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk. Mining database structure; or how to build a data quality browser. In *Proceedings of ACM SIGMOD Int. Conf. on Management of Data*, pages 240–251, Madison, WI, 2002.
92. W. H. E. Day. Analysis of quartet dissimilarity measures between undirected phylogenetic trees. *Systematic Zoology*, 35:325–333, 1986.
93. M. L. de Buyser, A. Morvan, S. Aubert, F. Dilasser, and N. El Solh. Evaluation of a ribosomal RNA gene probe for the identification of species and subspecies within the genus *Staphylococcus*. *J. Gen. Microbiol.*, 138:889–899, 1992.

94. L. Dehaspe, H. Toivonen, and R. D. King. Finding frequent substructures in chemical compounds. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 30–36, 1998.
95. H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1):67–103, 2002.
96. K. de Queiroz and J. Gauthier. Phylogenetic taxonomy. *Annual Review of Ecology and Systematics*, 23:449–480, 1992.
97. A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg. Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369–2376, 1999.
98. M. Deshpande and G. Karypis. Automated approaches for classifying structure. In *Proceedings of the 2nd ACM SIGKDD Workshop on Data Mining in Bioinformatics*, 2002.
99. M. Deshpande and G. Karypis. Using conjunction of attribute values for classification. In *Proceedings of the 11th ACM Conference of Information and Knowledge Management*, pages 356–364, 2002.
100. B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29:311–322, 1995.
101. B. Devlin, N. Risch, and K. Roeder. Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics*, 36:1–16, 1996.
102. P. W. Diaconis and S. P. Holmes. Matchings and phylogenetic trees. *Proceedings of the National Academy of Sciences of the USA*, 95:14600–14602, 1998.
103. T. G. Dietterich. Ensemble methods in machine learning. In *Proceedings of the 1st International Workshop on Multiple Classifier Systems*, pages 1–15, Cagliari, Italy, 2000.
104. J. Ding, D. Berleant, D. Nettleton, and E. Wurtele. Mining MEDLINE: abstracts, sentences, or phrases. In *Proceedings of Pacific Symposium on Biocomputing*, pages 326–337, 2002.
105. A. Drawid and M. Gerstein. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *Journal of Molecular Biology*, 301:1059–1075, 2000.
106. Dtp aids antiviral screen dataset. http://dtp.nci.nih.gov.
107. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*, 2nd ed. John Wiley, New York, 2000.
108. B. Dunkel and N. Soparkar. Data organization and access for efficient data mining. In *Proceedings of the 15th IEEE International Conference on Data Engineering*, pages 522–529, 1999.
109. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probability Models of Proteins and Nucleic Acids*. Cambridge University Press, New York, 1998.
110. R. Edwards and L. Glass. Combinatorial explosion in model gene networks. *Chaos*, 10(3):691–704, 2000.
111. M. Eerola, H. Mannila, and M. Salmenkivi. Frailty factors and time-dependent hazards in modeling ear infections in children using Bassist. In *Proceedings of the XIII Symposium on Computational Statistics*, pages 287–292, 1998.
112. I. Eidhammer and I. Jonassen. Protein structure comparison and structure patterns—an algorithmic approach. Tutorial at the 9th International Conference on Intelligent Systems for Molecular Biology, 2001.
113. M. B. Eisen, P. T. Spellman, and P. O. Brown. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the USA*, 95:14863, 1998.

114. P. A. Evans. *Algorithms and Complexity for Annotated Sequence Analysis.* PhD thesis, University of Victoria, 1999.
115. W. J. Ewens and G. R. Grant. *Statistical Methods in Bioinformatics: An Introduction.* Springer-Verlag, New York, 2001.
116. R. Fagin. Combining fuzzy information from multiple systems. In *Proceedings of the 15th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 216–226, Montreal, Canada, 1996.
117. R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. In *Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, Santa Barbara, CA, 2001.
118. C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time series databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 419–429, Minneapolis, MN, 1994.
119. M. Farach, T. M. Przytycka, and M. Thorup. On the agreement of many trees. *Information Processing Letters*, 55:297–301, 1995.
120. P. L. Farber. *Finding Order in Nature: The Naturalist Tradition from Linnaeus to E. O. Wilson.* Johns Hopkins University Press, Baltimore, MD, 2000.
121. U. Fayyad, G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discovery.* Morgan Kaufmann, San Francisco, 2001.
122. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: an overview. In *Proceedings of Int. Conf. Knowledge Discovery and Data Mining*, Portland, Oregon, 1996.
123. J. Felsenstein. Phylip—phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989. http://evolution.genetics.washington.edu/phylip.html.
124. J. Felsenstein. Phylogenies from molecular sequences: inferences and reliability. *Annual Reviews of Genetics*, 22:521–565, 1998.
125. D. Feng and R. F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, 60:351–360, 1987.
126. W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.
127. S. T. Fitz-Gibbon and C. H. House. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, 27:4218–4222, 1999.
128. C. Forst and K. Schulten. Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomic information. *Journal of Computational Biology*, 6:343–360, 1999.
129. C. Forst and K. Schulten. Phylogenetic analysis of metabolic pathways. *Journal of Molecular Evolution*, 52:471–489, 2001.
130. I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–148, 1993.
131. Y. Freund and R. E. Schapire. A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.
132. N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3):601–620, 2000.
133. D. Frishman and P. Argos. Incorporation of non-local interactions in protein secondary structure prediction from amino acid sequence. *Protein Engineering*, 9(2):133–142, 1996.
134. H. N. Gabow. An efficient implementation of Edmonds' algorithm for maximum matching on graphs. *Journal of the ACM*, 23(2):221–234, 1976.

135. V. Ganti, J. Gehrke, and R. Ramakrishnan. Mining very large databases. *Computer*, 32:38–45, 1999.
136. H. Garcia-Molina, J. D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice Hall, Upper Saddle River, NJ, 2002.
137. J. Gasteiger, C. Rudolph, and J. Sadowski. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comp. Method*, 3:537–547, 1990.
138. T. A. Geissman. *Principles of Organic Chemistry*. W. H. Freeman and Company, San Francisco, 1968.
139. A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, New York, 1995.
140. M. Gerstein. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *Journal of Molecular Biology*, 274:562–576, 1997.
141. V. Gewin. All living things, online. *Nature*, 418:362–363, 2002.
142. J.-F. Gibrat, T. Madej, and S. H. Bryant. Surprising similarities in structure comparison. *Current Opinion in Structural Biology*, 6:377–385, 1996.
143. D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977.
144. R. Gilmour. Taxonomic markup language: applying XML to systematic data. *Bioinformatics*, 16:406–407, 2000.
145. W. Gish. WU-blast. http://blast.wustl.edu.
146. H. C. J. Godfray. Challenges for taxonomy. *Nature*, 417:17–19, 2002.
147. P. A. Goloboff and D. Pol. Semi-strict supertrees. *Cladistics*, 18:514–525, 2002.
148. J. Gonzalez, L. Holder, and D. Cook. Application of graph based concept learning to the predictive toxicology domain. In *Proceedings of the Workshop on Predictive Toxicology Challenge at the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2001.
149. B. Goryachev, P. F. MacGregor, and A. M. Edwards. Unfolding microarray data. *J. Computational Biology*, 8:443–461, 2001.
150. S. Goto, H. Bono, H. Ogata, W. Fujibuchi, T. Nishioka, K. Sato, and M. Kanehisa. Organizing and computing metabolic pathway data in terms of binary relations. In *Proceedings of the 2nd Pacific Symposium on Biocomputing*, pages 175–186, 1996.
151. O. Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982.
152. J.-L. Gouze. Positive and negative circuits in dynamical systems. *Journal of Biological Systems*, 6(1):11–15, 1998.
153. L. Gravano, P. Ipeirotis, H. Jagadish, N. Koudas, S. Muthukrishnan, and D. Srivastava. Approximate string joins in a database (almost) for free. In *Proceedings of Int. Conf. Very Large Data Bases*, pages 491–500, Rome, Italy, 2001.
154. P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
155. M. Gribskov, R. Luethy, and D. Eisenberg. Profile analysis. *Methods in Enzymology*, 183:146–159, 1989.
156. R. Grossi and J. Vitter. Compressed suffix arrays and suffix trees with applications to text indexing and string matching. In *Proceedings of ACM Symposium on Theory of Computing*, pages 397–406, Crete, Greece, 2001.
157. S. R. Gunn. Support vector machines for classification and regression. Technical Report ISIS-1-98, Department of Electronics and Computer Science, University of Southampton, 1998.

158. S. W. Guo. Linkage disequilibrium measures for fine-scale mapping: a comparison. *Human Heredity*, 47:301–314, 1997.
159. D. Gusfield. *Algorithms on Strings, Trees and Sequences, Computer Science and Computation Biology*. Cambridge University Press, New York, 1997.
160. A. Gut. *An Intermediate Course in Probability*. Springer-Verlag, New York, 1995.
161. S. Guthe, M. Wand, J. Gonser, and W. Straer. Interactive rendering of large volume data sets. In *Proceedings of IEEE Visualization Conference*, pages 53–60, Boston, MA, 2002.
162. P. Guttorp. *Stochastic Modeling of Scientific Data*. Chapman & Hall, London, 1995.
163. L. Hammel and J. M. Patel. Searching on the secondary structure of protein sequences. In *Proceedings of the 28th International Conference on Very Large Data Bases*, pages 634–645, Hong Kong, China, 2002.
164. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2001.
165. J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proceedings of ACM SIGMOD Int. Conf. Management of Data*, pages 1–12, Dallas, TX, 2000.
166. C. Hansch, P. P. Maolney, T. Fujita, and R. M. Muir. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature*, 194:178–180, 1962.
167. C. Hansch, R. M. Muir, T. Fujita, C. F. Maloney, and M. Streich. The correlation of biological activity of plant growth-regulators and chloromycetin derivatives with hammett constants and partition coefficients. *Journal of the American Chemical Society*, 85:2817–2824, 1963.
168. R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67:786–804, 1979.
169. J. Hartigan. Direct clustering of a data matrix. *J. American Stat. Assoc.*, 67:123–129, 1972.
170. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, 2001.
171. W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.
172. D. K. Heidary, Jr., J. C. O'Neill, M. Roy, and P. A. Jennings. An essential intermediate in the folding of dihydrofolate reductase. *Proceedings of the National Academy of Sciences of the USA*, 97(11):5866–5870, 2000.
173. I. Herman, G. Melançon, and M. S. Marshall. Graph visualization in information visualization: a survey. *IEEE Transactions on Visualization and Computer Graphics*, 6:24–44, 2000.
174. M. Heymans and A. K. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, pages 138–146, Brisbane, Australia, 2003.
175. D. G. Higgins and P. M. Sharp. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73:237–244, 1988.
176. D. G. Higgins, J. T. Thompson, and T. J. Gibson. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acid Research*, 22:4673–4680, 1994.
177. D. G. Higgins, J. T. Thompson, and T. J. Gibson. Using Clustal for multiple sequence alignments. *Methods in Enzymology*, 266:383–402, 1996.

178. L. Hirschman, J. Park, J. Tsujii, L. Wong, and C. H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18:1553–1561, 2002.
179. L. Holder, D. Cook, and S. Djoko. Substructure discovery in the Subdue system. In *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases*, pages 169–180, 1994.
180. A. Holloway, R. K. van Laar, R. W. Tothill, and D. Bowtell. Options available—from start to finish—for obtaining data from DNA microarrays II. *Nature Genetics Supplement*, 32:481–489, 2002.
181. L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, 233:123–138, 1993.
182. L. Holm and C. Sander. 3D lookup: fast protein structure database searches at 90% reliability. In *Proceedings of the 3rd International Conference on Intelligent Systems for Molecular Biology*, pages 179–187, 1995.
183. S. Holmes. Statistics for phylogenies. *Theoretical Population Biology*, 63:17–32, 2003.
184. S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17:721–728, 2001.
185. K. Huang, M. Velliste, and R. F. Murphy. Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. In *Proceedings of the SPIE (International Society for Optical Engineering)*, pages 307–318, 2003.
186. X. Huang and A. Madan. CAP3: a DNA sequence assembly program. *Genome Research*, 9:868–877, 1999.
187. M. Hucka et al. The system biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19:524–531, 2003.
188. J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17:754–755, 2001.
189. T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.*, 2:343–372, 2001.
190. J. R. Iglesias, G. Gupta, E. Pontelli, D. Ranjan, and B. Milligan. Interoperability between bioinformatics tools: a logic programming approach. In *Proceedings of the 3rd International Symposium on Practical Aspects of Declarative Languages*, pages 153–168, 2001.
191. A. Inokuchi, T. Washio, and H. Motoda. An apriori-based algorithm for mining frequent substructures from graph data. In *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 13–23, Lyon, France, 2000.
192. Y. E. Ioannidis. Universality of serial histograms. In *Proceedings of the 19th International Conference on Very Large Data Bases*, Dublin, Ireland, 1993.
193. Y. E. Ioannidis and V. Poosala. Balancing histogram optimality and practicality for query result size estimation. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 233–244, San Jose, CA, 1995.
194. R. M. Jackson and R. B. Russell. The serine protease inhibitor canonical loop conformation: examples found in extracellular hydrolases, toxins, cytokines and viral proteins. *Journal of Molecular Biology*, 296(2), 2000.
195. R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.
196. H. V. Jagadish, O. Kapitskaia, R. T. Ng, and D. Srivastava. Multi-dimensional substring selectivity estimation. In *Proceedings of the 25th International Conference on Very Large Data Bases*, Edinburgh, Scotland, 1999.

197. H. M. Jamil, G. A. Modica, and M. A. Teran. Querying phylogenies visually. In *Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering*, pages 3–10, 2001.
198. J. W. Jarvik, S. A. Adler, C. A. Telmer, V. Subramaniam, and A. J. Lopez. Cd-tagging: a new approach to gene and protein discovery and analysis. *BioTechniques*, 20:896–904, 1996.
199. J. W. Jarvik, G. W. Fisher, C. Shi, L. Hennen, C. Hauser, S. Adler, and P. B. Berget. In vivo functional proteomics: mammalian genome annotation using cd-tagging. *BioTechniques*, 33:852–867, 2002.
200. P. A. Jennings, B. E. Finn, et al. A reexamination of the folding mechanism of dihydrofolate reductase from *Escherichia coli*: verification and refinement of a four-channel model. *Biochemistry*, 32(14):3783–3789, 1993.
201. R. I. Jennrich. Stepwise discriminant analysis. In *Statistical Methods for Digital Computers*, pages 77–95. John Wiley, New York, 1977.
202. T. Jiang, G.-H. Lin, B. Ma, and K. Zhang. The longest common subsequence problem for arc-annotated sequences. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, pages 154–165, 2000.
203. T. Jiang, G. H. Lin, B. Ma, and K. Zhang. A general edit distance between RNA structures. *Journal of Computational Biology*, 9(2):371–388, 2002.
204. T. Joachims. *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, MA, 1999.
205. A. K. Joshi. An introduction to tree adjoining grammars. In *Mathematics of Language*, pages 87–115. John Benjamins, Amsterdam, 1987.
206. T. Kahveci and A. Singh. An efficient index structure for string databases. In *Proceedings of International Conference on Very Large Data Bases*, pages 351–360, Rome, Italy, 2001.
207. T. Kahveci and A. Singh. Variable length queries for time series data. In *Proceedings of International Conference on Data Engineering*, pages 273–282, Heidelberg, Germany, 2001.
208. T. Kahveci and A. K. Singh. MAP: searching large genome databases. In *Proceedings of Pacific Symposium on Biocomputing*, pages 303–314, 2003.
209. M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30:42–46, 2002.
210. P. Karp, M. Riley, M. Saier, I. Paulsen, J. Collado-Vides, S. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castro. The EcoCyc database. *Nucleic Acids Research*, 30:56–58, 2002.
211. K. Karplus, C. Barrett, and R. Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10), 1999.
212. G. Karypis. CLUTO: a clustering toolkit. Technical Report 02-017, Department of Computer Science, University of Minnesota, 2002.
213. G. Karypis, E.-H. Han, and V. Kumar. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *Computer*, 32:68–75, 1999.
214. L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, New York, 1990.
215. S. L. Kazmirski and V. Daggett. Simulations of the structural and dynamical properties of denatured proteins: the molten coil state of bovine pancreatic trypsin inhibitor. *Journal of Molecular Biology*, 277:487–506, 1998.
216. W. J. Kent. BLAT: The BLAST-like Alignment Tool. *Genome Research*, 12(4):656–664, 2002.
217. E. Keogh and T. Folias. The UCR time series data mining archive. http://www.cs.ucr.edu/~eamonn/TSDMA/index.html, Computer Science and Engineering Department, University of California at Riverside, 2002.

218. A. Khotanzad and Y. H. Hong. Rotation invariant image recognition using features selected via a systematic method. *Pattern Recognition*, 23:1089–1101, 1990.
219. R. D. King, S. Muggleton, R. A. Lewis, and M. J. E. Sternberg. Drug design by machine learning: the use of inductive logic programming to model the structure-activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proceedings of the National Academy of Sciences of the USA*, 89:11322–11326, 1992.
220. R. D. King, S. H. Muggleton, A. Srinivasan, and M. J. E. Sternberg. Structure-activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenecity byd inductive logic programming. *Proceedings of the National Academy of Sciences of the USA*, 93:438–442, 1996.
221. R. D. King, A. Srinivasan, and L. Dehaspe. Warmr: a data mining tool for chemical data. *Journal of Computer Aided Molecular Design*, 15:173–181, 2001.
222. H. Kitano. Computational systems biology. *Nature*, 420:206–210, 2002.
223. H. Kitano. Systems biology: a brief overview. *Science*, 295:1662–1664, 2002.
224. J. Kittler and K. Messer. Fusion of multiple experts in multimodal biometric personal identity verification systems. In *Proceedings of IEEE International Workshop on Neural Networks for Signal Processing*, pages 3–12, 2002.
225. D. K. Klimov and D. Thirumalai. Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins*, 43:465–475, 2001.
226. J. Klingner and N. Amenta. Case study: visualization of evolutionary trees. In *Proceedings of IEEE Information Visualization Conference*, pages 71–74, 2002.
227. S. Kotz, N. L. Johnson, and C. B. Read. *Encyclopedia of Statistical Sciences*. John Wiley, New York, 1981.
228. S. Kramer, L. De Raedt, and C. Helma. Molecular feature mining in HIV data. In *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining*, 2001.
229. U. Kressel. Pairwise classification and support vector machines. In B. Scholkopf, C. Burges, and A. J. Smola (eds.), *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, 1999.
230. L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics*, 58:1347–1363, 1996.
231. A. Kumar, K.-H. Cheung, P. Ross-Macdonald, P. S. R. Coelho, P. Miller, and M. Snyder. Triples: a database of gene function in *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 28:81–84, 2000.
232. M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Proceedings of IEEE International Conference on Data Mining*, pages 313–320, San Jose, CA, 2001.
233. M. Kuramochi and G. Karypis. Discovering frequent geometric subgraphs. In *Proceedings of IEEE International Conference on Data Mining*, pages 258–265, 2002.
234. M. Kuramochi and G. Karypis. An efficient algorithm for discovering frequent subgraphs. *IEEE Transactions on Knowledge and Data Engineering*, 16(6), 2004.
235. S. Kurtz and C. Schleiermacher. REPuter—fast computation of maximal repeats in complete genomes. *Bioinformatics*, 15(5):426–427, 1999.
236. K. Kuwata, R. Shastry, H. Cheng, M. Hoshino, C. A. Bhatt, Y. Goto, and H. Roder. Structural and kinetic characterization of early folding events of lactoglobulin. *Nature*, 8(2):151–155, 2001.

237. J. C. Lam, K. Roeder, and B. Devlin. Haplotype fine-mapping by evolutionary trees. *American Journal of Human Genetics*, 66:659–673, 2000.
238. J. Lamping, R. Rao, and P. Pirioli. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of International Conference on Human Factors in Computing Systems*, pages 401–408, 1995.
239. C. Lauk. An attempt for a genus-level supertree of birds. BSc(Hons) project, DEEB, IBLS, University of Glasgow, 2002.
240. T. Lazardis and M. Karplus. New view of protein folding reconciled with the old through multiple unfolding simulations. *Science*, 278:1928–1931, 1997.
241. L. C. Lazzeroni. A chronology of fine-scale gene mapping by linkage disequilibrium. *Statistical Methods in Medical Research*, 10:57–76, 2001.
242. S. Y. Le, R. Nussinov, and J. V. Mazel. Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res.*, 22:461–473, 1989.
243. S. Y. Le, J. Owens, R. Nussinov, J. H. Chen, B. Shapiro, and J. V. Mazel. RNA secondary structures: comparisons and determination of frequently recurring substructures by consensus. *Comput. Appl. Biosci.*, 5:205–210, 1989.
244. S. M. Le Grand and J. K. M. Merz. Rapid approximation to molecular surface area via the use of boolean logic look-up tables. *Journal of Computational Chemistry*, 14:349–352, 1993.
245. A. R. Leach. *Molecular Modeling: Principles and Applications.* Prentice Hall, Upper Saddle River, NJ, 2001.
246. A. M. Lesk. *Introduction to Bioinformatics.* Oxford University Press, New York, 2002.
247. C. Levinthal. Are there pathways for protein folding? *Journal of Chemical Physics*, 65:44–45, 1968.
248. B. Lewin. *Genes VII.* Oxford University Press, New York, 2000.
249. W. Li, J. Han, and J. Pei. CMAR: accurate and efficient classification based on multiple class-association rules. In *Proceedings of IEEE International Conference on Data Mining*, pages 369–376, 2001.
250. S. Liang, S. Fuhrman, and R. Somogyi. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. In *Proceedings of Pacific Symposium on Biocomputing*, pages 18–29, 1998.
251. L. Liao, S. Kim, and J.-F. Tomb. Genome comparisons based on profiles of metabolic pathways. In *Proceedings of the 6th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, Crema, Italy, 2002.
252. G.-H. Lin, B. Ma, and K. Zhang. Edit distance between two RNA structures. In *Proceedings of the 5th Annual International Conference on Computational Molecular Biology*, pages 200–209, 2001.
253. B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 80–86, 1998.
254. J. S. Liu and C. E. Lawrence. Bayesian inference on biopolymer models. *Bioinformatics*, 15(1):38–52, 1999.
255. V. Lombard, E. B. Cameron, H. E. Parkinson, P. Hingamp, G. Stoesser, and N. Redaschi. EMBL-Align: a new public nucleotide and amino acid multiple sequence alignment database. *Bioinformatics*, 18:763–764, 2002.
256. D. P. Lopresti and A. Tomkins. Block edit models for approximate string matching. *Theoretical Computer Science*, 181(1):159–179, 1997.
257. B. Ma, L. Wang, and K. Zhang. Computing similarity between RNA structures. *Theoretical Computer Science*, 276:111–132, 2002.

258. M. Ma, J. Tromp, and M. Li. PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18:1–6, 2002.
259. D. R. Maddison, D. L. Swofford, and W. P. Maddison. NEXUS: an extensible file format for systematic information. *Systematic Biology*, 46:590–621, 1997.
260. D. R. Maddison, W. P. Maddison, J. Frumkin, and K.-S. Schulz. The Tree of Life project: a multi-authored, distributed Internet project containing information about phylogeny and biodiversity. In H. Saarenmaa and E. S. Nielsen (eds.), *Towards a Global Biological Information Infrastructure: Challenges, Opportunities, Synergies, and the Role of Entomology*. European Environment Agency, Copenhagen, pages 5–14, 2002.
261. T. Madej, J.-F. Gibrat, and S. H. Bryant. Threading a database of protein cores. *Proteins: Structure, Function, and Genetics*, 23:356–369, 1995.
262. B. L. Maidak, J. R. Cole, T. G. Lilburn, C. T. Parker, P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje. The RDP-II (Ribosomal Database Project). *Nucleic Acids Research*, 29:173–174, 2001.
263. J. Mallet and K. Willmott. Taxonomy: renaissance or tower of babel. *Trends in Ecology and Evolution*, 18:57–59, 2003.
264. B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:837–842, 1996.
265. B. Mann, R. Williams, M. Atkinson, K. Brodlie, A. Storkey, and C. Willimans. Scientific data mining, integration and visualization. In *Report of the Workshop on Scientific Data Mining, Integration and Visualization*, The E-Science Institute, Edinburgh, 2002.
266. H. Mannila and M. Salmenkivi. Finding simple intensity descriptions from event sequence data. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 341–346, 2001.
267. R. Mao, D. P. Miranker, J. N. Sarvela, and W. Xu. Clustering sequences in a metric space, the Mobios Project. Technical Report, University of Texas, Austin, 2003.
268. M. K. Markey, M. V. Boland, and R. F. Murphy. Towards objective selection of representative microscope images. *Biophys. J.*, 76:2230–2237, 1999.
269. H. Matsuno, R. Murakani, R. Yamane, N. Yamasaki, S. Fujita, H. Yoshimori, and S. Miyano. Boundary formation by notch signaling in *Drosophila* multicellular systems: experimental observations and gene network modeling by genomic object net. In *Proceedings of Pacific Symposium on Biocomputing*, pages 152–163, 2003.
270. M. S. McPeek and A. Strahs. Assessment of linkage disequilibrium by the decay of haplotype sharing, with applications to fine-scale genetic mapping. *American Journal of Human Genetics*, 65:858–875, 1999.
271. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.
272. D. P. Minaker. Metric-space indexes as a basis for scalable biological databases. *Omics: A Journal of Integrative Biology*, 7:57–60, 2003.
273. T. M. Mitchell. *Machine Learning*. WCB/McGraw-Hill, New York, 1997.
274. Y. Mok, C. Kay, L. Kay, and J. Forman-Kay. NOE data demonstrating a compact unfolded state for an sh3 domain under non-denaturing conditions. *Journal of Molecular Biology*, 289(3):619–638, 1999.
275. B. Morgenstern, K. Frech, A. Dress, and T. Werner. Dialign: finding local similarities by multiple sequence alignment. *Bioinformatics*, 14:290–294, 1998.

276. K. Morik, P. Brockhausen, and T. Joachims. Combining statistical learning with a knowledge-based approach—a case study in intensive care monitoring. In *Proceedings of International Conference on Machine Learning*, pages 268–277, 1999.
277. A. P. Morris, J. C. Whittaker, and D. J. Balding. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *American Journal of Human Genetics*, 70:686–707, 2002.
278. R. Mott, J. Schultz, P. Bork, and C. P. Ponting. Predicting protein cellular localization using a domain projection method. *Genome Research*, 12:1168–1174, 2002.
279. D. Mount. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, Woodbury, NY, 2001.
280. S. Muggleton. Inverse entailment and Progol. *New Generation Computing*, 13:245–286, 1995.
281. S. Muggleton and L. De Raedt. Inductive logic programming: theory and methods. *Journal of Logic Programming*, 19:629–679, 1994.
282. S. H. Muggleton and C. Feng. Efficient induction of logic programs. In S. Muggleton (ed.), *Inductive Logic Programming*. Academic Press, London, pages 281–298, 1992.
283. T. Munzner. Exploring large graphs in 3D hyperbolic space. *IEEE Computer Graphics and Applications*, 18:18–23, 1998.
284. T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. TreeJuxtaposer: scalable tree comparison using focus+context with guaranteed visibility. *ACM Transactions on Graphics*, 22(3):453–462, 2003.
285. M. Muralikrishna and D. J. DeWitt. Equi-depth histograms for estimating selectivity factors for multi-dimensional queries. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, Chicago, IL, 1994.
286. R. F. Murphy, M. V. Boland, and M. Velliste. Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, pages 251–259, 2000.
287. R. F. Murphy, M. Velliste, and G. Porreca. Robust classification of subcellular location patterns in fluorescence microscope images. In *Proceedings of IEEE International Workshop on Neural Networks for Signal Processing*, pages 67–76, 2002.
288. R. F. Murphy, M. Velliste, and G. Porreca. Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. *Journal of VLSI (Very Large Scale Integrated) Signal Processing*, 35:311–321, 2003.
289. S. Muthukrishnan and S. Sahinalp. Approximate nearest neighbors and sequence comparison with block operations. In *Proceedings of ACM Symposium on Theory of Computing*, pages 416–422, Crete, Greece, 2001.
290. E. W. Myers and W. Miller. Optimal alignments in linear space. *Computer Applications in the Biosciences*, 4(1):11–17, 1988.
291. K. Nakai. Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.*, 54:277–344, 2000.
292. L. Nakhleh, D. Miranker, F. Barbancon, W. H. Piel, and M. J. Donoghue. Requirements of phylogenetic databases. In *Proceedings of the 3rd IEEE Symposium on Bioinformatics and Bioengineering*, pages 141–148, 2003.

293. A. Natsev, Y.-C. Chang, J. R. Smith, C.-S. Li, and J. S. Vitter. Supporting incremental join queries on ranked inputs. In *Proceedings of the 28th International Conference on Very Large Data Bases*, Rome, Italy, 2001.
294. 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.
295. M. Nei. Phylogenetic analysis in molecular evolutionary genetics. *Annual Review of Genetics*, 30:371–403, 1996.
296. C. Nishimura, S. Prytulla, H. J. Dyson, and P. E. Wright. Conservation of folding pathways in evolutionary distant globin sequences. *Nature Structural Biology*, 7(8):679–686, 2000.
297. K. C. Nixon and J. M. Carpenter. On the other "Phylogenetic Systematics." *Cladistics*, 16:298–318, 2000.
298. B. Nolting, R. Golbik, J. Neira, A. Soler-Gonzalez, G. Schreiber, and A. Fersht. The folding pathway of a protein at high resolution from microseconds to seconds. *Proceedings of the National Academy of Sciences of the USA*, 94(3):826–30, 1997.
299. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzyme-catalysed reactions. http://www.chem.qmul.ac.uk/iubmb/enzyme/.
300. C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302:205–217, 2000.
301. R. Nussinov and H. J. Wolfson. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proceedings of the National Academy of Sciences of the USA*, 88:10495–10499, 1991.
302. H. Ogata, H. Bono, W. Fujibuchi, S. Goto, and M. Kanehisa. Analysis of binary relations and hierarchies of enzymes in the metabolic pathways. *Genome Informatics*, 7:128–136, 1996.
303. V. Ollikainen. Simulation techniques for disease gene localization in isolated populations. Technical Report A-2002-2, University of Helsinki, 2002.
304. P. Onkamo, V. Ollikainen, P. Sevon, H. T. T. Toivonen, H. Mannila, and J. Kere. Association analysis for quantitative traits by data mining: QHPM. *Annals of Human Genetics*, 66:419–429, 2002.
305. C. A. Orengo, A. E. Todd, and J. M. Thornton. From protein structure to function. *Current Opinion in Structural Biology*, 9(3), 1999.
306. R. D. M. Page. TreeView: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, 12:357–358, 1996.
307. R. D. M. Page. Modified mincut supertrees. In *Proceedings of the 2nd Workshop on Algorithms in Bioinformatics*, pages 537–551, 2002.
308. R. D. M. Page and E. C. Holmes. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Scientific, Oxford, 1998.
309. C. H. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *J. Comput. System Sciences*, 43:425–440, 1991.
310. S. Park, D. Lee, and W. W. Chu. Fast retrieval of similar subsequences in long sequence databases. In *Proceedings of the 3rd IEEE Knowledge and Data Engineering Exchange Workshop*, Chicago, IL, 1999.

311. W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA*, 85:2444–2448, 1988.
312. W. H. Piel, M. J. Donoghue, and M. J. Sanderson. TreeBASE: a database of phylogenetic knowledge. In *To the Interoperable Catalogue of Life with Partners—Species 2000 Asia Oceania—Proceedings of the 2nd International Workshop of Species 2000 (Research Report for the National Institute of Environmental Studies)*, R-171-2002, 2002.
313. W. H. Piel, M. J. Donoghue, and M. J. Sanderson. The small-world dynamics of tree networks and data mining in phyloinformatics. *Bioinformatics*, 19:1162–1168, 2003.
314. P. Pirolli, S. K. Card, and M. M. van Der Wege. The effect of information scent on searching information visualizations of large tree structures. In *Proceedings of the 5th International Working Conference on Advanced Visual Interfaces*, pages 161–172, 2000.
315. C. Plaisant, J. Grosjean, and B. B. Bederson. SpaceTree: supporting exploration in large node link tree, design evolution and empirical evaluation. In *Proceedings of IEEE Symposium on Information Visualization*, pages 57–70, 2002.
316. J. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin dags for multiclass classification. *Adv. Neural Inform. Proc. Systems*, 12:547–553, 2000.
317. F. J. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.
318. M. Purrello, C. Di Pietro, M. Ragusa, A. Pulvirenti, G. Pigola, R. Giugno, E. Modica, V. Zimmitti, V. Di Pietro, T. Maugeri, G. Emmanuele, S. Travali, M. Scalia, D. Shasha, and A. Ferro. In vitro and in silico cloning of *Xenopus laevis* sod2 and its phylogenetic analysis with AntiClustAl, a new algorithm for multiple sequence alignment, demonstrate a very high amino acid sequence conservation during evolution. Submitted, 2003.
319. W. K. Purves, D. E. Sadava, G. H. Orians, and H. C. Heller. *Life, the science of biology*. Sinauer Associates, Sunderland, MA, 2001.
320. R. L. Pyle. Core data model for managing taxonomic names, concepts, and associated references. http://www2.bishopmuseum.org/PBS.schema/Pyle Schema.pdf.
321. J. Quackenbush. Computational analysis of microarray data. *Natural Review Genetics*, 2:418–427, 2001.
322. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.
323. M. A. Ragan. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, 1:53–58, 1992.
324. C. Raguenaud and J. Kennedy. Multiple overlapping classifications: issues and solutions. In *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, pages 77–86, 2002.
325. V. Raman and J. M. Hellerstein. Potter's wheel: an interactive data cleaning system. In *Proceedings of Int. Conf. on Very Large Data Bases*, pages 381–390, Rome, Italy, 2001.
326. V. E. Ramensky, V. J. Makeev, M. A. Roytberg, and V. G. Tumanyan. DNA segmentation through the Bayesian approach. *Journal of Computational Biology*, 7(1):215–231, 2000.
327. D. E. Reich, S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin, J. M. Higgins, D. J. Richter, E. S. Lander, and D. Altshuler. Human genome sequence variation, and the influence of gene history, mutation and recombination. *Nature Genetics*, 32:135–142, 2002.

328. G. W. Richards. Virtual screening using grid computing: the screensaver project. *Nature Reviews: Drug Discovery*, 1:551–554, 2002.
329. T. W. Ridler and S. Calvard. Picture thresholding using an iterative selection method. *IEEE Trans. Syst. Man Cybernet*, 8:630–632, 1978.
330. C. J. R. Robertson and G. B. Nunn. Towards a new taxonomy for albatrosses. In G. Robertson and R. Gales (eds.), *Albatross Biology and Conservation*. Surrey Beaty, Chipping Norton, UK, pages 13–19, 1997.
331. D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.
332. M. M. Rolls, P. A. Stein, S. S. Taylor, E. Ha, F. McKeon, and T. A. Rapoport. A visual screen of a gfp-fusion library identifies a new type of nuclear envelope membrane protein. *J. Cell Biol*, 146:29–44, 1999.
333. E. J. S. Roques and R. F. Murphy. Objective evaluation of differences in protein subcellular distribution. *Traffic*, 3:61–65, 2002.
334. B. Rost. Review: protein secondary structure prediction continues to rise. *Journal of Structural Biology*, 134(2), 2001.
335. B. Rost. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2), 2001.
336. U. Rost and E. Bornberg-Bauer. TreeWiz: interactive exploration of huge trees. *Bioinformatics*, 18:109–114, 2002.
337. D. A. Rozwarski, A. M. Groneborn, M. G. Clore, J. F. Bazan, A. Bohm, A. Wlodawer, M. Hatada, and P. A. Karplus. Structural comparisons among the short-chain helical cytokines. *Structure*, 2:159–173, 1994.
338. R. B. Russell, P. D. Sasieni, and M. J. Sternberg. Supersites within superfolds—binding site similarity in the absence of homology. *Journal of Molecular Biology*, 282(4), 1998.
339. D. A. Ruths, E. S. Chen and L. Ellis. Arbor 3D: an interactive environment for examining phylogenetic and taxonomic trees in multiple dimensions. *Bioinformatics*, 16:1003–1009, 2000.
340. N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.
341. S. Salzberg, A. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26:544–548, 1998.
342. R. Sánchez, U. Pieper, N. Mirkovi, P. I. W. de Bakker, E. Wittenstein, and A. Šali. MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Research*, 28(1):250–253, 2000.
343. M. J. Sanderson, A. Purvis, and C. Henze. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution*, 13:105–109, 1998.
344. S. Sarkar and P. Soundararajan. Supervised learning of large perceptual organization: graph spectral partitioning and learning automata. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:504–525, 2000.
345. M. A. Savageau. Power-law formalism: a canonical nonlinear approach to modeling and analysis. In *Proceedings of the 1st World Congress on Nonlinear Analysis*, pages 3323–3334, 1996.
346. J. Schaff, B. Slepchenko, and L. Loew. Physiological modeling with virtual cell framework. *Methods in Enzymology*, 321:1–23, 2000.
347. R. E. Schapire. The boosting approach to machine learning: an overview. In *Proceedings of the MSRI (Mathematical Sciences Research Institute) Workshop on Nonlinear estimation and Classification*, 2002.
348. R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336, 1999.

349. C. H. Schilling, D. Letscher, and B. P. Palsson. Theory for the systemic definition of pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203:229–248, 2000.
350. C. H. Schilling, D. Letscher, and B. P. Palsson. Assessment of the metabolic capabilities of *H. influenzae* Rd through a genome-scale pathway analysis. *Journal of Theoretical Biology*, 203:249–283, 2000.
351. B. Schoelkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.
352. B. Scholkopf, A. Smola, and K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10:1299–1319, 1998.
353. S. Schwartz, Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. PipMaker—a Web server for aligning two genomic DNA sequences. *Genome Research*, 10(4):577–586, 2000.
354. P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price. Access path selection in a relational database management system. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, Boston, 1979.
355. C. Semple and M. Steel. A supertree method for rooted trees. *Discrete Applied Mathematics*, 105:147–158, 2000.
356. S. K. Service, D. W. T. Lang, N. B. Freimer, and L. A. Sandkuijl. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *American Journal of Human Genetics*, 64:1728–1738, 1999.
357. P. Sevon, P. Onkamo, V. Ollikainen, H. T. T. Toivonen, H. Mannila, and J. Kere. Mining the associations between phenotype, genotype, and covariates. Genetic Analysis Workshop 12, *Genetic Epidemiology*, 21(Suppl. 1):S588–S593, 2001.
358. P. Sevon, H. T. T. Toivonen, and V. Ollikainen. TreeDT: gene mapping by tree disequilibrium test. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 365–370, 2001 (extended version available at http://www.cs.helsinki.fi/TR/C.html).
359. H. Shan, K. G. Herbert, W. H. Piel, D. Shasha, and J. T. L. Wang. A structure-based search engine for phylogenetic databases. In *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, pages 7–10, 2002.
360. B. A. Shapiro. An algorithm for comparing multiple RNA secondary structures. *Comput. Appl. Biosci.*, 4(3):387–393, 1988.
361. B. A. Shapiro and K. Zhang. Comparing multiple RNA secondary structures using tree comparisons. *Comput. Appl. Biosci.*, 6(4):309–318, 1990.
362. D. Shasha, J. T. L. Wang, H. Shan, and K. Zhang. ATreeGrep: approximate searching in unordered trees. In *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, pages 89–98, 2002.
363. P. Shenoy, J. R. Haritsa, S. Sundarshan, G. Bhalotia, M. Bawa, and D. Shah. Turbo-charging vertical mining of large databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 22–33, 2000.
364. J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 731–737, 1997.

365. I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9):739–747, 1998.
366. A. P. Singh and D. L. Brutlag. Hierarchical protein structure superposition using both secondary structure and atomic representations. In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*, pages 284–293, 1997.
367. K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, 1997.
368. M. Sippl. Helmholtz free energy of peptide hydrogen bonds in proteins. *Journal of Molecular Biology*, 260(5):644–648, 1996.
369. M. Sipser. *Introduction to the Theory of Computation*. PWS, Boston, 1997.
370. J. Skolnick, A. Kolinski, and A. Ortiz. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins*, 38(1):3–16, 2000.
371. T. F. Smith and M. S. Waterman. Comparison of biosequences. *Adv. Appl. Math.*, 2:482–489, 1981.
372. T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.
373. P. H. A. Sneath and R. R. Sokal. *Numerical Taxonomy*, W. H. Freeman and Company, San Francisco, pages 230–234, 1973.
374. M. Sofer. Genealogical representation of trees in databases. Unpublished manuscript (http://www.utdt.edu/~mig/sql-trees/).
375. R. Sole, R. Ferrer-Cancho, J. Montoya, and S. Valverde. Selection, tinkering, and emergence in complex networks. *Complexity*, 8:20–33, 2003.
376. R. Somogyi, S. Fuhrman, and X. Wen. *Genetic network inference in computational models and applications to large-scale gene expression data*. MIT Press, Cambridge, MA, 2001.
377. R. Somogyi and C. A. Sniegoski. Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. *Journal of Computational Biology*, 6(1):45–63, 1996.
378. K. Sparck-Jones and J. Galliers. *Evaluating Natural Language Processing Systems, an analysis and review*. Springer-Verlag, New York, 1996.
379. P. T. Spellman et al. Design and implementation of microarray gene expression markup language (mage-ml). *Genome Biology*, 3:0046.1–9, 2002.
380. R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, 52:506–516, 1993.
381. A. Srinivasan, R. D. King, S. H. Muggleton, and M. Sternberg. The predictive toxicology evaluation challenge. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pages 1–6, 1997.
382. A. Sriniviasan and R. D. King. Feature construction with inductive logic programming: a study of quantitative predictions of biological activity aided by structural attributes. *Knowledge Discovery and Data Mining*, 3:37–57, 1999.
383. N. Stahl and G. D. Yancopoulos. The alphas, betas, and kinases of cytokine receptor complexes. *Cell*, 74:587–590, 1993.
384. B. Stapley and G. Benoit. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in MEDLINE abstracts. In *Proceedings of Pacific Symp. Biocomputing*, pages 529–540, 2000.

385. D. J. States and P. Agarwal. Compact encoding strategies for DNA sequence similarity search. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology*, pages 211–217, 1996.
386. D. J. Stephens and V. J. Allan. Light microscopy techniques for live cell imaging. *Science*, 300:82–86, 2003.
387. M. J. E. Sternberg and J. M. Thornton. On the conformation of proteins: the handedness of the connection between parallel beta strands. *Journal of Molecular Biology*, 110:269–283, 1977.
388. M. Stoer and F. Wagner. A simple min-cut algorithm. *Journal of the ACM*, 44(4):585–591, 1997.
389. A. M. Sugden, B. R. Jasny, E. Culotta, and E. Pennisi. Charting the evolutionary history of life. *Science*, 300:1691, 2003.
390. K. C. Tai. The tree to tree correction problem. *Journal of the ACM*, 26(3):422–433, 1979.
391. T. A. Tatusova and T. L. Madden. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters*, pages 247–250, 1999.
392. W. R. Taylor. Protein structure comparison using iterated double dynamic programming. *Protein Science*, 8:654–665, 1999.
393. W. R. Taylor and C. O. Orengo. Protein structure alignment. *Journal of Molecular Biology*, 208:1–22, 1989.
394. C. A. Telmer, P. B. Berget, B. Ballou, R. F. Murphy, and J. W. Jarvik. Epitope tagging genomic DNA using a cd-tagging tn10 minitransposon. *BioTechniques*, 32:422–430, 2002.
395. J. D. Terwilliger. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American Journal of Human Genetics*, 56:777–787, 1995.
396. R. Thomas and R. D'Ari. *Biological Feedback*. CRC Press, Boca Raton, FL, 1990.
397. R. Thomas, D. Thieffry, and M. Kaufman. Dynamical behavior of biological regulatory networks: I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bulletin of Mathematical Biology*, 57(2):247–276, 1995.
398. J. D. Thompson, F. Plewniak, and O. Poch. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*, 27(13):2682–2690, 1999.
399. J. L. Thorley and R. D. M. Page. RadCon: phylogenetic tree comparison and consensus. *Bioinformatics*, 16:486–487, 2000.
400. L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22(4):1701–1728, 1994.
401. H. T. T. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, M. Herr, and J. Kere. Data mining applied to linkage disequilibrium mapping. *American Journal of Human Genetics*, 67:133–145, 2000.
402. H. T. T. Toivonen, P. Onkamo, K. Vasko, V. Ollikainen, P. Sevon, H. Mannila, and J. Kere. Gene mapping by haplotype pattern mining. In *Proceedings of IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pages 99–108, 2000.
403. M. Tomita et al. E-cell: software environment for while-call simulation. *Bioinformatics*, 15:72–84, 1999.
404. B. Tower. Docking topical hierarchies: a comparison of two algorithms for reconciling keyword structures. Technical Report CS2001-0669, Department of Computer Science and Engineering, University of California at San Diego, 2001.

405. C. Traina, A. Traina, L. Wu, and C. Faloutsos. Fast feature selection using the fractal dimension. In *Proceedings of the XV Brazilian Symposium on Databases*, 2000.
406. E. C. Uberbacher and R. J. Mural. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proceedings of the National Academy of Sciences of the USA*, 88:11261–11265, 1991.
407. G. Valiente. Constrained tree inclusion. In *Proceedings of the 14th Annual Symposium on Combinatorial Pattern Matching*, pages 361–371, 2003.
408. V. Vapnik. *Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1995.
409. V. Vapnik. *Statistical Learning Theory.* John Wiley, New York, 1998.
410. E. O. Voit. *Computational Analysis of Biochemical Systems. A Practical Guide for Biochemists and Molecular Biologists.* Cambridge University Press, Cambridge, UK, 2000.
411. M. Velliste. Image interpretation methods for a systematic of protein subcellular location. Technical Report, Carnegie Mellon University, Pittsburgh, PA, 2002.
412. M. Velliste and R. F. Murphy. Automated determination of protein subcellular locations from 3D fluorescence microscope images. In *Proceedings of IEEE International Symposium on Biomedical Imaging*, pages 867–870, 2002.
413. M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, 1997.
414. M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus. Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409:641–645, 2001.
415. V. Villegas, J. C. Martinez, F. X. Aviles, and L. Serrano. Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *Journal of Molecular Biology*, 283:1027–1036, 1998.
416. G. von Heijne, H. Nielsen, J. Engelbrecht, and S. Brunak. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 10:1–6, 1997.
417. R. Waagepetersen and D. Sorensen. A tutorial on Reversible Jump MCMC with a view toward applications in QTL-mapping. *International Statistical Review*, 69:49–62, 2001.
418. H. M. Wain, M. Lush, F. Ducluzeau, and S. Povey. Genew: the human nomenclature database. *Nucleic Acids Research*, 30:169–171, 2002.
419. H. Wang, W. Wang, J. Yang, and P. S. Yu. Clustering by pattern similarity in large data sets. In *Proceedings of ACM SIGMOD Int. Conf. on Management of Data*, pages 418–427, Madison, WI, 2002.
420. J. T. L. Wang, H. Shan, D. Shasha, and W. H. Piel. TreeRank: a similarity measure for nearest neighbor searching in phylogenetic databases. In *Proceedings of the 15th International Conference on Scientific and Statistical Database Management*, pages 71–180, 2003.
421. Z. Wang and K. Zhang. Alignment between RNA structures. In *Proceedings of the 26th International Symposium on Mathematical Foundations of Computer Science*, pages 690–702, 2001.
422. S. R. Waterhouse. Classification and regression using mixtures of experts. Technical Report, Cambridge, UK, 1997.
423. D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28:31–36, 1988.

424. O. S. Weislow, R. Kiser, D. L. Fine, J. Bader, R. H. Shoemaker, and M. R. Boyd. New soluble formazan assay for HIV-1 cytopathic effects: application to high flux screening of synthetic and natural products for AIDS antiviral activity. *Journal of the National Cancer Institute*, 81:577–586, 1989.
425. D. R. Westhead, T. W. F. Slidel, T. P. J. Flores, and J. M. Thornton. Protein structural topology: automated analysis, diagrammatic representation and database searching. *Protein Science*, 8:897–904, 1999.
426. D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schulter, T. A. Tatusova, and B. A. Rapp. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 28:10–14, 2000.
427. H. E. Williams and J. Zobel. Indexing and retrieval for genomic databases. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):63–78, 2002.
428. WIT2. http://www-unix.mcs.anl.gov/compbio/.
429. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Mateo, CA, 2000.
430. Y. I. Wolf, I. B. Rogozin, N. V. Grishin, and E. V. Koonin. Genome trees and the Tree of Life. *Trends in Genetics*, 18:472–479, 2002.
431. H. J. Wolfson and I. Rigoutsos. Geometric hashing: an overview. *IEEE Computational Science and Engineering*, 4(4):10–21, 1997.
432. C. H. Wu, L. S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek, C. R. Vinayaka, J. Zhang, and W. C. Barker. The protein information resource. *Nucleic Acids Research*, 31(1):390–392, 2003.
433. Z. Wu and R. M. Leahy. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113, 1993.
434. X. Yan and J. Han. gSpan: graph-based substructure pattern mining. In *Proceedings of Int. Conf. on Data Mining*, pages 721–724, Maebashi, Japan, 2002.
435. X. Yan and J. Han. CloseGraph: mining closed frequent graph patterns. In *Proceedings of ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Washington, DC, 2003.
436. X. Yan, J. Han, and R. Afshar. CloSpan: mining closed sequential patterns in large datasets. In *Proceedings of SIAM Int. Conf. Data Mining*, pages 166–177, San Francisco, 2003.
437. M. D. Yandell and W. H. Majoros. Genomics and natural language processing. *Nature Review, Genetics*, 3:601–610, 2002.
438. J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. *IEEE Intell. Systems*, 13:44–49, 1998.
439. J. Yang, W. Wang, H. Wang, and P. S. Yu. $\delta$-cluster: capturing subspace correlation in a large data set. In *Proceedings of Int. Conf. Data Engineering*, pages 517–528, San Francisco, 2002.
440. J. Yang, W. Wang, and P. S. Yu. Mining asynchronous periodic patterns in time series data. In *Proceedings of ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, pages 275–279, Boston, 2000.
441. J. Yang, W. Wang, P. S. Yu, and J. Han. Mining long sequential patterns in a noisy environment. In *Proceedings of ACM SIGMOD Int. Conf. on Management of Data*, pages 406–417, Madison, WI, 2002.
442. H. Yu, J. Yang, and J. Han. Classifying large data sets using SVM with hierarchical clusters. In *Proceedings of ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, Washington, DC, 2003.

443. M. J. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(2):372–390, 2000.
444. M. J. Zaki and K. Gouda. Fast vertical mining using diffsets. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 326–335, 2003.
445. K. Zhang. Computing similarity between RNA secondary structures. In *Proceedings of IEEE International Joint Symposia on Intelligence and Systems*, pages 126–132, Rockville, MD, 1998.
446. K. Zhang, L. Wang, and B. Ma. Computing similarity between RNA structures. In *Proceedings of the 10th Symposium on Combinatorial Pattern Matching*, pages 281–293, 1999.
447. K. Zhang and D. Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Computing*, 18(6):1245–1262, 1989.
448. K. Zhang, J. T. L. Wang, and D. Shasha. On the editing distance between undirected acyclic graphs. *International Journal of Foundations of Computer Science*, 7(1):43–57, 1996.
449. S. Zhang and H. Zhao. Linkage disequilibrium mapping with genotype data. *Genetic Epidemiology*, 22:66–77, 2002.
450. S. Zhang, K. Zhang, J. Li, and H. Zhao. On a family-based haplotype pattern mining method for linkage disequilibrium mapping. In *Proceedings of the 7th Pacific Symposium on Biocomputing*, pages 100–111, 2002.
451. T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of ACM SIGMOD Int. Conf. Management of Data*, pages 103–114, Montreal, Canada, 1996.
452. Z. Zhang, A. A. Schaffer, W. Miller, T. L. Madden, D. J. Lipman, E. V. Koonin, and S. F. Altschul. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Research*, 26(17):3986–3990, 1998.
453. Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, 7:203–214, 2000.
454. Y. Zhong, S. Jung, S. Pramanik, and J. H. Beaman. Data model and comparison and query methods for interacting classifications in a taxonomic database. *Taxon*, 45:223–241, 1996.
455. Y. Zhong, C. A. Meacham, and S. Pramanik. A general method for tree-comparison based on subtree similarity and its use in a taxonomic database. *BioSystems*, 42:1–8, 1997.
456. Y. Zhong, Y. Luo, S. Pramanik, and J. H. Beaman. HICLAS: a taxonomic database system for displaying and comparing biological classification and phylogenetic trees. *Bioinformatics*, 15:149–156, 1999.