REFERENCES

References

1. Practical data integration in biopharmaceutical R&D: Strategies and technologies. White paper, 3rd Millennium Inc., 125 Cambridge Park Drive, Cambridge, MA 02140, 2002.
2. S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann, 1999.
3. S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.
4. J. P. Abrahams, M. van den Berg, E. van Batenburg, and C. Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Research*, 18:3035–3044, 1990.
5. F. Achard, G. Vaysseix, and E. Barillot. XML, bioinformatics and data integration. *Bioinformatics*, 17(2):115–125, 2001.
6. J. Adachi and M. Hasegawa. *MOLPHY Version 2.3: Programs for Molecular Phylogenetics Based on Maximum Likelihood*. Number 28 in Computer Science Monographs. Institute of Statistical Mathematics, Tokyo, 1996.
7. M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, et al. The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–2195, 2000.
8. M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and Venter J.C. Complementary DNA sequencing: Expressed sequence tags and the human genome project. *Science*, 252:1651–1656, 1991.
9. P. Agarwal and V. Bafna. The ribosome scanning model for translation initiation: Implications for gene prediction and full-length cDNA detection. In *Proceedings of 6th International Conference on Intelligent Systems for Molecular Biology*, pages 2–7, 1998.
10. R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of ACM-SIGMOD International Conference on Management of Data*, pages 207–216, 1993.
11. R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In *Advances in knowledge discovery and data mining*, pages 1–34. MIT Press, 1996.
12. R. Agrawal and S. Srikant. Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases*, pages 487–499,

1994.

13. B. Aguado and R. D. Campbell. Characterization of a human lysophosphatidic acid acyltransferase that is encoded by a gene located in the class III region of the human major histocompatibility complex. *Journal of Biological Chemistry*, 273:4096–4105, 1998.
14. Y. Akiyama and M. Kanehisa. NeuroFold: An RNA secondary structure prediction system using a Hopfield neural network. In *Proceedings of 3rd Genome Informatics Workshop*, 1992. In Japanese.
15. T. Akutsu. Dynamic programming algorithm for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, pages 45–62, 2000.
16. T. Akutsu, S. Miyano, and S. Kuhara. Algorithms for inferring qualitative models of biological networks. In *Proceedings of Pacific Symposium on Biocomputing 2000*, pages 293–304, 2000.
17. T. Akutsu and S. Miyano. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Proceedings of Pacific Symposium on Biocomputing'99*, pages 17–28, 1999.
18. B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Publishing, 4th edition, 2002.
19. N. N. Alexandrov and A. A. Mironov. Application of a new method of pattern recognition in DNA sequence analysis: A study of *E. coli* promoters. *Nucleic Acids Research*, 18:1847–1852, 1990.
20. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.
21. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S.Y.D. Mack, and J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745–6750, 1999.
22. P. Aloy and R. B. Russell. Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. USA*, 99(9):5896–5901, 2002.
23. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
24. 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(17):3389–3402, 1997.
25. M. A. Andrade, S. I. O'Donoghue, and B. Rost. Adaptation of protein surfaces to subcellular location. *Journal of Molecular Biology*, 276:517–525, 1998.
26. M. A. Andrade and C. Sander. Bioinformatics: From genome data to biological knowledge. *Current Opinion in Biotechnology*, 8:675–683, 1997.
27. Angiosperm Phylogeny Group. An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden*, 85:531–553, 1998.
28. D. R. Appling. Genetic approaches to the study of protein-protein interactions. *Methods*, 19(2):338–349, 1999.
29. R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. R. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M Oinn, M. Pagni, F. Servant,

C. J. A. Sigrist, and E. M. Zdobnov. The InterPro database, an integrated documentation resource for protein families, domains, and functional sites. *Nucleic Acids Research*, 29:37–40, 2001.

30. R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, L. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischman, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sirgist, and E. M. Zdobnov. InterPro—an integrated documentation resource for protein families, domains, and functional sites. *Bioinformatics*, 16:1145–1150, 2000.
31. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408:796–815, 2000.
32. E. Ardini, R. Agresti, E. Tagliabue, M. Greco, P. Aiello, L. T. Yang, S. Menard, and J. Sap. Expression of protein tyrosine phosphatase alpha (RPTPalpha) in human breast cancer correlates with low tumor grade, and inhibits tumor cell growth *in vitro* and *in vivo*. *Oncogene*, 19(43):4979–4987, 2000.
33. S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30:41–47, 2002.
34. G. E. Arnold, A. K. Dunker, S. L. Johns, and R. J. Douthart. Use of conditional probabilities for determining relationships between amino acid sequence and protein secondary structure. *Proteins*, 12(4):382–399, 1992.
35. M. Ashburner et al. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.
36. W. R. Atchley and W. M. Fitch. A natural classification of the basic helix-loop-helix class of transcription factors. *Proc. Natl. Acad, Sci. USA*, 94:5172–5176, 1997.
37. T. K. Attwood, M. J. Blythe, D. R. Flower, A. Gaulton, J. E. Mabey, N. Maudling, L. McGregor, A. L. Mitchell, G. Moulton, K. Paine, and P. Scordis. PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Research*, 30:239–241, 2002.
38. S. Audic and J.-M. Claverie. Detection of eukaryotic promoters using Markov transition matrices. *computer & Chemistry*, 21(4):223–227, 1997.
39. T. B. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs. *Intelligent Systems for Molecular Biology*, 2:28–36, 1994.
40. A. Bairoch. The ENZYME database in 2000. *Nucleic Acids Research*, 28:304–305, 2000.
41. A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1):45–48, 2000.
42. I. V. Bajić. *Digital Signal Processing Techniques in the Analysis of DNA/RNA and Protein Sequences*. Bsceng thesis, University of Natal, Durban, South Africa, 1998.
43. V. B. Bajić and I. V. Bajić. *ANN in DNA regulatory region recognitions: The case of promoters*. International Joint Conference on Neural Networks, Washington, DC, tutorial, cd edition, July 1999.
44. V. B. Bajić and I. V. Bajić. Neural network system for promoter recognition. In *Future Directions for Intelligent Systems and Information Science*, chapter 14, pages 288–

305. Physica-Verlag, 2000.
45. V. B. Bajić, I. V. Bajić, and W. Hide. A new method of spectral analysis of DNA/RNA and protein sequences. In *Proceedings of 1st International Conference on Bioinformatics of Genome Regulation and Structure*, volume 1, pages 120–123, 1998.
46. V. B. Bajić and S. H. Seah. Dragon Gene Start Finder: An advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Research*, 13:1923–1929, 2003.
47. V. B. Bajić and S. H. Seah. Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes. *Nucleic Acids Research*, 31(13):3560–3563, 2003.
48. V. B. Bajić, S. H. Seah, A. Chong, S. P. T. Krishnan, J. L. Y. Koh, and V. Brusic. Computer model for recognition of functional transcription start sites in polymerase II promoters of vertebrates. *Journal of Molecular Graphics & Modeling*, 21(5):323–332, 2003.
49. V. B. Bajić, A. Chong, S. H. Seah, and V. Brusic. Intelligent system for vertebrate promoter recognition. *IEEE Intelligent Systems*, 17(4):64–70, 2002.
50. V. B. Bajić. Comparing the success of different prediction software in sequence analysis: A review. *Briefings in Bioinformatics*, 1(3):214–228, 2000.
51. V. B. Bajić, S. H. Seah, A. Chong, G. Zhang, J. L. Y. Koh, and V. Brusic. Dragon Promoter Finder: Recognition of vertebrate RNA polymerase II promoters. *Bioinformatics*, 18(1):198–199, 2002.
52. P. G. Baker and A. Brass. Recent development in biological sequence databases. *Current Opinion in Biotechnology*, 9:54–58, 1998.
53. P. Baldi. Computing with arrays of bell shaped and sigmoid functions. In *Proceedings of IEEE Neural Information Processing Systems*, volume 3, pages 728–734, 1990.
54. P. Baldi, S. Brunak, Y. Chauvin, and A. Krogh. Hidden Markov models for human genes: Periodic patterns in exon sequences. In *Theoretical and Computational Methods in Genome Research*, pages 15–32, 1997.
55. P. Baldi and Y. Chauvin. Hidden Markov models of G-protein-coupled receptor family. *Journal of Computational Biology*, 1:311–335, 1994.
56. P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, 1999.
57. P. Baldi and S. Brunak. *Bioinformatics: Adaptive Computation and Machine learning*. MIT Press, 2nd edition, 2001.
58. P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.
59. J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–822, 1993.
60. R. E. Banks, M. J. Dunn, D. F. Hochstrasser, J. C. Sanchez, W. Blackstock, D. J. Pappin, and P. J. Selby. Proteomics: New perspectives, new biomedical opportunities. *Lancet*, 356:1749–1756, 2000.
61. H. Bannai, Y. Tamada, O. Maruyama, K. Nakai, and S. Miyano. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, 18(2):298–305, 2002.
62. W. C. Barker, F. Pfeiffer, and D. G. George. Superfamily classification in PIR international protein sequence database. *Methods in Enzymology*, 266:59–71, 1996.
63. O. A. Bashkirov, E. M. Braverman, and I. B. Muchnik. Theoretical foundations of the

potential function method in pattern recognition learning. *Automation and Remote Control*, 25:629–631, 1964.

64. A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Research*, 30(1):276–280, 2002.
65. A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K. L. Howe, and E. L. L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Research*, 28:263–266, 2000.
66. A. Bateman, E. Birney, R. Durbin, S. R. Eddy, R. D. Finn, and E. L. L. Sonnhammer. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Research*, 27(1):260–262, 1999.
67. A. Bauer and B. Kuster. Affinity purification-mass spectrometry: Powerful tools for the characterization of protein complexes. *European Journal of Biochemistry*, 270(4):570–578, 2003.
68. S. M. Baxter and J. S. Fetrow. Sequence- and structure-based protein function prediction from genomic information. *Current Opinion in Drug Discovery & Development*, 4:291–295, 2001.
69. R. J. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large dense databases. In *Proceedings of 15th International Conference on Data Engineering*, pages 188–197, 1999.
70. R. J. Bayardo. Efficiently mining long patterns from databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 85–93, 1998.
71. C. R. Beals, C. M. Sheridan, C. W. Turck, P. Gardner, and G. R. Crabtree. Nuclear export of nf-atc enhanced by glycogen synthase kinase-3. *Science*, 275:1930–1933, 1997.
72. E. Beaudoing, S. Freier, J. R. Wyatt, J.-M. Claverie, and D. Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome Research*, 10:1001–1010, 2000.
73. A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6:281–297, 1999.
74. A. Bender and J. R. Pringle. Use of a screen for synthetic lethal and multicopy suppressee mutants to identify two new genes involved in morphogenesis in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology*, 11(3):1295–1305, 1991.
75. C. Benoist, K. O'Hare, R. Breathnach, and P. Chambon. The ovalbumin gene—sequence of putative control regions. *Nucleic Acids Research*, 8:127–142, 1980.
76. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. GenBank. *Nucleic Acids Research*, 28:15–18, 2000.
77. D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. Genbank. *Nucleic Acids Research*, 30:17–20, 2002.
78. D. Benton. Recent changes in the GenBank on-line services. *Nucleic Acids Research*, 18(6):1517–1520, 1990.
79. C. E. Bessey. The phylogenetic taxonomy of flowering plants. *Annals of the Missouri Botanical Garden*, 2:109–164, 1915.
80. J. C. Bezdek. *Fuzzy Mathematics in Pattern Classification*. Cornell University, Ithaca, N. Y., 1973.
81. A. P. Bird, M. H. Taggart, R. D. Nicholas, and D. R. Higgs. Non-methylated CpG-

rich islands at the human alpha-globin locus: Implications for evolution of the alpha-globin pseudogene. *EMBO Journal*, 6:999–1004, 1986.
82. C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.
83. M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406:536–540, 2000.
84. C. L. Blake and C. J. Marz. UCI machine learning repository, 1998. See `http://www.ics.uci.edu/~mlearn/MLRepository.html`.
85. T. Blumenthal. Gene clusters and polycistronic transcription in eukaryotes. *Bioessays*, 20(6):480–487, 1998.
86. B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.
87. M. S. Boguski, T. M. J. Lowe, and C. M. Tolstoshev. dbEST—database for "expressed sequence tags". *Nature Genetics*, 4:332–333, 1993.
88. M. S. Boguski and C. M. Tolstoshev. Gene discovery in dbEST. *Science*, 265:1993–1994, 1994.
89. 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(12):1213–1223, 2001.
90. M. F. Bonaldo, G. Lennon, and M. B. Soares. Normalization and substraction: Two approaches to facilitate gene discovery. *Genome Research*, 6:791–806, 1996.
91. P. Bork and E. Koonin. Ready for a motif submission? a proposed checklist. *Trends in Biochemical Sciences*, 20:104, 1995.
92. P. Bork and E. V. Koonin. Predicting functions from protein sequences—where are the bottlenecks? *Nature Genetics*, 18:313–318, 1998.
93. P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan. Predicting function: From genes to genomes and back. *Journal of Molecular Biology*, 283:707–725, 1998.
94. M. Borodovsky and J. D. McIninch. GENEMARK: Parallel gene recognition for both DNA strands. *Computers and Chemistry*, 17(2):123–133, 1993.
95. M. Borodovsky, J. D. McIninch, E.V. Koonin, K.E. Rudd, C. Medigue, and A. Danchin. Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Research*, 23:3554–3562, 1995.
96. N. K. Bose and P. Liang. *Neural Network Fundamentals with Graphs, Algorithms, and Applications*. McGraw-Hill, 1996.
97. H. Bourlard and N. Morgan. Continuous speech recognition by connectionist statistical methods. *IEEE Transactions on Neural Networks*, 4(6):893–909, 1993.
98. P. S. Bradley, U. M. Fayyad, and C. A. Reina. Scaling EM (expectation-maximization) clustering to large databases. Technical Report MSR-TR-98-35, Microsoft Research, November 1998.
99. R. K. Brayton, S. W. Director, G. D. Hachtel, and L. Vidigal. A new algorithm for statistical circuit design based on quasi-Newton methods and function splitting. *IEEE Transactions on Circuits and Systems*, CAS-26:784–794, 1979.
100. P. Brazhnik, A. de la Fuente, and P. Mendes. Gene networks: How to put the function

in genomics. *Trends in Biotechnology*, 20:467–472, 2002.

101. L. Breiman, L. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.
102. B. Brejová, D. Brown, and T. Vinař. Optimal spaced seeds for hidden Markov models, with application to homologous coding regions. In *Proceedings of 14th Annual Symposium on Combinatorial Pattern Matching*, pages 42–54, 2003.
103. B. Brejová, D. Brown, and T. Vinař. Vector seeds: An extension to spaced seeds allows substantial improvements in sensitivity and specificity. In *Proceedings of 3rd Annual Workshop on Algorithms in Bioinformatics*, pages 39–54, 2003.
104. S. E. Brenner. Errors in genome annotation. *Trends in Genetics*, 15(4):132–133, 1999.
105. S. Brin, R. Motwani, and C. Silverstein. Beyond market basket: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2:39–68, 1998.
106. S. Brin, R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of ACM-SIGMOD International Conference on Management of Data*, pages 255–264, 1997.
107. R. J. Brooker. *Genetics: Analysis and Principles*. Addison-Wesley, 1999.
108. B. R. Brooks, R. E. Bruccokeri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.
109. D. Broomhead and D. Lowe. Multivariable function interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.
110. M. Brown and C. Wilson. RNA pseudoknot modeling using intersections of stochastic context free grammars with applications to database search. In *Proceedings of Pacific Symposium on Biocomputing*, pages 109–125, 1996.
111. M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, 97(1):262–267, 2000.
112. N. P. Brown, C. Leroy, and C. Sander. MView: A web-compatible database search for multiple alignment viewer. *Bioinformatics*, 14:380–381, 1998.
113. R. K. Brummitt, editor. *Vascular Plant Families and Genera*. Royal Botanical Gardens, Kew, England, 1992.
114. S. J. Brunak, J. Engelbrecht, and S. Knudsen. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology*, 220:49–65, 1991.
115. P. Bucher and E. N. Trifonov. Compilation and analysis of eukaryotic Pol II promoter sequences. *Nucleic Acids Research*, 14:10009–10026, 1986.
116. P. Bucher. Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of Molecular Biology*, 212:563–578, 1990.
117. J. Buhler. Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics*, 17:419–428, 2001.
118. J. Buhler, U. Keich, and Y. Sun. Designing seeds for similarity search in genomic DNA. In *Proceedings of 7th Annual International Conference on Computational Biology*, pages 67–75, 2003.

119. J. Buhler and M. Tompa. Finding motifs using random projections. In *Proceedings of 5th Annual International Conference on Computational Biology*, pages 69–76, 2001.
120. F. R. Burden and D. A. Winkler. A quantitative structure-activity relationships model for the acute toxicity of substituted benzenes to *tetrahymena pyriformis* using Bayesian-regularised neural networks. *Chemical Research in Toxicology*, 13:436–440, 2000.
121. C. Burge. *Identification of genes in human genomic DNA*. PhD thesis, Stanford University, Stanford, CA, 1997.
122. C. Burge and S. Karlin. Finding the genes in genomic DNA. *Current Opinion on Structural Biology*, 8:346–354, 1998.
123. C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94, 1997.
124. C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.
125. S. K. Burley and R. G. Roeder. Biochemistry and structural biology of transcription factor IID (TFIID). *Annual Review of Biochemistry*, 65:769–799, 1996.
126. D. M. Burns, V. Horn, J. Paluh, and C. Yanofsky. Evolution of the tryptophan synthetase of fungi: Analysis of experimentally fused *Escherichia coli* tryptophan synthetase alpha and beta chains. *Journal of Biological Chemistry*, 265(4):2060–2069, 1990.
127. M. Burset and R. Guigo. Evaluation of gene structure prediction programs. *Genomics*, 34:353–367, 1996.
128. C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*, 282(5396):2012–2018, 1998.
129. A. Caccone and J. R. Powell. DNA divergence among hominoids. *Evolution*, 43:925–942, 1989.
130. Y.-D. Cai and K.-C. Chou. Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. *Biochemical and Biophysical Research Communications*, 305:407–411, 2003.
131. Y.-D. Cai, X.-J. Liu, X. Xu, and K.-C. Chou. Support vector machines for prediction of protein subcellular location. *Molecular Cell Biology Research Communications*, 4:230–233, 2001.
132. C. R. Cantor. Orchestrating the human genome project. *Science*, 248:49–51, 1990.
133. M. Caria. *Measurement Analysis: An Introduction to the Statistical Analysis of Laboratory Data in Physics, Chemistry, and the Life Sciences*. Imperial College Press, 2000.
134. R. Casadio, P. Fariselli, G. Finocchiaro, and P. L. Martelli. Fishing new proteins in the twilight zone of genomes: The test case of outer membrane proteins in *escherichia coli* k12, *escherichia coli* o157:h7, and other gram-negative bacteria. *Protein Science*, 12:1158–1168, 2003.
135. M. Caudill. *Neural Networks Primer*. Miller Freeman Publications, 1989.
136. J. Cedano, J. A. Pérez-Ponsa, and E. Querol. Relation between amino acid composition and cellular location of proteins. *Journal of Molecular Biology*, 266(3):594–600, 1997.
137. S. M. Chabregas, D. D. Luche, and L. P. Farias. Dual targeting properties of the N-

terminal signal sequence of *arabidopsis thaliana* thi1 protein to mitochondria and chloroplasts. *Plant Molecular Biology*, 46:639–650, 2001.

138. D. M. Chao and R. A. Young. Activation without a vital ingredient. *Nature*, 383:119–120, 1996.
139. M. W. Chase, D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y.-L. Qui, K. A. Kron, J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Sytsma, H. J. Michaels, H. J. Kress, W. J. Karol, K. G. Clark, M. Hedrén, B. S. Gaut, R. K. Jansen, K.-J. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q.-Y. Xiang, G. M. Plunkett, P. S. Soltis, S. M. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learns, S. W. Graham, S. C. H. Barrett, S. Dayanandan, and V. Albert. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene rbcL. *Annals of the Missouri Botanical Garden*, 80:528–580, 1993.
140. Y. Chauvin. A back-propagation algorithm with optimal use of hidden units. *Advances in Neural Information Processing Systems*, 1:519–526, 1989.
141. Y. Chauvin and D. Rumelhart. *Backpropagation: Theory, Architectures, and Applications*. Lawrence Erlbaum, 1995.
142. M. Chee, R. Yang, E. Hubbell, A. Berno, X. C. Huang, D. Stern, J. Winkler, D. J. Lockhart, M. S. Morris, and S. P. Fodor. Accessing genetic information with high-density DNA arrays. *Science*, 274(5287):610–614, 1996.
143. C. H. Chen, editor. *Fuzzy Logic and Neural Network Handbook*. McGraw-Hill, 1996.
144. F. Chen, C. C. MacDonald, and J. Wilusz. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Research*, 23:2614–2620, 1995.
145. I.-M. A. Chen and Victor M. Markowitz. An overview of the object-protocol model (OPM) and OPM data management tools. *Information Systems*, 20(5):393–418, 1995.
146. J. Chen, N.-H. Chua, D. Strauss, and L. Wong. Extracting Kozak consensus sequence using Kleisli. In *Proceedings of 1st International Conference on Bioinformatics of Genome Regulation and Structure*, pages 218–223, 1998.
147. J. Chen, D. Strauss, and L. Wong. Using Kleisli to bring out features in BLASTP results. In *Genome Informatics 1998*, pages 102–111, 1998.
148. J. Chen, L. Zhang, and L. Wong. A protein patent query system powered by Kleisli. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 593–595, 1998.
149. P. P. S. Chen. The entity-relationship model: Toward a unified view of data. *ACM Transaction on Database Systems*, 1(1):9–36, 1976.
150. Q. Chen, G. Hertz, and G. D. Stormo. PromFD 1.0: A computer program that predicts eukaryotic pol II promoters using strings and IMD matrices. *Computer Applications in the Biosciences*, 13:29–35, 1997.
151. Q. Chen, G. Z. Hertz, and G. D. Stormo. MATRIX SERACH 1.0: A computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Computer Applications in the Biosciences*, 13:29–35, 1995.
152. S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function network. *IEEE Transactions on Neural Networks*, 2(2):302–309, 1991.
153. T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential

equations. In *Proceedings of Pacific Symposium on Biocomputing'99*, pages 29–40, 1999.
154. V. Cherkassky and F. Mulier. *Learning from Data*. John Wiley & Sons, 1998.
155. D. Cheung, J. Han, V. Ng, A. Fu, and Y. Fu. A fast distributed algorithm for mining association rules. In *Proceedings of 4th International Conference on Parallel and Distributed Information Systems*, pages 31–44, 1996.
156. G. Chiaromonte, V. B. Yap, and W. Miller. Scoring pairwise genomic sequence alignments. In *Proceedings of Pacific Symposium on Biocomputing*, pages 115–126, 2002.
157. R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.
158. K. P. Choi and L. Zhang. Sensitive analysis and efficient method for identifying optimal spaced seeds. *Journal of Computer and System Sciences*, 2003. To appear.
159. K.-C. Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. *PROTEINS: Structure, Function and Genetics*, 44:60, 2001.
160. K.-C. Chou. Using subsite coupling to predict signal peptides. *Protein Engineering*, 14(2):75–79, 2001.
161. K.-C. Chou and Y.-D. Cai. Using functional domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry*, 48:45765–45769, 2002.
162. S. Y. Chung and L. Wong. Kleisli, a new tool for data integration in biology. *Trends in Biotechnology*, 17(9):351–355, 1999.
163. P. Cirri, P. Chiarugi, G. Camici, G. Manao, G. Raugei, G. Cappugi, and G. Ramponi. The role of Cys12, Cys17, and Arg18 in the catalytic mechanism of low-m(r) cytosolic phosphotyrosine protein phosphatase phosphatase. *European Journal of Biochemistry*, 214:647–657, 1993.
164. M. G. Claros, S. Brunak, and G. von Heijne. Prediction of N-terminal protein sorting signals. *Current Opinion in Structural Biology*, 7:394–398, 1997.
165. M. G. Claros and P. Vincens. Computational method to predict mitochondrially imported proteins and their targeting sequences. *European Journal of Biochemistry*, 241:779–786, 1996.
166. J. Claverie. Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics*, 6(10):1735–1744, 1997.
167. J. Claverie, O. Poirot, and F. Lopez. The difficulty of identifying genes in anonymous vertebrate sequences. *Computer & Chemistry*, 21(4):203–214, 1997.
168. J. Claverie and I. Sauvaget. Assessing the biological significance of primary structure consensus patterns using sequence databanks: I. Heat-shock and glucocorticoid control elements in eukaryotic promoters. *Computer Applications in the Biosciences*, 1:95–104, 1985.
169. D. Cochrane, C. Webster, G. Masih, and J. McCafferty. Identification of natural ligands for SH2 domains from a phage display cDNA library. *Journal of Molecular Biology*, 297(1):89–97, 2000.
170. E. F. Codd. A relational model for large shared data bank. *Communications of the ACM*, 13(6):377–387, 1970.
171. F. Collins and D. Galas. A new five-year plan for the U.S. Human Genome Project.

Science, 262:43–46, 1993.

172. R. R. Copley, T. Doerks, I. Letunic, and P. Bork. Protein domain analysis in the era of complete genomes. *FEBS Letters*, 513:129–134, 2002.
173. J. Corden, B. Wasylyk, A. Buchwalder, P. Sassone-Corsi, C. Kedinger, and P. Chambon. Promoter sequence of eukaryotic protein-coding genes. *Science*, 209:1406–1414, 1980.
174. F. Corpet and B. Michot. RNAlign program: Alignment of RNA sequences using both primary and secondary structure. *Computer Applications in the Biosciences*, 7:347–352, 1994.
175. F. Corpet, F. Servant, J. Gouzy, and D. Kahn. ProDom and ProDom-CG: Tools for protein domain analysis and whole-genome comparisons. *Nucleic Acids Research*, 28:267–269, 2000.
176. C. Creighton and S. Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79–86, 2003.
177. F. H. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin. General nature of the genetic code for proteins. *Nauchni trudove na Visshiia meditsinski institut, Sofiia*, 192:1227–1232, 1961.
178. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
179. A. Cronquist. *An Integrated System of Classification of Flowering Plants*. Columbia University Press, 1981.
180. S. H. Cross and A. P. Bird. CpG islands and genes. *Current Opinion in Genetics and Development*, 5:309–314, 1995.
181. S. H. Cross, V. H. Clark, and A. P. Bird. Isolation of CpG islands from large genomics clones. *Nucleic Acids Research*, 27:2099–2107, 1999.
182. M. Cserzö, F. Eisenhaber, B. Eisenhaber, and I. Simon. On filtering false positive transmembrane protein predictions. *Protein Engineering*, 15(9):745–752, 2002.
183. E. Dam, K. Pleij, and D. Draper. Structural and functional aspects of RNA pseudoknots. *Biochemistry*, 31(47):11665–11676, 1992.
184. T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: A fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324–328, 1998.
185. R. Das, J. Junker, D. Greenbaum, and M. B. Gerstein. Global perspectives on proteins: Comparing genomes in terms of folds, pathways and beyond. *Pharmacogenomics Journal*, 1:115–125, 2001.
186. S. Datta. Exploring relationships in gene expressions: A partial least squares approach. *Gene Expression*, 9:257–264, 2001.
187. S. Datta and S. Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459–466, 2003.
188. S. B. Davidson, C. Overton, and P. Buneman. Challenges in integrating biological data sources. *Journal of Computational Biology*, 2(4):557–572, 1995.
189. S. Davidson, C. Overton, V. Tannen, and L. Wong. BioKleisli: A digital library for biomedical researchers. *International Journal of Digital Libraries*, 1(1):36–53, 1997.
190. R. V. Davuluri, I. Grosse, and M. Q. Zhang. Computational identification of promoters and first exons in the human genome. *Nature Genetics*, 29(4):412–417, 2001.
191. M. O. Dayhoff. *Atlas of Protein Sequence and Structure, volumes 1–5, supplements*

1–3. National Biomedical Research Foundation, Washington, DC., 1965-1978.
192. M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5(Suppl. 3):345–352, 1978.
193. M. O. Dayhoff. The origin and evolution of protein superfamilies. *Federation Proceedings*, 35:2132–2138, 1976.
194. A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg. Alignment of whole genomes. *Nucleic Acids Research*, 27:2369–2376, 1999.
195. G. Deleage, C. Combet, C. Blanchet, and C. Geourjon. ANTHEPROT: An integrated protein sequence analysis software with client/server capabilities. *Computers in Biology and Medicine*, 31:259–267, 2001.
196. B. Demeler and G. W. Zhou. Neural network optimization for *E.coli* promoter prediction. *Nucleic Acids Research*, 19:1593–1599, 1991.
197. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.
198. J. Demsar, B. Zupan, M.W. Kattan, J.R. Beck, and I. Bratko. Naive Bayesian-based nomogram for prediction of prostate cancer recurrence. *Studies in Health Technolnology Informatics*, 68:436–441, 1999.
199. M. Deng, S. Mehta, F. Sun, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, 12(10):1540–1548, 2002.
200. J. M. Denu and J. E. Dixon. Protein tyrosine phosphatases: mechanisms of catalysis and regulation. *Current Opinion in Chemistry & Biology*, 2(5):633–641, 1998.
201. J. M. Denu, J. A. Stuckey, M. A. Saper, and J. E. Dixon. Form and function in protein dephosphorylation. *Cell*, 87:361–364, 1996.
202. S. J. DeRose. XQuery: A unified syntax for linking and querying general XML documents. In *Position Papers of QL'98—The Query Languages Workshop*, 1998.
203. D. Devos and A. Valencia. Intrinsic errors in genome annotation. *Trends in Genetics*, 17:429–431, 2001.
204. L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.
205. P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Mining the gene expression matrix: Inferring gene relationships from large scale gene expression data. In *Information Processing in Cells and Tissues*, pages 203–212. Plenum, 1998.
206. K. Dolinski et al. Saccharomyces genome database, 2003. See `http://www.yeastgenome.org`.
207. G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of 5th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 15–18, 1999.
208. G. Dong, J. Li, and X. Zhang. Discovering jumping emerging patterns and experiments on real datasets. In *Proceedings of 9th International Database Conference on Heterogeneous and Internet Databases*, pages 15–17, 1999.
209. G. Dong, X. Zhang, L. Wong, and J. Li. CAEP: Classification by aggregating emerging patterns. In *Proceedings of 2nd International Conference on Discovery Science*, pages 30–42, 1999.
210. S. Dong and D. B. Searls. Gene structure prediction by linguistic methods. *Genomics*, 162:705–708, 1994.

211. T. A. Down and T. J. Hubbard. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Research*, 12(3):458–461, 2002.
212. A. Drawid, R. Jansen, and M. Gerstein. Genome-wide analysis relating expression level with protein subcellular localization. *Trends in Genetics*, 16:426–430, 2000.
213. 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.
214. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.
215. R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, editors. *Biological Sequence Analysis: Probabilistics Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.
216. W. S. Dynan and R. Tjian. Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. *Nature*, 316:774–778, 1985.
217. B. A. Eckman, J. S. Aaronson, J. A. Borkowski, W. J. Bailey, K. O. Elliston, A. R. Williamson, and R. A. Blevins. The Merck Gene Index browser: An extensible data integration system for gene finding, gene characterization, and EST data mining. *Bioinformatics*, 14(1):2–13, 1998.
218. A. Economou. Bacterial secretome: The assembly manual and operating instructions. *Molecular Membrane Biology*, 19(3):159–169, 2002.
219. S. R. Eddy. Hidden Markov models. *Current Opinion in Structural Biology*, 6:361–365, 1996.
220. S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22:2079–2088, 1994.
221. S. R. Eddy, G. Mitchison, and R. Durbin. Maximum discrimination hidden Markov models of sequence consensus. *Journal of Computational Biology*, 2(9–23), 1995.
222. M. Edman, T. Jarhede, M. Sjöström, and A. Wieslander. Different sequence patterns in signal peptides from mycoplasmas, other gram-positive bacteria, and escherichia coli: A multivariate data analysis. *PROTEINS: Structure, Function, and Genetics*, 35:195–205, 1999.
223. A. Efstratiadis, J. W. Posakony, T. Maniatis, R. M. Lawn, C. O'Connell, R. A. Spritz, J. K. DeRiel, B. G. Forget, S. M. Weissman, J. L. Slightom, A. E. Blechl, O. Smithies, F. E. Baralle, C. C. Shoulders, and N. J. Proudfoot. The structure and evolution of the human beta-globin gene family. *Cell*, 21:653–668, 1980.
224. C. Ehresmann, F. Baudin, M. Mougel, P. Romby, J. P. Ebel, and B. Ehresmann. Probing the structure of RNAs in solution. *Nucleic Acids Research*, 15:9109–9112, 1987.
225. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Bostein. Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci. USA*, 95:14863–14868, 1998.
226. F. Eisenhaber, B. Eisenhaber, W. Kubina, S. Maurer-Stroh, G. Neuberger, G. Schneider, and M. Wildpaner. Prediction of lipid posttranslational modifications and localization signals from protein sequences: Big-Π, NMT and PTS1. *Nucleic Acids Research*, 31(13):3631–3634, 2003.
227. F. Eisenhaber and P. Bork. Wanted: Subcellular localization of proteins based on sequence. *Trends in Cell Biology*, 8:169–170, 1998.
228. F. Eisenhaber and P. Bork. Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, 15:528–535, 1999.

229. J. L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.
230. O. Emanuelsson, A. Elofsson, G. von Heijne, and S. Cristobal. In silico prediction of the peroxisomal proteome in fungi, plants, and animals. *Journal of Molecular Biology*, 330(2):443–456, 2003.
231. O. Emanuelsson, H. Nielsen, and G. von Heijne. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science*, 8(5):978–984, 1999.
232. O. Emanuelsson. Predicting protein subcellular localisation from amino acid sequence information. *Briefings in Bioinformatics*, 3(4):361–376, 2002.
233. A. J. Enright and C. A. Ouzounis. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biology*, 2(9):RESEARCH0034, 2001.
234. A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90, 1999.
235. T. Etzold and P. Argos. SRS: Information retrieval system for molecular biology data banks. *Methods in Enzymology*, 266:114–128, 1996.
236. B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 8(3):175–178, 1998.
237. S. Faisst and S. Meyer. Compilation of vertebrate encoded transcription factors. *Nucleic Acids Research*, 20:3–16, 1992.
238. L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. A. Sigrist, K. Hofmann, and A. Bairoch. The PROSITE database, its status in 2002. *Nucleic Acids Research*, 30(1):235–238, 2002.
239. J. B. Fant, C. D. Preston, and J. A. Barrett. Isozyme evidence of the parental origin and possible fertility of the hybrid *Potamogeton* x *fluitans Roth*. *Plant Systematics and Evolution*, 229(1/2):45–57, 2001.
240. R. Farber, A. Lapedes, and K. Sirotkin. Determination of eukaryotic protein coding regions using neural networks and information theory. *Journal of Molecular Biology*, 226:471–479, 1992.
241. L. Fausett. *Fundamentals of Neural Networks*. Prentice-Hall, 1994.
242. U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of 13th International Joint Conference on Artificial Intelligence*, pages 1022–1029, 1993.
243. B. Fazi, M. J. Cope, A. Douangamath, et al. Unusual binding properties of the SH3 domain of the yeast actin-binding protein Abp1: Structural and functional analysis. *Journal of Biological Chemistry*, 277(7):5290–5298, 2002.
244. Z.-P. Feng. An overview on predicting the subcellular location of a protein. *In Silico Biology*, 2(3):291–303, 2002.
245. Z.-P. Feng and C.-T. Zhang. Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. *International Journal of Biological Macromolecules*, 28:255–261, 2001.
246. M. F. Fernandez, A. Morishima, and D. Suciu. Efficient evaluation of XML middleware queries. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 103–114, 2001.
247. J. W. Fickett. Finding genes by computer: The state of the art. *Trends in Genetics*, 12(8):316–320, 1996.

248. J. W. Fickett and R. Guigo. Computational gene identification. In *Internet for Molecular Biologist*, pages 73–100. Horizon Scientific Press, 1996.
249. J. W. Fickett and A. G. Hatzigeorgiou. Eukaryotic promoter recognition. *Genome Research*, 7(9):861–878, 1997.
250. S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, 1989.
251. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.
252. D. Florescu and D. Kossmann. Storing and querying XML data using RDBMS. *Data Engineering Bulletin*, 22(3):27–34, 1999.
253. S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas. Light-directed spatially addressable parallel chemical synthesis. *Science*, 251:767–773, 1991.
254. D. B. Fogel. *Evolutionary Computation*. IEEE Press, 2nd edition, 2000.
255. P. L. Forey, C. J. Humphries, I. J. Kitching, R. W. Scotland, D. J. Siebert, and D. M. Williams. *Cladistics: A Practical Course in Systematics*. Clarendon Press, 1992.
256. V. Di Francesco, J. Granier, and P.J. Munson. Protein topology recognition from secondary structure sequences—applications of the hidden Markov models to the alpha class proteins. *Journal of Molecular Biology*, 267:446–463, 1997.
257. S. Frank, A. Lustig, T. Schulthess, J. Engel, and R. A. Kammerer. A distinct seven-residue trigger sequence is indispensable for proper coiled-coil formation of the human macrophage scavenger receptor oligomerization domain. *Journal of Biological Chemistry*, 275:11672–11677, 2000.
258. H. Franke, P. Hochschild, P. Pattnaik, and M. Snir. An efficient implementation of MPI on IBM SP1. In *Proceedings of 23rd Annual International Conference on Parallel Processing*, volume 3, pages 197–201, 1994.
259. K. Frech, P. Dietze, and T. Werner. ConsInspector 3.0: New library and enhanced functionality. *Computer Applications in the Biosciences*, 13:109–110, 1997.
260. K. Frech, G. Herrmann, and T. Werner. Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Research*, 21:1655–1664, 1993.
261. K. Frech, K. Quandt, and T. Werner. Muscle actin genes: A first step towards computational classification of tissue specific promoters. *In Silico Biology*, 1:29–38, 1998.
262. K. Frech and T. Werner. Specific modelling of regulatory units in DNA sequences. In *Proceedings of Pacific Symposium on Biocomputing*, pages 151–162, 1997.
263. N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyse expression data. *Journal of Computational Biology*, 7:601–620, 2000.
264. T.-T. Friess, N. Cristianini, and C. Campbell. The kernel adatron algorithm: A fast and simple learning procedure for support vector machines. In *Proceedings of 15th International Conference on Machine Learning*, 1998.
265. D. Frishman, A. Mironov, and M. Gelfand. Starts of bacterial genes: Estimating the reliability of computer predictions. *Gene*, 234:257–265, 1999.
266. K. J. Fryxell. The coevolution of gene family trees. *Trends in Genetics*, 12(9):364–369, 1996.
267. Y. Fujiwara, M. Asogawa, and K. Nakai. Prediction of mitochondrial targeting signals using hidden Markov models. In *Proceedings of 8th International Workshop on Genome Informatics*, pages 53–60, 1997.

268. Y. Fukunishi and Y. Hayashizaki. Amino-acid translation program for full-length cDNA sequences with frame-shift errors. *Physiological Genomics*, 5:81–87, 2001.
269. T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.
270. J. S. Garavelli, Z. Hou, N. Pattabiraman, and R. M. Stephens. The RESID database of protein structure modifications and the NRL-3D sequence-structure database. *Nucleic Acids Research*, 29:199–201, 2001.
271. M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *Journal of Molecular Biology*, 196:261–282, 1987.
272. J. L. Gardy, C. Spencer, K. Wang, M. Ester, G. E. Tusnády, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, and F. S. L. Brinkman. PSORT-B: Improving protein subcellular localization for gram-negative bacteria. *Nucleic Acids Research*, 31(13):3613–3617, 2003.
273. M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of $\mathcal{NP}$-completeness*. W. H. Freeman, 1979.
274. R. B. Gary and G. D. Stormo. Graph-theoretic approach to RNA modeling using comparative data. *Intelligent Systems for Molecular Biology*, 3:75–80, 1995.
275. A. P. Gasch, M. Huang, S. Metzner, D. Botstein, S. J. Elledge, and P. O. Brown. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. *Molecular Biology of the Cell*, 12:2987–3003, 2001.
276. A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.
277. A. Gasch and M. B. Eisen. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biology*, 3(11):research0059.1–0059.22, 2002.
278. A. C. Gavin, M. Bosche, R. Krause, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.
279. A. J. Gavin, T. E. Scheetz, C. A. Roberts, B. O'Leary, T. A. Braun, V. C. Sheffield, M. B. Soares, J. P. Robinson, and T. L. Casavant. Pooled library tissue tags for est-based gene discovery. *Bioinformatics*, 18(9):1162–1166, 2002.
280. H. Ge, Z. Liu, G. Church, and M. Vidal. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genetics*, 29(4):482–486, 2001.
281. J. Gehrke, V. Ganti, R. Ramakrishnan, and W. Y. Loh. BOAT—optimistic decision tree construction. In *Proceedings of ACM-SIGMOD International Conference on Management of Data*, pages 169–180, 1999.
282. M. S. Gelfand. Prediction of function in DNA sequence analysis. *Journal of Computational Biology*, 2(1):87–115, 1995.
283. M. S. Gelfand, A. A. Mironov, and P. A. Pezner. Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA*, 93:9061–9066, 1996.
284. F. W. Gembicki. *Vector optimization for Control with Performance and Parameter Sensitivity Indices*. PhD thesis, Case Western Reserve University, Cleveland, Ohio, 1974.

285. Gene Ontology Consortium. Creating the gene ontology resource: Design and implementation. *Genome Research*, 11:1425–1433, 2001.
286. H.-H. Gerdes and C. Kaether. Green fluorescent protein: Applications in cell biology. *FEBS Letters*, 389:44–47, 1996.
287. D. Ghosh. Status of the transcription factors database. *Nucleic Acids Research*, 21:2091–2093, 1993.
288. S. Ghosh and P. P. Majumder. Mapping a quantitative trait locus via the EM algorithm and Bayesian classification. *Genetic Epidemiology*, 19(2):97–126, 2000.
289. G. J. Gibson and C. F. N. Cowan. On the decision regions of multilayer perceptrons. *Proceedings of the IEEE*, 78(10):1590–1594, 1990.
290. W. Gish and D. J. States. Identification of protein coding regions by database similarity search. *Nature Genetics*, 3(3):266–272, 1993.
291. C. A. Goble, R. Stevens, G. Ng, S. Bechhofer, N. M. Paton, P. G. Baker, M. Peim, and A. Brass. Transparent access to multiple bioinformatics information sources. *IBM Systems Journal*, 40:532–552, 2001.
292. A. Goffeau, R. Aert, M. L. Agostini-Carbone, A. Ahmed, M. Aigle, L. Alberghina, K. Albermann, et al. The yeast genome directory. *Nature*, 387(6632 Supplement):5, 1997.
293. A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.
294. C. S. Goh, A. A. Bogan, M. Joachimiak, et al. Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, 299(2):283–293, 2000.
295. D. S. Goldfarb, J. Gariepy, G. Schoolnik, and R. D. Kornberg. Synthetic peptides as nuclear localization signals. *Nature*, 322:641–644, 1986.
296. E. Golemis. *Protein-Protein Interactions: A Molecular Cloning Manual*. Cold Spring Harbor Laboratory Press, 2002.
297. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Misirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(15):531–537, 1999.
298. S. M. Gomez and A. Rzhetsky. Towards the prediction of complete protein–protein interaction networks. In *Proceedings of Pacific Symposium on Biocomputing*, pages 413–424, 2002.
299. D. Gorlich. Nuclear protein import. *Current Opinion in Cell Biology*, 9(3):412–419, 1997.
300. O. Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982.
301. A. Grigoriev. A relationship between gene expression and protein interactions on the proteome scale: Analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 29(17):3513–3519, 2001.
302. U. Grob and K. Stuber. Recognition of ill-defined signals in nucleic acid sequences. *Computer Applications in the Biosciences*, 4:79–88, 1988.
303. S. Guha, R. Rastogi, and K. Shim. CURE: an efficient clustering algorithm for large databases. In *Proceedings of ACM-SIGMOD International Conference on Manage-*

ment of Data, pages 73–84, 1998.
304. R. Guigo. Computational gene identification: An open problem. *Computers & Chemistry*, 21(4):215–222, 1997.
305. R. Guigo. Assembling genes from predicted exons in linear time with dynamic programming. *Journal of Computational Biology*, 5:681–702, 1998.
306. R. Guigo, S. Knudsen, N. Drake, and T. Smith. Prediction of gene structure. *Journal of Molecular Biology*, 226:141–157, 1992.
307. A. P. Gultyaev, F. H. D. van Batenburg, and C. W. A. Pleij. The computer simulation of RNA folding pathways using a genetic algorithm. *Journal of Molecular Biology*, 250:37–51, 1995.
308. M. Gunduz, M. Ouchida, K. Fukushima, H. Hanafusa, T. Etani, S. Nishioka, K. Nishizaki, and K. Shimizu. Genomic structure of the human ING1 gene and tumor-specific mutations detected in head and neck squamous cell carcinomas. *Cancer Research*, 60:3143–3146, 2000.
309. D. Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, 1997.
310. S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mRNA abundance in yeast. *Molecular and Cellular Biology*, 19(3):1720–1730, 1999.
311. L. M. Haas, P. M. Schwarz, P. Kodali, E. Kotlar, J. E. Rice, and W. C. Swope. DiscoveryLink: A system for integrated access to life sciences data sources. *IBM Systems Journal*, 40(2):489–511, 2001.
312. G. Habeler, K. Natter, G. G. Thallinger, M. E. Crawford, S. D. Kohlwein, and Z. Trajanoski. Ypl.db: The yeast protein localization database. *Nucleic Acids Research*, 30:80–83, 2002.
313. D. H. Haft, B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen, and O. White. TIGRFAMS: A protein family resource for the functional identification of proteins. *Nucleic Acids Research*, 29:41–43, 2001.
314. S. Hahn, S. Buratowski, P. A. Sharp, and L. Guarente. Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus sequences. *Proc. Natl. Acad. Sci. USA*, 86:5718–5722, 1989.
315. J. M. Hall, M. K. Lee, B. Newman, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684–1689, 1990.
316. M. A. Hall. *Correlation-based feature selection machine learning*. PhD thesis, Department of Computer Science, University of Waikato, New Zealand, 1998.
317. R. Hamming. *Coding and Information Theory*. Prentice Hall, 1982.
318. A. Hamosh, A. F. Scott, J. Amberger, C Bocchini, V. Valle, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52–55, 2002.
319. J. Han and Y. Fu. Discovery of multiple-level association rules from large databases. In *Proceedings of 21st International Conference on Very Large Data Bases*, pages 420–431, 1995.
320. J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidates generation. In *Proceedings of ACM-SIGMOD International Conference on Management of Data*, pages 1–12, 2000.
321. S. Hannenhalli and S. Levy. Promoter prediction in the human genome. *Bioinformat-*

ics, 17:S90–S96, 2001.
322. R. Harr, M. Haggstrom, and P. Gustafsson. Search algorithm for pattern match analysis of nucleic acid sequences. *Nucleic Acids Research*, 11:2943–2957, 1983.
323. S. Harroch, G. C. Furtado, W. Brueck, J. Rosenbluth, J. Lafaille, M. Chao, J. D. Buxbaum, and J. Schlessinger. A critical role for the protein tyrosine phosphatase receptor type Z in functional recovery from demyelinating lesions. *Nature Genetics*, 32(3):411–414, 2002.
324. J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.
325. B. Hassibi and D. G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in Neural Information Processing Systems*, 5:164–172, 1993.
326. T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.
327. T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):607–616, 1996.
328. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.
329. A. Hatzigeorgiou, T. Harrer, N. Mache, and M. Reczko. The gene sequence analysis system DIANA. In *Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism*, pages 19–28. Wiley-VCH Verlag, 1995.
330. A. Hatzigeorgiou, N. Mache, and M. Reczko. Functional site recognition of the DNA sequence by artificial neural networks. In *Proceedings of IEEE International Joint Symposia on Intelligence and Systems*, pages 12–17, 1996.
331. A. Hatzigeorgiou, N. Mache, J. Wieland, M. Reczko, and A. Zell. Erkennung von promotoren und kodierenden bereichen in eukaryontischen genomischen sequenzen mit neuronalen netzen. In *Proceedings of Bioinformatik 94*, pages 70–74, 1994.
332. A. Hatzigeorgiou and M. Reczko. Recognition of protein coding regions and reading frames in DNA using neural networks. In *Proceedings of World Congress on Neural Networks*, volume 3, pages 136–138, 1995.
333. A. Hatzigeorgiou and M. Reczko. Gene identification with neural networks. In *Proceedings of Symposium on Control, Optimization, and Supervision*, volume 1, pages 140–143, 1996.
334. A. G. Hatzigeorgiou. *Mathematical models for feature recognition in nucleotide sequences*. Dr. rer. nat. dissertation, University of Jena, Germany, 2000.
335. A. G. Hatzigeorgiou. Translation initiation start prediction in human cDNAs with high accuracy. *Bioinformatics*, 18(2):343–350, 2002.
336. W. Hayes and M. Borodovsky. How to interpret anonymous genome? Machine learning approach to gene identification. *Genome Research*, 8:1154–1171, 1998.
337. T. R. Hazbun and S. Fields. Networking proteins in yeast. *Proc. Natl. Acad. Sci. USA*, 98(8):4277–4278, 2001.
338. R. Hecht-Nielsen. *Neurocomputing*. Addison-Wesley, 1990.
339. D. Heckerman. Bayesian networks for knowledge discovery. In *Advances in Knowledge Discovery and Data Mining*, pages 273–305, 1996. MIT Press.
340. C. Helbing, C. Veilette, K. Riabowol, R. N. Johnston, and I. Garkavetsev. A novel

candidate tumor suppressor, ING1, is involved in the regulation of apoptosis. *Cancer Research*, 57:1255–1258, 1997.

341. J. Henderson, S. Salzberg, and K. Fasman. Finding genes in human DNA with a hidden markov model. *Journal of Computational Biology*, 4(2):119–126, 1997.
342. R. Henderson and D. Tweten. Portable batch system: External reference specification. Technical report, NASA Ames Research Center, 1996.
343. J. G. Henikoff, E. A. Greene, S. Pietrovski, and L. Henikoff. Increased coverage of protein families with the Blocks database servers. *Nucleic Acids Research*, 28:228–230, 2000.
344. S. Henikoff. Scores for sequence searches and alignments. *Current Opinion in Structural Biology*, 6:353–360, 1996.
345. S. Henikoff, E. A. Greene, S. Pietrokovski, P. Bork, T. K. Attwood, and L. Hood. Gene families: The taxonomy of protein paralogs and chimeras. *Science*, 278:609–614, 1997.
346. S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89(22):10915–10919, 1992.
347. S. Henikoff and J.G. Henikoff. Automated assembly of protein blocks for database searching. *Nucleic Acids Research*, 19:6565–6572, 1991.
348. D. Hennessy, B. Buchanan, D. Subramanian, P. A. Wilkosz, and J. M. Rosenberg. Statistical methods for the objective design of screening procedures for macromolecular crystallization. *Acta Crystallogr. D Biol. Crystallogr.*, 56(7):817–827, 2000.
349. W. Hennig. *Phylogenetic Systematics*. University of Illinois Press, Urbana, Illinois, 1966.
350. J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*. Adison-Wesley, 1991.
351. G. R. Hicks and N. V. Raikhel. Protein import into the nucleus: an integrated view. *Annual Review of Cell and Developmental Biology*, 11:155–188, 1995.
352. D. M. Hillis and J. J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42:182–192, 1993.
353. T. Hirokawa, B.-C. Seah, and S. Mitaku. Sosui: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14(4):378–379, 1998.
354. D. S. Hirschberg. Algorithms for the longest common subsequence problems. *Journal of the ACM*, 24(4):664–675, 1977.
355. L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18:1553–1561, 2002.
356. J. D. Hirst and M. J. Sternberg. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry*, 31:7211–7218, 1992.
357. H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 18(6):523–521, 2001.
358. Y. Ho, A. Gruhler, A. Heilbut, et al. Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415(6868):180–183, 2002.
359. D. S. Hochbaum. Approximating covering and packing problems. In D. S. Hochbaum, editor, *Approximation Algorithms for NP-hard Problems*, chapter 3, pages 94–143. PWS, 1997.

360. K. M. Hoffmann, N. K. Tonks, and D. Barford. The crystal structure of domain 1 of receptor protein-tyrosine phosphatase mu. *Journal of Biological Chemistry*, 272(44):27505–27508, 1997.
361. K. Hofmann, P. Bucher, L. Falquet, and A. Bairoch. The PROSITE database, its status in 1999. *Nucleic Acids Research*, 27(1):215–219, 1999.
362. M. C. Honeyman, V. Brusic, N. Stone, and L. C. Harrison. Neural network-based prediction of candidate T-cell epitopes. *Nature Biotechnology*, 16(10):966–969, 1998.
363. S. B. Hoot, S. Magallón, and P. R. Crane. Phylogeny of basal eudicots based on three molecular data sets: atpB, *rbcL*, and 18S nuclear ribosomal DNA sequences. *Annals of the Missouri Botanical Garden*, 86:1–32, 1999.
364. S. Horai, K. Hayasaka, R. Kondo, K. Tsuganes, and N. Takahata. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. USA*, 92:532–536, 1995.
365. P. B. Horton and M. Kanehisa. An assessment of neural network and statistical approaches for prediction of *E.coli* promoter sites. *Nucleic Acids Research*, 20:4331–4338, 1992.
366. P. Horton. *String Algorithms and Machine Learning Applications for Computational Biology*. PhD thesis, University of California-Berkeley, Berkeley, CA, 1997.
367. P. Horton and K. Nakai. A probabilitistic classification system for predicting the cellular localization sites of proteins. *Intelligent Systems for Molecular Biology*, 4:109–115, 1996.
368. P. Horton and K. Nakai. Better prediction of protein cellular localization sites with the k nearest neighbours classifier. *Intelligent Systems for Molecular Biology*, 5:147–152, 1997.
369. S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.
370. H. Huang, C. Xiao, and C. H. Wu. ProClass protein family database. *Nucleic Acids Research*, 28:273–276, 2000.
371. X. Huang, M. D. Adams, H. Zhou, and A. R. Kerlavage. A tool for analyzing and annotating genomic sequences. *Genomics*, 46:37–45, 1997.
372. T. Hubbard et al. The Ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.
373. T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.
374. W.-K. Huh, J. V. Falvo, L. G. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425:686–691, 2003.
375. A. D. Huitema, R. A. Mathot, M. M. Tibben, J. H. Schellens, S. Rodenhuis, and J. H. Beijnen. Validation of techniques for the prediction of carboplatin exposure: application of Bayesian methods. *Clinical Pharmacology & Therapeutics*, 67(6):621–630, 2000.
376. B. Hussain and M. R. Kabuka. A novel feature recognition neural network and its application to character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16:98–106, 1994.
377. G. B. Hutchinson. The prediction of vertebrate promoter regions using different hex-

amer frequency analysis. *Computer Applications in the Biosciences*, 12:391–398, 1996.
378. G. B. Hutchinson and M. R. Hayden. The prediction of exons through an analysis of spliceable open reading frames. *Nucleic Acids Research*, 20:3453–3462, 1992.
379. J. Hutchinson. *The Families of Flowering Plants*. Clarendon Press, 2nd edition, 1959.
380. T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(Suppl. 1):S233–S240, 2002.
381. S. Ieong, M. Y. Kao, T. W. Lam, W. K. Sung, and S. M. Yiu. Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs. In *Proceedings of 2nd IEEE International Symposium on Bioinformatics and Bioengineering*, pages 183–190, 2001.
382. P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of 30th Annual ACM Symposium on Theory of Computing*, pages 604–613, 1998.
383. I. P. Ioshikhes and M. Q. Zhang. Large-scale human promoter mapping using CpG islands. *Nature Genetics*, 26(1):61–63, 2000.
384. C. Iseli, C. V. Jongeneel, and P. Bucher. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Intelligent Systems for Molecular Biology*, 7:138–148, 1999.
385. M. Ishikawa. A structural learning algorithm with forgetting of link weights. Technical Report TR-90-7, Electrotechnical Laboratories, Tsukuba City, Japan, 1990.
386. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, 98(8):4569–4574, 2001.
387. L. M. Iyer, L. Aravind, P. Bork, K. Hofmann, A. R. Mushegian, I. B. Zhulin, and E. V. Koonin. Quod erat demonstrandum? the mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biology*, 2:research0051, 2001.
388. T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7((1-2)):95–114, 2000.
389. B. Jagla and J. Schuchhardt. Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites. *Bioinformatics*, 16(3):245–250, 2000.
390. R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole genome expression data with protein-protein interactions. *Genome Research*, 12(1):37–46, 2002.
391. J. W. Jarvik and C. A. Telmer. Epitope tagging. *Annual Review of Genetics*, 32:601–618, 1998.
392. R. Javahery, A. Khachi, K. Lo, B. Zenzie-Gregory, and S. T. Smale. DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Molecular and Cellular Biology*, 14:116–127, 1994.
393. F. V. Jensen. *An Introduction to Bayesian Networks*. Springer-Verlag, 1996.
394. F. V. Jensen. *Bayesian Networks and Decision Graphs*. Springer, 2001.
395. C. Ji, R. R. Snapp, and D. Psaltis. Generalizing smoothness constraints from discrete samples. *Neural Computation*, 2(2):188–197, 1990.
396. Z. Jia, D. Barford, A. J. Flint, and N. K. Tonks. Structural basis for phosphotyrosine

peptide recognition by protein tyrosine phosphatase 1B. *Science*, 268:1754–1758, 1995.

397. J. Jiang and H. J. Jacob. EbEST: An automatic tool using expressed sequence tags to delineate gene structure. *Genome Research*, 8(3):268–275, 1998.
398. G. H. John. *Enhancements to the Data Mining Process*. PhD thesis, Stanford University, 1997.
399. I. T. Jolliffe. *Principal Component Analysis*. Springer Verlag, Berlin, 1986.
400. D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, 8:275–282, 1992.
401. J. Jones, J. K. Field, and J. M. Risk. A comparative guide to gene prediction tools for the bioinformatics amateur. *International Journal of Oncology*, 20:697–705, 2002.
402. N. C. Jones, P. W. J. Rigby, and E. B. Ziff. Trans-acting protein factors and the regulation of eukaryotic transcription: lessons from studies on DNA tumor viruses. *Genes & Development*, 2:267–281, 1988.
403. C. P. Joshi, H. Zhou, X. Huang, and V. L. Chiang. Context sequences of translation initiation codon in plants. *Plant Molecular Biology*, 35:993–1001, 1997.
404. J. Jurka. RepBase Update: A database and an electronic journal of repetitive elements. *Trends in Genetics*, 9:418–420, 2000.
405. G. V. Kaas. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29:119–127, 1980.
406. K. Kaiser and M. Meisterernst. The human general co-factors. *Trends in Biochemical Sciences*, 21:342–345, 1996.
407. R. A. Kammerer, T. Schulthess, R. Landwehr, A. Lustig, J. Engel, U. Aebi, and M. O. Steinmetz. An autonomous folding unit mediates the assembly of two-stranded coiled coils. *Proc. Natl. Acad. Sci. USA*, 95:13419–13424, 1998.
408. Z. Kan, E. C. Rouchka, W. R. Gish, and D. J. States. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Research*, 115:889–900, 2001.
409. M. D. Kane, T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Research*, 28:4552–4557, 2000.
410. M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG database at GenomeNet. *Nucleic Acids Research*, 30(1):42–46, 2002.
411. N. Kaplan, A. Vaaknin, and M. Linial. PANDORA: Keyword-based analysis of protein sets by integration of annotation sources. *Nucleic Acids Research*, 31:5617–5626, 2003.
412. N. B. Karayiannis. Reformulated radial basis neural networks trained by gradient descent. *IEEE Transactions on Neural Networks*, 10(3):657–671, 1999.
413. S. Karlin and S. F. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA*, 90:5873–5877, 1993.
414. S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, 87:2264–2268, 1990.
415. E. D. Karnin. A simple procedure for pruning back-propagation trained neural networks. *IEEE Transactions on Neural Networks*, 1(2):239–242, 1990.

416. P. D. Karp. Database links are a foundation for interoperability. *Trends in Biotechnology*, 14:273–279, 1996.
417. P. D. Karp, M. Krummenacker, S. Paley, and J. Wagg. Integrated pathway-genome databases and their role in drug discovery. *Trends in Biotechnology*, 17:275–281, 1999.
418. G. Karypis, E.-H. Han, and V. Kumar. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *IEEE COMPUTER*, 32(8):68–75, 1999.
419. J. Kasanov, G. Pirozzi, A. J. Uveges, and B. K. Kay. Characterizing class I WW domains defines key specificity determinants and generates mutant domains with novel specificities. *Chemistry & Biology*, 8(3):231–241, 2001.
420. S. Kasif and A. L. Delcher. Modeling biological data and structure with probabilistic networks. In *Computational Methods in Molecular Biology*, pages 335–352. Elsevier, 1998.
421. L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, 1990.
422. J. Kaufmann and T. S. Smale. Direct recognition of initiator elements by a component of the transcription factor IID complex. *Genes & Development*, 8:821–829, 1994.
423. J. Kaufmann, C. P. Verrijzer, J. Shao, and S. T. Smale. CIF, an essential cofactor for TFIID-dependent initiator function. *Genes & Development*, 10:873–886, 1996.
424. H. Kawaji, C. Schönbach, Y. Matsuo, J. Kawai, Y. Okazaki, Y. Hayashizaki, and H. Matsuda. Exploration of novel motifs derived from mouse cDNA sequences. *Genome Research*, 12:367–378, 2002.
425. U. Keich, M. Li, B. Ma, and J. Tromp. On spaced seeds of similarity search. *Discrete Applied Mathematics*, 2003. To appear.
426. P. Kemmeren, N. L. van Berkum, J. Vilo, et al. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular Cell*, 9(5):1133–1143, 2002.
427. W. J. Kent. BLAT—the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, 2002.
428. W. J. Kent and A. M. Zahler. Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. *Genome Research*, 10(8):1115–1125, 2000.
429. B. Kerem, J. M. Rommens, J. A. Buchanan, et al. Identification of the cystic fibrosis gene: Genetic analysis. *Science*, 245(4922):1073–1080, 1989.
430. J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westerman, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679, 2001.
431. S. H. Kim, G. J. Suddath, G. J. Quigley, A. McPherson, J. L. Sussman, A. H. J. Wang, N. C. Seeman, and A. Rich. Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science*, 185:435–439, 1974.
432. R. D. King, M. Saqi, R. Sayle, and M. J. Sternberg. DSC: Public domain protein secondary structure predication. *Computer Applications in the Biosciences*, 13(4):473–474, 1997.
433. H. Kishino, T. Miyata, and M. Hasegawa. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 17:368–

376, 1990.
434. D. Kisman, B. Ma, and M. Li. tPatternHunter: Gapped, fast, and translated homology search. Manuscript, Bioinformatics Solutions Inc., 2003.
435. J. Kleffe, K. Hermann, W. Vahrson, B. Wittig, and V. Brendel. GeneGenerator—a flexible algorithm for gene prediction and its application to maize sequences. *Bioinformatics*, 14(3):232–243, 1998.
436. M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. Verkamo. Finding interesting rules from large sets of discovered association rules. In *Proceedings of 3rd International Conference on Information and Knowledge Management*, pages 401–408, 1994.
437. P. S. Klosterman, M. Tamura, S. R. Holbrook, and S. E. Brenner. SCOR: a structural classification of RNA database. *Nucleic Acids Research*, 30(1):392–394, 2002.
438. A. Klug and J. W. Schwabe. Protein motifs 5: Zinc fingers. *FASEB Journal*, 9(8):597–604, 1995.
439. S. Knudsen. Promoter2.0: For the recognition of Pol II promoter sequences. *Bioinformatics*, 15(5):356–361, 1999.
440. T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.
441. T. Kohonen. Learning vector quantization for pattern recognition. Technical Report TKK-F-A601, Helsinki University of Technology, 1986.
442. T. Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, 2nd edition, 1989.
443. T. Kohonen. Improved versions of learning vector quantizations. In *Proceedings of International Joint Conference on Neural Networks*, volume I, pages 545–550, 1990.
444. T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
445. T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, 2nd edition, 1997.
446. Y. V. Kondrakhin, A. E. Kel, N. A. Kolchanov, A. G. Romashchenko, and L. Milanesi. Eukaryotic promoter recognition by binding sites for transcription factors. *Computer Applications in the Biosciences*, 11:477–488, 1995.
447. D. A. Konings and R. R. Gutell. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA*, 1(6):559–574, 1995.
448. R. D. Kornberg. RNA polymerase II transcription control. *Trends in Biochemical Sciences*, 21:325–326, 1996.
449. Z. Kou, W. W. Cohen, and R. F. Murphy. Extracting information from text and image proteomics. In *Proceedings of 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics*, pages 2–9, 2003.
450. M. Kozak. An analysis of vertebrate mRNA sequences: Intimations of translational control. *The Journal of Cell Biology*, 115:887–903, 1991.
451. M. Kozak. Initiation of translation in prokaryotes and eukaryotes. *Gene*, 234:187–208, 1999.
452. M. Kozak. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Research*, 15:8125–8148, 1987.
453. A. Krogh. Gene finding: Putting the parts together. In *Guide to Human Genome Computing*, chapter 11, pages 261–274. Academic Press, 2nd edition, 1998.
454. A. Krogh. Two methods for improving performance of an HMM and their application

for gene finindg. *Intelligent Systems for Molecular Biology*, 5:179–186, 1997.

455. A. Krogh. An introduction to hidden Markov models for biological sequences. In *Computational Methods in Molecular Biology*, pages 45–62. Elsevier, 1998.
456. A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.
457. A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, 305:567–580, 2001.
458. N. X. Krueger, M. Streuli, and H. Saito. Structural diversity and evolution of human receptor-like protein tyrosine phosphatases. *EMBO Journal*, 9(10):3241–3252, 1990.
459. J. K. Kruschke. Creating local and distributed bottlenecks in hidden layers of back-propagation networks. In *Proceedings of 1988 Connectionist Models Summer School*, pages 120–126, 1988.
460. J. K. Kruschke. Improving generalization in back-propagation networks with distributed bottlenecks. In *Proceedings of International Joint Conference on Neural Networks*, volume I, pages 443–447, 1989.
461. M. Kukar, I. Kononenko, and T. Silvester. Machine learning in prognosis of the femoral neck fracture recovery. *Artificial Intelligence in Medicine*, 8(5):431–451, 1996.
462. A. Kumar and M. Snyder. Protein complexes take the bait. *Nature*, 415(6868):123–124, 2002.
463. A. Kumar, S. Agarwal, J. A. Heyman, S. Matson, M. Heidtman, S. Piccirillo, L. Umansky, A. Drawid, R. Jansen, Y. Liu, K.-H. Cheung, P. Miller, M. Gerstein, G. S. Roeder, and M. Snyder. Subcellular localization of the yeast proteome. *Genes & Development*, 16:707–719, 2002.
464. S. Kumar, K. Tamura, I. B. Jakobsen, and M. Nei. Mega: Molecular evolutionary genetics analysis, version 2.0, 2000. Published by authors. Pennsylvania State University, University Park; and Arizona State University, Tempe.
465. S. Y. Kung. *Digital Neural Networks*. Prentice-Hall, 1993.
466. J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *journal of Molecular Biology*, 157:105–132, 1982.
467. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):861–921, 2001.
468. W. H. Landschulz, P. F. Johnson, and S. L. McKnight. The leucine zipper: A hypothetical structure common to a new class of DNA binding proteins. *Science*, 240:1759–1764, 1988.
469. W. H. Landschulz, P. F. Johnson, and S. L. McKnight. The DNA binding domain of the rat liver nuclear protein C/EBP is bipartite. *Science*, 243:1681–1688, 1989.
470. K. J. Lang and A. H. Waibel. A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3:23–43, 1990.
471. P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifier. In *Proceedings of 10th National Conference on Artificial Intelligence*, pages 223–228. AAAI Press, 1992.
472. P. Langley. *Elements of Machine Learning*. Morgan Kaufmann, 1996.

473. A. S. Lapedes, C. Barnes, C. Burks, R. M. Farber, and K. M. Sirotkin. Applications of neural networks and other machine learning algorithms to DNA sequence analysis. In *Computers and DNA*, pages 157–182. Addison-Wesley, 1989.
474. F. Larsen, G. Gundersen, R. Lopez, and H. Prydz. CpG islands as gene markers in the human genome. *Genomics*, 13:1095–1107, 1992.
475. R. A. Laskowski. PDBsum: Summaries and analyses of PDB structures. *Nucleic Acids Research*, 29:221–222, 2001.
476. D. S. Latchman. *Eukaryotic Transcription Factors*. Academic Press, 1991.
477. R. P. Laura, A. S. Witt, H. A. Held, et al. The Erbin PDZ domain binds with high affinity and specificity to the carboxyl termini of delta-catenin and ARVCF. *Journal of Biological Chemistry*, 277(15):12906–12914, 2002.
478. S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19:191–201, 1995.
479. Y. Le Cun, J. S. Denker, and S. A. Solla. Optimal brain damage. *Advances in Neural Information Processing Systems*, II:598–605, 1990.
480. P. Leder and M. Nirenberg. RNA codewords and protein synthesis II: Nucleotide sequence of a valine RNA codeword. *Proc. Natl. Acad. Sci. USA*, 52:420–427, 1964.
481. S. D. Lee, D. W. Cheung, and B. Kao. Is sampling useful in data mining? A case in the maintenance of discovered association rules. *Data Mining and Knowledge Discovery*, 2:233–262, 1998.
482. S.-J. Lee and H.-L. Tsai. Pattern fusion in feature recognition neural networks for handwritten character recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 28(4):612–625, 1998.
483. P. Legrain, J. Wojcik, and J. M. Gauthier. Protein–protein interaction maps: A lead towards cellular functions. *Trends in Genetics*, 17(6):346–352, 2001.
484. L.-G. Lei, S.-S. Zhang, and Z.-Y. Yu. The karyotype and the evolution of *Gymnotheca*. *Acta Botanica Boreali-Occidentalis Sinica*, 11(6):41–46, 1991.
485. G. G. Lennon, C. Auffray, M. Polymeropoulos, and M. B. Soares. The I.M.A.G.E. Consortium: An integrated molecular analysis of genomes and their expression. *Genomics*, 33:151–152, 1996.
486. H. L. Levin. *Ancient Invertebrates and Their Living Relatives*. Prentice Hall, 1999.
487. M. Levitt and M. Gerstein. A unified statistical framework for sequence comparison and structural comparison. *Proc. Natl. Acad. Sci. USA*, 95:5913–5920, 1998.
488. S. Levy, L. Compagnoni, E. W. Myers, and G. D. Stormo. Xlandscape: The graphical display of word frequencies in sequences. *Bioinformatics*, 14:74–80, 1998.
489. S. Lewis, M. Ashburner, and M. G. Reese. Annotating eukaryote genomes. *Current Opinion in Structural Biology*, 10:349–354, 2000.
490. F. Li and G. D. Stormo. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, 17:1067–1076, 2001.
491. J. Li, G. Dong, and K. Ramamohanarao. DeEPs: Instance-based classification using emerging patterns. In *Proceedings of 4th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 191–200, 2000.
492. J. Li, H. Liu, J. R. Downing, A. E.-J. Yeoh, and L. Wong. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics*, 19:71–78, 2003.
493. J. Li, H. Liu, and L. Wong. A comparative study on feature selection and classifi-

cation methods using a large set of gene expression profiles. In *Proceedings of 13th International Conference on Genome Informatics*, pages 51–60, 2002.

494. J. Li, S.-K. Ng, and L. Wong. Bioinformatics adventures in database research. In *LNCS 2572: Proceedings of 9th International Conference on Database Theory*, pages 31–46, 2003.
495. J. Li, K. R., and G. Dong. The space of jumping emerging patterns and its incremental maintenance algorithms. In *Proceedings of 17th International Conference on Machine Learning*, pages 551–558, 2000.
496. J. Li and L. Wong. Corrigendum: Identifying good diagnostic genes or genes groups from gene expression data by using the concept of emerging patterns. *Bioinformatics*, 18:1407–1408, 2002.
497. J. Li and L. Wong. Geography of differences between two classes of data. In *Proceedings 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 325–337, 2002.
498. J. Li and L. Wong. Identifying good diagnostic genes or genes groups from gene expression data by using the concept of emerging patterns. *Bioinformatics*, 18:725–734, 2002.
499. J. Li and L. Wong. Solving the fragmentation problem of decision trees by discovering boundary emerging patterns. In *Proceedings of IEEE International Conference on Data Mining*, pages 653–656, 2002.
500. M. Li, B. Ma, D. Kisman, and J. Tromp. PatternHunter II: Highly sensitive and fast homology search. *Journal of Bioinformatics and Computational Biology*, 2003. To appear.
501. W.-H. Li. *Molecular Evolution*. Sinauer Associates, 1997.
502. H.-X. Liang. Karyomorphology of *Gymnotheca* and phylogeny of four genera in Saururaceae. *Acta Botanica Yunnanica*, 13(3):303–307, 1991.
503. H.-X. Liang. Study on the pollen morphology of Saururaceae. *Acta Botanica Yunnaninca*, 14(4):401–404, 1992.
504. H.-X. Liang. On the systematic significance of floral organogenesis in Saururaceae. *Acta Phytotaxonomica Sinica*, 32(5):425–432, 1994.
505. H.-X. Liang. On the evolution and distribution in Saururaceae. *Acta Botanica Yunnanica*, 17(3):255–267, 1995.
506. H.-X. Liang and S. C. Tucker. Comparative study of the floral vasculature in Saururaceae. *American Journal of Botany*, 77:607–623, 1990.
507. H.-X. Liang and S. C. Tucker. Floral ontogeny of *Zippelia Begoniaefolia* and its familial affinity: Saururaceae or Piperaceae? *American Journal of Botany*, 82(5):681–689, 1995.
508. D.-I. Lin and Z. M. Kedem. Pincer-search: A new algorithm for discovering the maximum frequent set. In *Proceedings of 6th International Conference on Extending Database Technology*, pages 105–119, 1998.
509. K. Lin, A. E. Ting, J. Wang, and L. Wong. Hunting TPR domains using Kleisli. In *Proceedings of 9th International Workshop on Genome Informatics*, pages 173–182, 1998.
510. M. Linial. How incorrect annotation evolved—the case of short ORFs. *Trends in Biotechnology*, 21:298–300, 2003.
511. M. Linial and G. Yona. Methodologies for target selection in structural genomics.

Progress in Biophysics & Molecular Biology, 73:297–320, 2000.

512. D. Lipman and W. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.
513. R. P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, pages 4–22, 1987.
514. H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of IEEE 7th International Conference on Tools with Artificial Intelligence*, pages 338–391, 1995.
515. H. Liu and L. Wong. Data mining tools for biological sequences. *Journal of Bioinformatics and Computational Biology*, 1(1):139–168, 2003.
516. M. D. Lledó, P. O. Karis, M. B. Crespo, M. F. Fay, and M. W. Chase. Phylogenetic position and taxonomic status of the genus *Aegialitis* and subfamilies Staticoideae and Plumbaginoideae (Plumbaginaceae): Evidence from plastid DNA sequences and morphology. *Plant Systematics and Evolution*, 229:107–124, 2001.
517. L. Lo Conte, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin. SCOP database in 2002: Refinements accommodate structural genomics. *Nucleic Acids Research*, 30:264–267, 2002.
518. D. J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680, 1996.
519. W. Y. Loh and Y. S. Shih. Split selection methods for classification trees. *Statistica Sinica*, 7:815–840, 1997.
520. W. Y. Loh and N. Vanichsetakul. Tree-structured classification via generalized discriminant analysis. *Journal of American Statistical Association*, 83:715–728, 1988.
521. T. A. Longacre, M. H. Chung, D. N. Jensen, and M. R. Hendrickson. Proposed criteria for the diagnosis of well-differentiated endometrial carcinoma. A diagnostic test for myoinvasion. *American Journal of Surgical Pathology*, 19(4):371–406, 1995.
522. A. Loria and T. Pan. Domain structure of the ribozyme from eubacterial ribonuclease P. *RNA*, 2(6):551–563, 1996.
523. I. S. Lossos, R. Breuer, O. Intrator, and A. Lossos. Cerebrospinal fluid lactate dehydrogenase isoenzyme analysis for the diagnosis of central nervous system involvement in hematooncologic patients. *Cancer*, 88(7):1599–1604, 2000.
524. B. G. Louis and M. C. Ganoza. Signals determining translational start-site recognition in eukaryotes and their role in prediction of genetic reading frames. *Molecular Biology Reports*, 13:103–115, 1988.
525. T. M. Lowe and S. R. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5):955–964, 1997.
526. T. M. Lowe and S. R. Eddy. A computational screen for methylation guide snoRNAs in yeast. *Science*, 283:1168–1171, 1999.
527. L. Lu, H. Lu, and J. Skolnick. MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49(3):350–364, 2002.
528. A. V. Lukashin, V. V. Anshelevich, B. R. Amirikyan, A. I. Gragerov, and M. D. Frank-Kamenetskii. Neural network models for promoter recognition. *Journal of Biomolec-*

ular Structure & Dynamics, 6:1123–1133, 1989.

529. A. V. Lukashin and M. Borodovsky. GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research*, 26(4):1107–1115, 1998.
530. R. B. Lyngso and C. N. S. Pedersen. RNA pseudoknot prediction in energy-based models. *Journal of Computational Biology*, 7(3-4):409–427, 2000.
531. R. B. Lyngso, M. Zuker, and C. N. S. Pedersen. Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, 15(6):440–445, 1999.
532. T. J. Lyons, A. P. Gasch, L. A. Gaither, D. Botstein, P. O. Brown, and D. J. Eide. Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. *Proc. Natl. Acad. Sci. USA*, 97:7957–7962, 2000.
533. B. Ma, J. Tromp, and M. Li. PatternHunter: Faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, 2002.
534. G. MacBeath and S. L. Schreiber. Printing proteins as microarrays for high-throughput function determination. *Science*, 289(5485):1760–1763, 2000.
535. N. Mache and P. Levi. Detection of eukaryotic POL II promoters with multi-state time-delay neural network. In *IMISE Report No. 1: Proceedings of German Conference on Bioinformatics*, Leipzip, 1996. Institut fuer Medizinische Informatik, Statistik und Epidemilogie.
536. R. Mack and M. Hehenberger. Text-based knowledge discovery: Search and mining of life-sciences documents. *Drug Discovery Today*, 7(11 Supplement):S89–S98, 2002.
537. J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1967.
538. H. Mangalam. The Bio* toolkits—a brief overview. *Briefings in Bioinformatics*, 3(3):296–302, 2002.
539. O. L. Mangasarian, W. Nick Street, and W. H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4):570–577, 1995.
540. H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. In *Proceedings of AAAI'94 Workshop on Knowledge Discovery in Databases (KDD'94)*, pages 181–192, 1994.
541. E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285:751–753, 1999.
542. E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–86, 1999.
543. E. M. Marcotte, I. Xenarios, and D. Eisenberg. Mining literature for protein-protein interactions. *Bioinformatics*, 17(4):359–363, 2001.
544. E. Marshall and E. Pennisi. NIH launches the final push to sequence the genome. *Science*, 272(5259):188–189, 1996.
545. M. M. Martínez-Ortega and E. Rico. Seed morphology and its systematics significance in some *Veronica* species (Scrophulariaceae) mainly from the Western Mediterranean. *Plant Systematics and Evolution*, 228:15–32, 2001.
546. C. Mathe, M. F. Sagot, T. Schiex, and P. Rouze. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, 30(19):4103–4117, 2002.

547. C. K. Mathews and K. E. Van Holde. *Biochemistry*. Benjamin Cummings, 2nd edition, 1996.
548. D. M. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.
549. S. Matis, Y. Xu, M. Shah, X. Guan, J. R. Einstein, R. Mural, and E. Uberbacher. Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. *Computers & Chemistry*, 20:135–140, 1996.
550. H. Matsuda. Detection of conserved domains in protein sequences using a maximum-density subgraph algorithm. *IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Sciences*, E83-A:713–721, 2000.
551. H. Matsuda, T. Ishihara, and A. Hashimoto. Classifying molecular sequences using a linkage graph with their pairwise similarities. *Theoretical Computer Science*, 210:305–325, 1999.
552. L. R. Matthews, P. Vaglio, J. Reboul, et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, 11(12):2120–2126, 2001.
553. V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, et al. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, 2003.
554. L. A. McCue, K. A. McDonough, and C. E. Lawrence. Functional classification of cnmp-binding proteins and nucleotide cyclases with implications for novel regulatory pathways in mycobacterium tuberculosis. *Genome Research*, 10(2):204–219, 2000.
555. P. McGarvey, H. Huang, W. C. Barker, B. C. Orcutt, and C. H. Wu. The PIR website: New resource for bioinformatics. *Bioinformatics*, 16:290–291, 2000.
556. D. J. McGeoch. On the predictive recognition of signal peptide sequences. *Virus Research*, 3:271–286, 1985.
557. S. McKnight and R. Tjian. Transcriptional selectivity of viral genes in mammalian cells. *Cell*, 46:795–805, 1986.
558. M. Mehta, R. Agrawal, and J. Rissanen. SLIQ: A fast scalable classifier for data mining. In *Proceedings of International Conference on Extending Database Technology*, pages 18–32, 1996.
559. K. Melén, A. Krogh, and G. von Heijne. Reliability measures for membrane protein topology prediction algorithms. *Journal of Molecular Biology*, 327:735–744, 2003.
560. S.-W. Meng, Z.-D. Chen, D.-Z. Li, and H.-X. Liang. Phylogeny of Saururaceae inferred from matR sequence data. *Acta Botanica Sinica*, 43(6):653–656, 2001.
561. S.-W. Meng and H.-X. Liang. Comparative embryology on Saururaceae. *Acta Botanica Yunnanica*, 19(1):67–74, 1997.
562. K. M. L. Menne, H. Hermjakob, and R. Apweiler. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, 16:741–742, 2000.
563. H. W. Mewes, D. Frishman, C. Gruber abd B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schuller, S. Stocker, and B. Weil. MIPS: A database for genomes and protein sequences. *Nucleic Acids Research*, 28:37–40, 2000.
564. H. W. Mewes, D. Frishman, U. Guldener, et al. MIPS: A database for genomes and

protein sequences. *Nucleic Acids Research*, 30(1):31–34, 2002.

565. L. Milanesi, M. Muselli, and P. Arrigo. Hamming-clustering method for signal prediction in 5' and 3' regions of eukaryotic genes. *Computer Applications in the Biosciences*, 12:399–404, 1996.
566. L. Milanesi and I. Rogozin. Prediction of human gene structure. In *Guide to Human Genome Computing*, pages 215–259. Academic Press, 2nd edition, 1998.
567. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.
568. J. M. Mingot, E. A. Espeso, E. Díez, and M. A. Penalva. Ambient ph signaling regulates nuclear localization of the *Aspergillus nidulans* pacc transcription factor. *Molecular and Cellular Biology*, 21(5):1688–1699, 2001.
569. M. Minsky and S. Papert. *Perceptrons*. MIT Press, 1969.
570. P. J. Mitchell and R. Tjian. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245:371–245, 1998.
571. T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
572. K. Miyahara and F. Yoda. *Printed Japanese Character Recognition Based on Multiple Modified LVQ Neural Network*. IEEE Press, 1996.
573. B. Modrek, A. Resch, C. Grasso, and C. Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Research*, 29:2850–2859, 2001.
574. K. V. K. Mohan and C. D. Atreya. Novel organelle-targeting signals in viral proteins. *Bioinformatics*, 19:10–13, 2003.
575. S. Möller, M. D. R. Croning, and R. Apweiler. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, 17(7):646–653, 2001.
576. J. Moody and C. Darken. Learning with localized receptive fields. In *Proceedings of 1988 Connectionist Models Summer School*, pages 133–143, 1988.
577. J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.
578. D. P. Mortlock, P. Sateesh, and J. W. Innis. Evolution of N-terminal sequences of the vertebrate HOXA13 protein. *Mammalian Genome*, 11:151–158, 2000.
579. R. Mott. Maximum likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bulletin of Mathematical Biology*, 54:59–75, 1992.
580. J. Moult, K. Fidelis, A. Zemla, and T. Hubbard. Critical assessment of methods of protein structure prediction (CASP): Round IV. *Proteins*, Suppl. 5:2–7, 2001.
581. M. C. Mozer and P. Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. *Advances in Neural Information Processing*, 1:107–115, 1989.
582. N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. A. Sigrist, R. Vaughan, and E. M. Zdobnov. The interpro database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, 31:315–318, 2003.

583. A. Muller, R. M. MacCallum, and M. J. Sternberg. Benchmarking PSI-BLAST in genome annotation. *Journal of Molecular Biology*, 293:1257–1271, 1999.
584. M. E. Mulligan and W. R. McClure. Analysis of the occurrence of promoter-sites in DNA. *Nucleic Acids Research*, 14:109–126, 1986.
585. K. Mullis, F. Faloona, S. Scharf, et al. Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbour Symposium on Quantitative Biology*, 51(Pt. 1):263–273, 1986.
586. K. B. Mullis. The unusual origin of the polymerase chain reaction. *Scientific American*, 262(4):56–61, 64–5, 1990.
587. K. Murakami and T. Takagi. Gene recognition by combination of several gene-finding programs. *Bioinformatics*, 14(8):665–675, 1998.
588. V. Murino. Structured neural networks for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 28(4):553–561, 1998.
589. 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. *Intelligent Systems for Molecular Biology*, 8:251–259, 2000.
590. J. Murvai, K. Vlahovicek, E. Barta, and S. Pongor. The sbase protein domain library, release 8.0: A collection of annotated protein sequence segments. *Nucleic Acids Research*, 29(1):58–60, 2001.
591. M. T. Musavi, W. Ahmed, K. H. Chan, K. B. Faris, and D. M. Hummels. On the training of radial basis function classifiers. *Neural Networks*, 5(4):595–603, 1992.
592. T. Nagashima, D. Silva, L. Socha, N. Petrovsky, H. Suzuki, R. Saito, T. Kasukawa, I. Kurochkin, A. Konagaya, and C. Schönbach. Inferring higher functional information for RIKEN mouse full-length cDNA clones with FACTS. *Genome Research*, 13(6):1520–1533, 2003.
593. R. Nair, P. Carter, and B. Rost. NLSdb: Database of nuclear localization signals. *Nucleic Acids Research*, 31:397–399, 2003.
594. R. Nair and B. Rost. Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, 18(Suppl. 1):S78–S86, 2002.
595. R. Nair and B. Rost. Sequence conserved for subcellular localization. *Protein Science*, 11:2836–2847, 2002.
596. R. Nair and B. Rost. LOC3D: Annotate sub-cellular localization for protein structures. *Nucleic Acids Research*, 31:3337–3340, 2003.
597. K. Nakai and P. Horton. PSORT: A program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences*, 24:34–36, 1999.
598. K. Nakai. Refinement of the prediction methods of signal peptides for the genome analyses of *Saccharomyces cerevisiae* and *Bacillus subtilis*. In *Proceedings of 7th International Workshop on Genome Informatics*, pages 72–81, 1996.
599. K. Nakai. Protein sorting signals and prediction of subcellular localization. *Advances in Protein Chemistry*, 54:277–344, 2000.
600. K. Nakai. Prediction of *in vivo* fates of proteins in the era of genomics and proteomics. *Journal of Structural Biology*, 134:103–116, 2001.
601. K. Nakai and M. Kanehisa. Expert system for predicting protein localization sites in gram-negative bacteria. *PROTEINS: Structure, Function, and Genetics*, 11:95–110,

1991.

602. K. Nakai and M. Kanehisa. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, 14:897–911, 1992.
603. M. Nakao. Improved accuracy of PSORTII with feature selection and DANN, November 2003. Private Communication.
604. H. Nakashima and K. Nishikawa. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequences. *Journal of Molecular Biology*, 238:54–61, 1994.
605. K. Nakata, M. Kanehisa, and J. V. Maizel. Discriminant analysis of promoter regions in *Escherichia coli* sequences. *Computer Applications in the Biosciences*, 4:367–371, 1988.
606. Growth of genbank, 2002. Available at `www.ncbi.nlm.nih.gov/Genbank/genbankstats.html`.
607. 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*, 48:444–453, 1970.
608. S.-K. Ng and M. Wong. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Informatics*, 10:104–112, 1999.
609. S. K. Ng, Z. Zhang, and S. H. Tan. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, 19(8):923–929, 2003.
610. W. Nickel. The mystery of nonclassical protein secretion: A current view on cargo proteins and potential export routes. *European Journal of Biochemistry*, 270(10):2109–2119, 2003.
611. H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 10:1–6, 1997.
612. H. Nielsen and A. Krogh. Prediction of signal peptides and signal anchors by a hidden Markov model. *Intelligent Systems for Molecular Biology*, 6:122–130, 1998.
613. H. Nielsen. Hot papers in bioinformatics, interview by Eugene Russo. *The Scientist*, 13(13):8, 1999.
614. H. Nielsen. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Engineering*, 12(1):3–9, 1999.
615. J. Nilsson, S. Stahl, J. Lundeberg, et al. Affinity fusion strategies for detection, purification, and immobilization of recombinant proteins. *Protein Expression and Purification*, 11(1):1–16, 1997.
616. M. W. Nirenberg and J. H. Matthaei. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA*, 47:1588–1602, 1961.
617. K. Nishikawa and T. Ooi. Correlation of the amino acid composition of a protein to its structural and biological characters. *Journal of Biochemistry*, 91(5):1821–1824, 1982.
618. T. Nishikawa, T. Ota, and T. Isogai. Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. *Bioinformatics*, 16:960–967, 2000.
619. K. C. Nixon. Winclada (beta) ver. 0.9.9, 1999. Published by author. Ithaca, NY.
620. N. K. Gray and M. Wickens. Control of translation initiation in animals. *Annual*

Review of Cells & Developmental Biology, 14:399–458, 1998.

621. C. D. Novina and A. L. Roy. Core promoters and transcriptional control. *Trends in Genetics*, 9:351–355, 1996.
622. S. J. Nowlan and G. E. Hinton. Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4(4):473–493, 1992.
623. R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA*, 77(11):6309–6313, 1980.
624. R. Nussinov, J. Owens, and J. V. Maizel. Sequence signals in eukaryotic upstream regions. *Biochimica et Biophysica Acta*, 866:109–119, 1986.
625. N. Ogawa, J. DeRisi, and P. O. Brown. New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Molecular Biology of the Cell*, 11:4309–4321, 2000.
626. U. Ohler, S. Harbeck, H. Niemann, E. Noth, and M. G. Reese. Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics*, 15(5):362–369, 1999.
627. U. Ohler, G. C. Liao, H. Niemann, and G. M. Rubin. Computational analysis of core promoters in the Drosophila genome. *Genome Biology*, 3(12):RESEARCH0087, 2002.
628. U. Ohler, H. Niemman, G.-C. Liao, and G. M. Rubin. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, 17(Suppl 1):S199–S206, 2001.
629. U. Ohler and M. G. Reese. Detection of eukaryotic promoter regions using stochastic language models. In *Molekulare Bioinformatik*, pages 89–100. Shaker, 1998.
630. U. Ohler, G. Stemmer, S. Harbeck, and H. Niemann. Stochastic segment models of eukaryotic promoter regions. In *Proceedings of Pacific Symposium on Biocomputing*, pages 380–391, 2000.
631. H. Okada. Karyomorphology and relationship in some genera of Saururaceae and Piperaceae. *Botanical Magazine Tokyo*, 99:289–299, 1986.
632. T. Okamoto, T. Suzuki, and N. Yamamoto. Microarray fabrication with covalent attachment of DNA using bubble jet technology. *Nature Biotechnology*, 18(4):438–441, 2000.
633. S. Oliver. Guilt-by-association goes global. *Nature*, 403:601–603, 2000.
634. A. B. Olshen and A. N. Jain. Deriving quantitative conclusions from microarray expression data. *Bioinformatics*, 18(7):961–970, 2002.
635. M. V. Olson. The human genome project. *Proc. Natl. Acad. Sci. USA*, 90:4338–4344, 1993.
636. E. M. O'Neill, A. Kaffman, E. R. Jolly, and E. K. O'Shea. Regulation of pho4 nuclear localization by the pho80-pho85 cyclin-cdk complex. *Science*, 271:209–212, 1996.
637. M. C. O'Neill. Training back-propagation neural networks to define and detect DNA-binding sites. *Nucleic Acids Research*, 19(2):313–318, 1991.
638. M. C. O'Neill. Consensus methods for finding and ranking DNA binding sites: Application to *Escherichia coli* promoters. *Journal of Molecular Biology*, 207:301–310, 1989.
639. E. K. O'Shea, R. Rutkowski, and P. S. Kim. Evidence that the leucine zipper is a coiled coil. *Science*, 243:538–542, 1989.
640. A. O'Shea-Greenfield and S. T. Smale. Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription.

Journal of Biological Chemistry, 267(2):1391–1402, 1992.

641. R. Overbeek, M. Fonstein, M. D'Souza, et al. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, 96(6):2896–2901, 1999.
642. R. Overbeek, N. Larsen, G. D. Pusch, M. D'Souza, E. Selkov Jr., N. Kyrpides, M. Fonstein, N. Maltsev, and E. Selkov. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research*, 28:123–125, 2000.
643. T. Oyama, K. Kitano, K. Satou, and T. Ito. Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics*, 18(5):705–714, 2002.
644. S. Pages, A. Belaich, J. P. Belaich, et al. Species-specificity of the cohesin-dockerin interaction between *clostridium thermocellum* and *clostridium cellulolyticum*: Prediction of specificity determinants of the dockerin domain. *Proteins*, 29(4):517–527, 1997.
645. C. Papanicolaou, M. Gouy, and J. Ninio. An energy model that predicts the correct folding of both the tRNA and 5S RNA molecules. *Nucleic Acids Research*, 12:31–44, 1984.
646. J. Park, M. Chen, and P. Yu. An effective hash-based algorithm for mining association rules. In *Proceedings of ACM-SIGMOD International Conference on Management of Data*, pages 175–186, 1995.
647. J. Park, S. A. Teichmann, T. Hubbard, and C. Chothia. Intermediate sequences increase the detection of homology between sequences. *Journal of Molecular Biology*, 273:349–354, 1997.
648. K.-J. Park and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 19(13):1656–1663, 2003.
649. T. Park, S.-G. Yi, S. Lee, S. Y. Lee, D.-H. Yoo, J.-I. Ahn, and Y.-S. Lee. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, 19(6):694–703, 2003.
650. F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, 14(9):609–614, 2001.
651. F. M. G. Pearl, N. Martin, J. E. Bray, D. W. A. Buchan, A. P. Harrison, D. Lee, G. A. Reeves, A. J. Shepherd, I. Sillitoe, A. E. Todd, J. M. Thornton, and C. A. Orengo. A rapid classification protocol for the CATH domain database to support structural genomics. *Nucleic Acids Research*, 29:223–227, 2001.
652. J. Pearl. *Causality*. Cambridge University Press, 2000.
653. D. A. Pearlman, D. A. Case, J. C. Caldwell, G. L. Seibel, C. Singh, P. Weiner, and P. A. Kollman. *AMBER 4.0*. University of California, San Francisco, 1991.
654. W. R. Pearson. Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, 11:635–650, 1991.
655. W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.
656. A. G. Pedersen, P. Baldi, Y. Chauvin, and S. Brunak. The biology of eukaryotic promoter prediction—a review. *Computers & Chemistry*, 23:191–207, 1999.
657. A. G. Pedersen, P. Baldi, S. Brunak, and Y. Chauvin. Characterization of prokary-

otic and eukaryotic promoters using hidden Markov models. *Intelligent Systems for Molecular Biology*, 4:182–191, 1996.

658. A. Gorm Pedersen and H. Nielsen. Neural network prediction of translation initiation sites in eukaryotes: Perspectives for EST and genome analysis. *Intelligent Systems for Molecular Biology*, 5:226–233, 1997.
659. H. R. B. Pelham. The retention signal for soluble proteins of the endoplasmic reticulum. *Trends in Biochemical Sciences*, 15:482–486, 1990.
660. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, 96:4285–4288, 1999.
661. F. E. Penotti. Human DNA TATA boxes and transcription initiation sites. *Journal of Molecular Biology*, 213:37–52, 1990.
662. R. C. Perier, V. Praz, T. Junier, C. Bonnard, and P. Bucher. The eukaryotic promoter database (EPD). *Nucleic Acids Research*, 28:302–303, 2000.
663. C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, et al. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.
664. A. E. Pertiz, R. Kierzek, N. Sugimoto, and D. H. Turner. Thermodynamic study of internal loops in oligoribonucleotides: Symmetric loops are more stable than asymmetric loops. *Biochemistry*, 30:6428–5436, 1991.
665. G. Pesole, S. Liuni, G. Grillo, F. Licculli, A. Larizza, W. Makalowski, and C. Saccone. UTRdb and UTRsite: Specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. *Nucleic Acids Research*, 28:193–196, 2000.
666. J. D. Peterson, L. A. Umayam, T. Dickinson, E. K. Hickey, and O. White. The comprehensive microbial resource. *Nucleic Acids Research*, 29:123–125, 2001.
667. J. Platt. A resource-allocating network for function interpolation. *Neural Computation*, 3(2):213–225, 1991.
668. J. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods—Support Vector Learning*. MIT Press, 1998.
669. D. C. Plaut, S. J. Nowlan, and G. E. Hinton. Experiments on learning by back propagation. Technical Report CMU-CS-86-126, Carnegie-Mellon University, Pittsburgh, PA 15213, 1986.
670. T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.
671. R. M. Polhill, P. H. Raven, and C. H. Stirton. Evolution and systematics of the Leguminosae. In *Advances in Legume Systematics: Part I.*, pages 1–26. Royal Botanic Gardens, Kew, England, 1981.
672. M. H. Polymeropoulos, J. J. Higgins, L. I. Golbe, et al. Mapping of a gene for parkinson's disease to chromosome 4q21-q23. *Science*, 274(5290):1197–1199, 1996.
673. L. Ponger, L. Duret, and Mouchiroud. Determination of CpG islands: Expression in early embryo and isochore structure. *Genome Research*, 11:1854–1860, 2001.
674. L. Ponger and D. Mouchiroud. CpGProD: Identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, 18:631–633, 2002.
675. E. Portugaly, I. Kifer, and M. Linial. Selecting targets for structural determination by

navigating in a graph of protein families. *Bioinformatics*, 18:899–907, 2002.
676. M. J. D. Powell. Radial basis functions for multivariable interpolation: A review. In *Algorithms for Approximation*. Clarendon Press, 1987.
677. V. Praz, R. Perier, C. Bonnard, and P. Bucher. The eukaryotic promoter database, EPD: New entry types and links to gene expression data. *Nucleic Acids Research*, 30(1):322–324, 2002.
678. D. S. Prestridge. SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements. *Computer Applications in the Biosciences*, 7:203–206, 1991.
679. D. S. Prestridge. Predicting Pol II promoter sequences using transcription factor binding sites. *Journal of Molecular Biology*, 249:923–932, 1995.
680. D. S. Prestridge. SIGNAL SCAN 4.0: Additional databases and sequence formats. *Computer Applications in the Biosciences*, 12:157–160, 1996.
681. D. S. Prestridge. Computer software for eukaryotic promoter analysis: Review. *Methods in Molecular Biology*, 130:265–295, 2000.
682. D. S. Prestridge and C. Burks. The density of transcriptional elements in promoter and non-promoter sequences. *Human Molecular Genetics*, 2:1449–1453, 1993.
683. D. S. Prestridge and G. Stormo. SIGNAL SCAN 3.0: New database and program features. *Computer Applications in the Biosciences*, 9:113–115, 1993.
684. K.D. Pruitt and D.R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29:137–140, 2001.
685. M. Ptashne and A. Gann. Transcriptional activation by recruitment. *Nature*, 386:567–577, 1997.
686. C. H. Pui and W. E. Evans. Acute lymphoblastic leukemia. *New England Journal of Medicine*, 339:605–615, 1998.
687. N. Qian and T.J. Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, 202:865–884, 1988.
688. Y. L. Qiu, J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. The earliest angiosperms: evidence from mitochondrial, plastid, and nuclear genomes. *Nature*, 402:404–407, 1999.
689. K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner. MatInd and MatInspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research*, 23:4878–4884, 1995.
690. K. Quandt, K. Grote, and T. Werner. GenomeInspector: A new approach to detect correlation patterns of elements on genomic sequences. *Computer Applications in the Biosciences*, 12:405–413, 1996.
691. K. Quandt, K. Grote, and T. Werner. GenomeInspector: Basic software tools for analysis of spatial correlations between genomic structures within megabase sequences. *Genomics*, 33:301–304, 1996.
692. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
693. J. R. Quinlan. *C4.5: Program for Machine Learning*. Morgan Kaufmann, 1993.
694. G. Raddatz, M. Dehio, T.F. Meyer, and C. Dehio. PrimeArray: genome-scale primer design for DNA-microarray construction. *Bioinformatics*, 17:98–99, 2001.
695. J. C. Rain, L. Selig, H. De Reuse, et al. The protein-protein interaction map of *Helicobacter pylori*. *Nature*, 409(6817):211–215, 2001.

696. T. H. Rainer, P. K. Lam, E. M. Wong, and R. A. Cocks. Derivation of a prediction rule for post-traumatic acute lung injury. *Resuscitation*, 42(3):187–196, 1999.
697. S. Rampone. Recognition of splice junctions on DNA sequences by BRAIN learning algorithm. *Bioinformatics*, 14(8):676–684, 1998.
698. G. Ramsay. DNA chips: State-of-the art. *Nature Biotechnology*, 16:40–44, 1998.
699. R. Rastogi and K. Shim. Public: A decision tree classifier that integrates building and pruning. In *Proceedings of 24th International Conference on Very Large Data Bases*, pages 404–415, 1998.
700. S. Raudys. How good are support vector machines? *Neural Network*, 13(1):17–19, 2000.
701. R. Reed. Pruning algorithms—a survey. *IEEE Transactions on Neural Networks*, 4(5):740–747, 1993.
702. M. G. Reese. *Erkennung von Promotoren in pro- und eukaryontischen DNA-Sequenzen durch Künstliche Neuronale Netze*. Diploma work, University of Heidelberg, Germany, 1994.
703. M. G. Reese. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Computers & Chemistry*, 26(1):51–56, 2001.
704. M. G. Reese and F. H. Eeckman. Novel neural network prediction system for human promoters and splice sites. In *Proceedings of Workshop on Gene-Finding and Gene Structure Prediction*, 1995.
705. M. G. Reese and F. H. Eeckman. Time-delay neural networks for eukaryotic promoter prediction, 1999. Unpublished.
706. M. G. Reese and F.H. Eeckman. Novel neural network algorithms for improved eukaryotic promoter site recognition. *Genome Science and Technology*, 1(1):45, 1995.
707. M. G. Reese, N. L. Harris, and F. H. Eeckman. Large scale sequencing specific neural networks for promoter and splice site recognition. In *Proceedings of Pacific Symposium on Biocomputing*, 1996.
708. M. G. Reese, G. Hartzell, N. l. Harris, U. Ohler, J. F. Abril, and S. E. Lewis. Genome annotation assessment in Drosophila melanogaster. *Genome Research*, 10:483–501, 2000.
709. A. Rehm, P. Stern, H. L. Ploegh, and D. Tortorella. Signal peptide cleavage of a type I membrane protein, hcmv us11, is dependent on its membrane anchor. *EMBO Journal*, 20(7):1573–1583, 2001.
710. A. Reinhardt and T. Hubbard. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Research*, 26(9):2230–2236, 1998.
711. J. Rice. *Mathematical Statistics and Data Analysis*. Wadsworth, 1988.
712. M. D. Richard and R. P. Lippmann. Neural networks classifiers estimate Bayesian *a posteriori* probabilities. *Neural Computation*, 3:461–483, 1991.
713. S. K. Riis and A. Krogh. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *Journal of Computational Biology*, 3:163–183, 1996.
714. RIKEN Genome Exploration Research Group Phase II Team and FANTOM Consortium. Functional annotation of a full-length mouse cDNA collection. *Nature*, 409:685–690, 2001.
715. E. Rivas and S. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285:2053–2068, 1999.

716. J. Robie, J. Lapp, and D. Schach. XML Query Language (XQL). In *Position Papers of QL'98—The Query Languages Workshop*, 1998.
717. R. G. Roeder. The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences*, 21:327–335, 1996.
718. I. B. Rogozin, A. V. Kochetov, F. A. Kondrashov, E. V. Koonin, and L. Milanesi. Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. *Bioinformatics*, 17(10):890–900, 2001.
719. D. A. Rosenblueth, D. Thieffry, A. M. Huerta, H. Salgado, and J. Collado-Vides. Syntactic recognition of regulatory regions in *Escherichia coli*. *Computer Applications in the Biosciences*, 12(5):415–422, 1996.
720. C. M. Ross, J. B. Kaplan, M. E. Winkler, and B. P. Nichols. An evolutionary comparison of *Acinetobacter calcoaceticus* trpF with trpF genes of several organisms. *Molecular Biology and Evolution*, 7(1):74–81, 1990.
721. P. Ross-Macdonald, P. S. R. Coelho, T. Roemer, S. Agarwal, A. Kumar, R. Jansen, K.-H. Cheung, A. Sheehan, D. Symoniatis, L. Umansky, M. Heidtman, F. K. Nelson, H. Iwasaki, K. Hager, M. Gerstein, P. Miller, G. S. Roeder, and M. Snyder. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, 402:413–418, 1999.
722. B. Rost. PHD: Predicting one-dimensional protein structure by profile based neural networks. *Methods in Enzymology*, 266:525–539, 1996.
723. B. Rost. Did evolution leap to create the protein universe? *Current Opinion in Structural Biology*, 12:409–416, 2002.
724. B. Rost, P. Fariselli, and R.Casadio. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Science*, 5:1704–1718, 1996.
725. B. Rost and C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19:55–72, 1994.
726. A. Roy, S. Govil, and R. Miranda. Algorithm to generate radial basis function (RBF)-like nets for classification problems. *Neural Networks*, 8(2):179–201, 1995.
727. M. A. Roytberg, T. V. Astahova, and M. S. Gelfand. Combinatorial approaches to gene recognition. *Computer & Chemistry*, 21(4):229–235, 1997.
728. D. E. Rumelhart, G. E. Hinton, and R. J. Wiliams. Learning internal representations by error propagation. In *Parallel Distributed Processing*, volume 1, pages 318–362. M.I.T. Press, 1986.
729. D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.
730. D. E. Rumelhart, J .L. McClelland, and the PDP Research Group, editors. *Parallel Distributed Processing*, volume 1 & 2. MIT Press, 1986.
731. S. Russell, J. Binder, D. Koller, and K. Kanazawa. Local learning in probabilistic networks with hidden variables. In *Proceedings of 14th Joint International Conference on Artificial Intelligence, volume 2*, pages 1146–1152, 1995.
732. R. Rymon. Search through systematic set enumeration. In *Proceedings of 3rd International Conference on Principles of Knowledge Representation and Reasoning*, pages 539–550, 1992.
733. A. Saito, T. Furukawa, S. Fukushige, S. Koyama, M. Hoshi, Y. Hayashi, and A. J. Horii. p24/ING1-ALT1 and p47/ING1-ALT2, distinct alternative transcripts of p33/ING1. *Human Genetics*, 45:177–181, 2000.

734. R. Saito, H. Suzuki, and Y. Hayashiazaki. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Research*, 30:1163–1168, 2002.
735. R. Saito, H. Suzuki, and Y. Hayashizaki. Construction of reliable protein-protein interaction networks with a new interaction generality measure. *Bioinformatics*, 19(6):756–763, 2003.
736. H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26:43–49, 1987.
737. A. A. Salamov, T. Nishikawa, and M. A. Swindells. Assessing protein coding region integrity in cDNA sequencing projects. *Bioinformatics*, 14:384–390, 1998.
738. S. L. Salzberg. Decision trees and Markov chains for gene finding. In *Computational Methods in Molecular Biology*, pages 187–206. Elsevier, 1998.
739. S. L. Salzberg, A. L. Delcher, K. H. Fasman, and J. Henderson. A decision tree system for finding genes in DNA. *Journal of Computational Biology*, 5(4):667–680, 1998.
740. S. L. Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1(3):317–328, 1997.
741. S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2):544–548, 1998.
742. R. Sandy. *Statistics for Business and Economics*. McGrawHill, 1989.
743. D. Sankoff. Simultaneous solution of RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):810–825, 1985.
744. V. M. Sarich and A. C. Wilson. Immunological time scale for hominid evolution. *Science*, 158:1200–1203, 1967.
745. O. Sasson, N. Linial, and M. Linial. The metric space of proteins-comparative study of clustering algorithms. *Bioinformatics*, 18(Supplement):S14–S21, 2002.
746. O. Sasson, A. Vaaknin, H. Fleischer, E. Portugaly, Y. Bilu, N. Linial, and M. Linial. ProtoNet: Hierarchical classification of the protein space. *Nucleic Acids Research*, 31:348–352, 2003.
747. K. Satou, G. Shibayama, T. Ono, Y. Yamamura, E. Furuichi, S. Kuhara, and T. Takagi. Finding association rules on heterogeneous genome data. In *Proceedings of Pacific Symposium on Biocomputing*, pages 397–480, 1997.
748. A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *Proceedings of 21st International Conference on Very Large Data Bases*, pages 432–443, 1995.
749. C. Scharfe, P. Zaccaria, K. Hoertnagel, M. Jaksch, T. Klopstock, R. Lilland H. Prokisch, K.-D. Gerbitz, H. W. Mewes, and T. Meitinger. MITOP: Database for mitochondria-related proteins, genes and diseases. *Nucleic Acids Research*, 27:153–155, 1999.
750. T. E. Scheetz, N. Trivedi, C. A. Roberts, T. Kucaba, B. Berger, N. L. Robinson, C. L. Birkett, A .J. Gavin, B. O'Leary, T. A. Braun, M. F. Bonaldo, J. P. Robinson, V. C. Sheffield, M. B. Soares, and T. L. Casavant. ESTprep: Preprocessing cDNA sequence reads. *Bioinformatics*, 19(11):1318–1324, 2003.
751. T. E Scheetz and J. J. Laffin and B. Berger and S. Mackerly and S. A. Baumes, R. Brown II, S. Chang, J. Coco, J. Conklin, K. Crouch, M. Donohue, G. Doonan,

C. Estes, M. Eyestone, K. Fishler, J. Gardiner, L. Guo, B. Johnson, C. Keppel, R. Kreger, M. Lebeck, R. Marcelino, V. Miljkovich, M. Perdue, L. Qui, J. Rehmann, R.S. Reiter, B. Rhoads, K. Schaefer, C. Smith, I. Sunjevaric, K. Trout, N. Wu, C. L. Birkett, J. Bischof, B. Gackle, A. Gavin, B. Mokrzycki, C. Moressi, B. OLeary, K. Pedretti, C. Roberts, M. Smith, D. Tack, N. Trivedi, T. Kucaba, T. Freeman, J. Lin, M.F. Bonaldo, T. L. Casavant, V. C. Sheffield, M. B. Soares. High-throughput gene discovery in the rat. *Genome Research*, in press.

752. M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.
753. M. Scherf, A. Klingenhoff, and T. Werner. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: A novel context analysis approach. *Journal of Molecular Biology*, 297:599–606, 2000.
754. K. Schittowski. NLQPL: A FORTRAN-subroutine solving constrained nonlinear programming problems. *Annals of Operations Research*, 5:485–500, 1985.
755. G. Schneider, S. Sjöling, E. Wallin, P. Wrede, E. Glaser, and G. von Heijne. Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. *PROTEINS*, 30:49–60, 1998.
756. C. Schönbach, P. Kowalski-Saunders, and V. Brusic. Data warehousing in molecular biology. *Briefings in Bioinformatics*, 1:190–198, 2000.
757. B. Scholkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.
758. G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans. Entrez: Molecular biology database and retrieval system. *Methods in Enzymology*, 266:141–162, 1996.
759. B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18:1257–1261, 2000.
760. B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18:1257–1261, 2000.
761. J. A. Scott, E. L. Palmer, and A. J. Fischman. How well can radiologists using neural network software diagnose pulmonary embolism? *American Journal of Roentgenology*, 175(2):399–405, 2000.
762. B. E. Segee and M. J. Carter. Fault tolerance of pruned multilayer networks. In *Proceedings of International Joint Conference on Neural Networks*, volume II, pages 447–452, 1991.
763. H. P. Selker, J. L. Griffith, S. Patil, W. J. Long, and R. B. D'Agostino. A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. *Journal of Investigative Medicine*, 43(5):468–476, 1995.
764. J. Selletin and B. Mitschang. Data-intensive intra- & internet applications—Experiences using Java and CORBA in the World Wide Web. In *Proceedings of 14th IEEE International Conference on Data Engineering*, pages 302–311, 1998.
765. F. Servant, C. Bru, S. Carrere, et al. ProDom: Automated clustering of homologous domains. *Briefings in Bioinformatics*, 3(3):246–251, 2002.
766. D. Sha and V. B. Bajić. Adaptive on-line ANN learning algorithm and application to identification of non-linear systems. *Informatica: An International Journal of Computing and Informatics*, 23(4):251–259, 1999.
767. D. Sha and V. B. Bajić. On-line adaptive learning rate BP algorithm for MLP and ap-

plication to an identification problem. *Journal of Applied Computer Science*, 7(2):67–82, 1999.

768. D. Sha and V. B. Bajić. On-line hybrid learning algorithm for MLP in identification problems. *Computers & Electrical Engineering, An International Journal*, 28(6):587–598, 2002.
769. J. Shafer, R. Agrawal, and M. Mehta. SPRINT: A scalable parallel classifier for data mining. In *Proceedings of 22nd International Conference on Very Large Data Bases*, pages 544–555, 1996.
770. D. Shalon, S. J. Smith, and P. O. Brown. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, 6(7):639–645, 1996.
771. J. Sietsma and R. J. F. Dow. Neural net pruning—why and how. In *Proceedings of IEEE International Conference on Neural Networks*, volume I, pages 325–333, 1988.
772. J. Sietsma and R. J. F. Dow. Creating artificial neural networks that generalize. *Neural Networks*, 4(1):67–69, 1991.
773. V. L. Singer, C. R. Wobbe, and K. Struhl. A wide variety of DNA sequences can functionally replace yeast TATA element for transcriptional activation. *Genes & Development*, 4:636–645, 1990.
774. S. T. Smale. Generality of a functional initiator consensus sequence. *Gene*, 182:13–22, 1997.
775. S. T. Smale. Transcription initiation from TATA-less promoters within eukaryotic protein coding genes. *Biochimica et Biophysica Acta*, 1351:73–88, 1997.
776. S. T. Smale and D. Baltimore. The initiator as a transcriptional control element. *Cell*, 57:103–111, 1989.
777. I. Small, H. Wintz, K. Akashi, and H. Mireau. Two birds with one stone: Genes that encode products targeted to two or more compartments. *Plant Molecular Biology*, 38:265–277, 1998.
778. G. P. Smith. Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–1317, 1985.
779. H. O. Smith and K. W. Wilcox. A restriction enzyme from *Hemophilus influenzae* I: Purification and general properties. *Journal of Molecular Biology*, 51(2):379–391, 1970.
780. T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.
781. E. E. Snyder and G. D. Stormo. Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks. *Nucleic Acids Research*, 21:607–613, 1993.
782. E. E. Snyder and G. D. Stormo. Identification of protein coding regions in genomic DNA. *Journal of Molecular Biology*, 248:1–18, 1995.
783. L. A. Soinov, M. A. Krestyaninova, and A. Brazma. Towards reconstruction of gene networks from expression data by supervised learning. *Genome Biology*, 4(1):R6.1–9, 2003.
784. V. Solovyev and A. Salamov. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Intelligent Systems for Molecular Biology*, 5:294–302, 1997.
785. V. V. Solovyev, A. A. Salamov, and C. B. Lawrence. Predicting internal exons by

oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Research*, 22:5156–5163, 1994.

786. D. E. Soltis, P. S. Soltis, M. W. Chase, M. E. Mort, D. C. Albach, M. Zanis, V. Savolainen, W. H. Hahn, S. B. Hoot, M. F. Fay, M. Axtell, S. M. Swensen, L. M. Prince, W. J. Kress, K. C. Nixon, and J. S. Farris. Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. *Botanical Journal of the Linnean Society*, 133(4):381–461, 2000.
787. D. E. Soltis, P. S. Soltis, D. L. Nickrent, L. A. Johnson, W. J. Hahn, S. B. Hoot, J. A. Sweere, R. K. Kuzoff, K. A. Kron, M. W. Chase, S. M. Swensen, E. A. Zimmer, S. M. Chaw, L. J. Gillespie, W. J. Kress, and K. J. Sytsma. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. *Annals of the Missouri Botanical Garden*, 84:1–49, 1997.
788. D. F. Specht. Probabilistic neural networks for classification, mapping or associative memory. In *Proceedings of IEEE International Conference on Neural Networks*, volume 1, pages 525–532, 1988.
789. D. F. Specht. Probabilistic neural networks. *Neural Networks*, 3:109–118, 1990.
790. D. F. Specht. A general regression neural network. *IEEE Transactions on Neural Networks*, 2:568–576, 1991.
791. D. F. Specht. Probabilistic neural networks and general regression neural networks. In *Fuzzy Logic and Neural Network Handbook*, chapter 3, McGraw-Hill, 1996.
792. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.
793. E. Sprinzak, S. Sattath, and H. Margalit. How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, 327(5):919–923, 2003.
794. R. Srikant and R. Agrawal. Mining generalized association rules. In *Proceedings of 21st International Conference on Very Large Data Bases*, pages 407–419, 1995.
795. R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Research*, 12:505–519, 1984.
796. R. Staden. Methods to define and locate patterns of motifs in sequences. *Computer Applications in the Biosciences*, 4:53–60, 1988.
797. J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611–1618, 2002.
798. B. J. Stapley, L. A. Kelley, and M. J. E. Sternberg. Predicting the subcellular location of proteins from text using support vector machines. In *Proceedings of Pacific Symposium on Biocomputing*, pages 374–385, 2002.
799. L. A. Stargell and K. Struhl. Mechanisms of transcriptional activation in vivo: Two steps forward. *Trends in Genetics*, 8:311–315, 1996.
800. E. W. Steeg. Neural networks, adaptive optimization, and RNA secondary structure prediction. In *Artificial Intelligence and Molecular Biology*, pages 121–160, 1993.
801. G. D. Stormo. DNA binding sites: Representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

802. G. D. Stormo. Gene finding approaches for eukaryotes. *Genome Research*, 10:394–397, 2000.
803. G. D. Stormo. Computer methods for analyzing sequence recognition of nucleic acids. *Annual Review of Biophysics and Biophysical Chemstry*, 17:241–63, 1988.
804. G. D. Stormo, T. D. Schneider, and L. M. Gold. Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Research*, 10(9):2971–2996, 1982.
805. G. D. Stormo, T. D. Schneider, L. M. Gold, and A. Ehrenfeucht. Use of the 'perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*, 10:2997–3010, 1982.
806. M. Streuli, N. X. Krueger, T. Thai, M. Tang, and H. Saito. Distinct functional roles of the two intracellular phosphatase-like domains of the receptor-linked protein tyrosine phosphatases LCA and LAR. *EMBO Journal*, 9:2399–2407, 1990.
807. M. Struyvé, M. Moons, and J. Tommassen. Carboxyl-terminal phenylalanine is essential for the correct assembly of a bacterial outer membrane protein. *Journal of Molecular Biology*, 218:141–148, 1991.
808. C. M. Stultz, R. Nambudripad, R. H. Lathrop, and J. V. White. Predicting protein structure with probabilistic models. In *Protein Structural Biology in Biomedical Research*, pages 447–506, 1997.
809. M. Suwa. Fraction of proteins annotatable by sequence similarity, November 2003. Private Communication.
810. D. L. Swofford. *PAUP: Phylogenetic Analysis Using Parsimony, Ver. 4.OB8*. Sinauer Associates, 2001.
811. S.-H. Sze and P. A. Pevzner. Las Vegas algorithms for gene recognition: Suboptimal and error-tolerant spliced alignment. *Journal of Computational Biology*, 4(3):297–310, 1997.
812. S.-H. Sze, M. Roytberg, M. Gelfand, A. Mironov, T. Astakhova, and P. Pevzner. Algorithms and software for support of gene identification experiments. *Bioinformatics*, 14(1):14–19, 1998.
813. J. E. Tabaska, R. B. Cary, H. N. Gabow, and G. D. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–699, 1998.
814. T. Takagi and M. Sugeno. Derivation of fuzzy control rules from human operator's control actions. In *Proceedings of IFAC Symposium on Fuzzy Information, Knowledge Representation, and Decision Analysis*, pages 55–60, 1983.
815. A. Takhtajan. *Diversity and Classification of Flowering Plants*. Columbia University Press, 1997.
816. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96:2907–2912, 1999.
817. C. K. Tang and D. E. Draper. Unusual mRNA pseudoknot structure is recognized by a protein translational repressor. *Cell*, 57:531–536, 1989.
818. R. L. Tatusov, A. R. Mushegian, P. Bork, N. R. Brown, W. S. Hayes, M. Borodovski, K. E. Rudd, and E. V. Koonin. Metabolism and evolution of *haemophilus influenzae* deduced from a whole genome comparison with *escherichia coli*. *Current Biology*, 6:279–291, 1996.

819. R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, 29:22–28, 2001.
820. T. A. Tatusova and T. L. Madden. BLAST 2 Sequences—a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters*, 174:247–250, 1999.
821. P. Taubert. Leguminosae. In *Die Naturlichen Pflanzenfamilien*. W. Engelmann, 1894.
822. N. R. Temkin, R. Holubkov, J. E. Machamer, H. R. Winn, and S. S. Dikmen. Classification and regression trees (CART) for prediction of function at 1 year following head trauma. *Journal of Neurosurgery*, 82(5):764–771, 1995.
823. M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. USA*, 98:8614–8619, 2001.
824. G. Thimm and E. Fiesler. Evaluating pruning methods. In *Proceedings of International Symposium on Artificial Neural Networks*, volume A2, pages 20–25, 1995.
825. J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. The CLUSTAL-X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, 25(24):4876–4882, 1997.
826. J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.
827. R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA*, 99(10):6567–6572, 2002.
828. B. Tinland, Z. Koukolikova-Nicola, M. N. Hall, and B. Hohn. The t-DNA-linked vird2 protein contains two distinct functional nuclear localization signals. *Proc. Natl. Acad. Sci. USA*, 89:7442–7446, 1992.
829. I. Tinoco, P. N. Borer, B. Dengler, M. D. Levine, O. C. Uhlenbeck, D. M. Crothers, and J. Gralla. Improved estimation of secondary structure in ribonucleic acids. *Nature New Biology*, 246:40–41, 1973.
830. I. Tinoco, O. C. Uhlenbeck, and M. D. Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230:362–367, 1971.
831. S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy. Prediction of probable genes by Fourier analysis of genomic sequences. *Computer Applications in the Biosciences*, 13(3):263–270, 1997.
832. H. Toivonen. Sampling large databases for association rules. In *Proceedings of 22rd International Conference on Very Large Data Bases*, pages 134–145, 1996.
833. A. H. Tong, B. Drees, G. Nardelli, et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–324, 2002.
834. N. K. Tonks and B. G. Neel. From form to function: Signaling by protein tyrosine phosphatases. *Cell*, 87(3):361–364, 1996.
835. E. N. Trifonov. Inferring context of regulatory sequence elements. *Computer Applications in the Biosciences*, 12:423–429, 1996.
836. N. Trivedi, J. Bischof, S. Davis, K. Pedretti, T. E. Scheetz, T. A. Braun, C. A. Roberts, N. L. Robinson, V. C. Sheffield, M. B. Soares, and T. L. Casavant. Parallel creation

of non-redundant gene indices from partial mRNA transcripts. *Future Generation Computer Systems*, 18(6):863–870, 2002.

837. O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525, 2001.
838. A. Tsuji and I. Tamai. Organic anion transporters. *Pharm. Biotechnol.*, 12:471–491, 1999.
839. S. C. Tucker. Floral development in *Saururus cernuus*: 1. Floral initiation and stamen development. *American Journal of Botany*, 62(3):289–301, 1975.
840. S. C. Tucker. Inflorescence and flower development in the Piperaceae: I. *Peperomia*. *American Journal of Botany*, 67(5):686–702, 1980.
841. S. C. Tucker. Inflorescence and floral development in *Houttuynia cordata* (Saururaceae). *American Journal of Botany*, 68(8):1017–1032, 1981.
842. S. C. Tucker. Inflorescence and flower development in the Piperaceae: II. Inflorescence development of *Piper*. *American Journal of Botany*, 69(5):743–752, 1982.
843. S. C. Tucker. Inflorescence and flower development in the Piperaceae: III. Inflorescence development of *Piper*. *American Journal of Botany*, 69(9):1389–1401, 1982.
844. S. C. Tucker. Initiation and development of inflorescence and flower in *Anemopsis californica* (Saururaceae). *American Journal of Botany*, 72(1):20–31, 1985.
845. S. C. Tucker and A. W. Douglas. Floral structure, development, and relationships of paleoherbs: *Saruma*, *Cabomba*, *Lactoris*, and selected Piperales. In D. W. Taylor and L. J. Hickey, editors, *Flowering Plants: Origin, Evolution, and Phylogeny*. Chapman & Hall Press, 1996.
846. S. C. Tucker, A. W. Douglas, and H.-X. Liang. Utility of ontogenetic and conventional characters in determining phylogenetic relationships of Saururaceae and Piperaceae (Piperales). *Systematic Botany*, 18(4):414–441, 1993.
847. E. P. Turton, D. J. Scott, M. Delbridge, S. Snowden, and R. C. Kester. Ruptured abdominal aortic aneurysm: A novel method of outcome prediction using neural network technology. *European Journal of Vascular and Endovascular Surgery*, 19(2):184–189, 2000.
848. E. C. Uberbacher and R. J. Mural. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA*, 88:11261–11265, 1991.
849. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.
850. J. D. Ullman. *Principles of Database and Knowledgebase Systems I*. Computer Science Press, 1989.
851. A. Ureta-Vidal, L. Ettwiller, and E. Birney. Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nature Reviews Genetics*, 4(4):251–262, 2003.
852. P. E. Utgoff. An incremental ID3. In *Proceedings of 5th International Conference on Machine Learning*, pages 107–120, 1988.
853. A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, 12(3):368–373, 2002.
854. K. H. van Wely, J. Swaving, R. Freudl, and A. J. Driessen. Translocation of proteins across the cell envelope of gram-positive bacteria. *FEMS Microbiology Reviews*,

25(4):437–454, 2001.
855. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
856. V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
857. V. Veljković, I. Ćosić, B. Dimitrijević, and Lalović. Is it possible to analyse DNA and protein sequences by the methodsof digital signal processing? *IEEE Transactions on Biomedical Engineering*, 32(5):337–341, 1985.
858. V. Veljković and I. Slavić. Simple general-model pseudopotential. *Physical Review Letters*, 29(5):105–107, 1972.
859. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.
860. C. P. Verrijzer and R. Tjian. TAFs mediate transcriptional activation and promoter selectivity. *Trends in Biochemical Sciences*, 21:338–342, 1996.
861. G. von Heijne. Patterns of amino acids near signal-sequence cleavage sites. *European Journal of Biochemistry*, 133:17–21, 1983.
862. G. von Heijne. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research*, 14:4683–4690, 1986.
863. G. von Heijne. *Sequence Analysis in Molecular Biology: Treasure Trove or Trivial Pursuit*. Academic Press, 1987.
864. C. von Mering, R. Krause, B. Snel, et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.
865. J. L. Vriesema, H. G. van der Poel, F. M. Debruyne, J. A. Schalken, L. P. Kok, and M. E. Boon. Neural network-based digitized cell image diagnosis of bladder wash cytology. *Diagnostic Cytopathology*, 23(3):171–179, 2000.
866. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339, 1989.
867. A. Waibel, H. Sawai, and K. Shikano. Modularity and scaling in large phonemic neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12):1888–1898, 1989.
868. A. J. Walhout, R. Sordella, X. Lu, et al. Protein interaction mapping in *c. elegans* using proteins involved in vulval development. *Science*, 287(5450):116–122, 2000.
869. M. Walker, V. Pavlovic, and S. Kasif. A comparative genomic method for computational identification of prokaryotic translation initiation sites. *Nucleic Acids Research*, 30(14):3181–3191, 2002.
870. L. Wang. Multi-associative neural networks and their applications to learning and retrieving complex spatio-temporal sequences. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, 29(1):73–82, 1999.
871. W. Q. Wang, J. P. Sun, and Z. Y. Zhang. An overview of the protein tyrosine phosphatase superfamily. *Current Topics in Medicinal Chemistry*, 3(7):739–748, 2003.
872. B. Wasylyk. Transcription elements and factors of RNA polymerase B promoters of higher eukaryotes. *Critical Reviews Biochemistry*, 23:77–120, 1988.
873. J. D. Watson. The human genome project: Past, present, and future. *Science*, 248:44–49, 1990.
874. J. D. Watson and R. M. Cook-Deegan. Origins of the human genome project. *FASEB Journal*, 5:8–11, 1991.

875. J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.
876. A. S. Weigend, D. E. Rumelhart, and B. A. Huberman. Back-propagation, weight-elimination, and time series prediction. In D. Touretzky, J. Elman, T. Sejnowski, and G. Hinton, editors, *Proceedings of Connectionist Models Summer School*, pages 105–116, 1990.
877. A. S. Weigend, D. E. Rumelhart, and B. A. Hubermann. Generalization by weight-elimination with application to forecasting. *Advances in Neural Information Processing*, 3:875–882, 1991.
878. R. O. J. Weinzierl. *Mechanism of Gene Expression*. Imperial College Press, 1999.
879. T. Werner. Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome*, 10:168–175, 1999.
880. J. Westbrook, Z. Feng, S. Jain, T. N. Bhat, N. Thanki, V. Ravichandran, G. L. Gilliland, W. Bluhm, H. Weissig, D. S. Greer, P. E. Bourne, and H. E. Berman. The Protein Data Bank: Unifying the archive. *Nucleic Acids Research*, 30(1):245–248, 2002.
881. J. Weston, F. Perez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff, and B. Scholkopf. Feature selection and tranduction for prediction of molecular bioactivity for drug design. *Bioinformatics*, 19(6):764–771, 2003.
882. D. L. Wheeler, D. M. Church, A. E. Lash, D. D. Leipe, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, T. A. Tatusova, L. Wagner, and B. A. Rapp. Database resources of the national center for biotechnology information: 2002 update. *Nucleic Acids Research*, 30:13–16, 2002.
883. D. Whitley and C. Bogart. The evolution of connectivity: Pruning neural networks using genetic algorithms. In *Proceedings of International Joint Conference on Neural Networks*, volume I, pages 134–137, 1990.
884. B. Widrow and M. A. Lehr. 30 years of adaptive neural networks: Perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78(9):1415–1442, 1990.
885. B. Widrow, R. G. Winter, and R. A. Baxter. Layered neural nets for pattern recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36:1109–1118, 1988.
886. S. R. Wiley, R. J. Kraus, and J. E. Mertz. Functional binding of TATA box binding component of transcription factor TFIID to the –30 region of TATA-less promoters. *Proc. Natl. Acad. Sci. USA*, 89:5814–5818, 1992.
887. E. Wingender. Transcription regulating proteins and their recognition sequences. *Critical Reviews in Eukaryotic Gene Expression*, 1:11–48, 1990.
888. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann, 2000.
889. C. R. Wobbe and K. Struhl. Yeast and human TATA-binding proteins have nearly identical DNA sequence requirement for transcription in vitro. *Molecular and Cellular Biology*, 10:3859–3867, 1990.
890. L. Wodicka, H. Dong, M. Mittmann, M.-H. Ho, and D. J. Lockhart. Genome-wide expression monitoring in *saccharomyces cerevisiae*. *Nature Biotechnology*, 15:1359–1367, 1997.
891. C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA*,

87:4576–4579, 1990.

892. J. Wojcik and V. Schächter. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17(Supplement 1):S296–S305, 2001.
893. S. L. Wolfe. *Introduction to Cell and Molecular Biology*. Wadsworth, 1995.
894. L. Wong. Kleisli, a functional query system. *Journal of Functional Programming*, 10(1):19–56, 2000.
895. L. Wong. Kleisli, its exchange format, supporting tools, and an application in protein interaction extraction. In *Proceedings of IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pages 21–28, 2000.
896. L. Wong. PIES, a protein interaction extraction system. In *Proceedings of Pacific Symposium on Biocomputing*, pages 520–531, 2001.
897. L. Wong. Technologies for integrating biological data. *Briefings in Bioinformatics*, 3(4):389–404, 2002.
898. V. Wood, R. Gwilliam, M.A. Rajandream, M. Lyne, R. Lyne, A. Stewart, J. Sgouros, N. Peat, et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature*, 415(6874):871–880, 2002.
899. R. Wooster, S. L. Neuhausen, J. Mangion, et al. Localization of a breast cancer susceptibility gene, brca2, to chromosome 13q12-13. *Science*, 265(5181):2088–2090, 1994.
900. J. Wootton and S. Federhen. Statistics of local complexity in amino acid sequences and sequence databases. *Computers and Chemistry*, 17:149–163, 1993.
901. C. T. Workman and G. D. Stormo. ANN-Spec: A method for discovering transcription factor binding sites with improved specificity. In *Proceedings of Pacific Symposium on Biocomputing*, pages 112–123, 2000.
902. C. H. Wu, S. Zhao, and H. L. Chen. A protein class database organized with PROSITE protein groups and PIR superfamilies. *Journal of Computational Biology*, 3:547–562, 1996.
903. C. H. Wu, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, R. S. Ledley, K. C. Lewis, H.-W. Mewes, B. C. Orcutt, B. E. Suzek, A. Tsugita, C. R. Vinayaka, L.-S. Yeh, J. Zhang, and W. C. Barker. The Protein Information Resource: An integrated public resource of functional annotation of proteins. *Nucleic Acids Research*, 30(1):35–37, 2002.
904. C. H. Wu, C. Xiao, Z. Hou, H. Huang, and W. C. Barker. iProClass: An integrated, comprehensive, and annotated protein classification database. *Nucleic Acids Research*, 29:52–54, 2001.
905. K. C. Wu, J. T. Bryan, M. I. Morasso, S. I. Jang, J. H. Lee, J. M. Yang, L. N. Marekov, D. A. Parry, and P. M. Steinert. Coiled-coil trigger motifs in the 1B and 2B rod domain segments are required for the stability of keratin intermediate filaments. *Molecular Biology of the Cell*, 11:3539–3558, 2000.
906. Z.-Y. Wu. An outline of phytogeography. *The Society of Botanists in Yunnan Province*, 1:44–45, 1984.
907. Z.-Y. Wu, Y.-C. Tang, A.-M. Lu, and Z.-D. Chen. On primary subdivisions of the magnoliophyta—towards a new scheme for an eight-class system of classification. *Acta Phytotaxonomica Sinica*, 36:385–402, 1998.
908. Z.-Y. Wu and W.-C. Wang. A preliminary study on tropical and subtropical flora in Yunnan I. *Acta Phytotaxonomica Sinica*, 6(2):183–254, 1957.

909. Z.-Y. Wu and W.-C. Wang. Some corrections on the paper "A preliminary study on tropical and subtropical flora in Yunnan I". *Acta Phytotaxonomica Sinica*, 7(2):193–196, 1958.
910. I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303–305, 2002.
911. D. Xu, G. Li, L. Wu, J. Zhou, and Y. Xu. PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics*, 18(11):1432–1437, 2002.
912. Y. Xu, R. J. Mural, and E. C. Uberbacher. Inferring gene structures in genomic sequences using pattern recognition and expressed sequence tags. *Intelligent Systems for Molecular Biology*, 5:344–353, 1997.
913. Y. Xu and E. C. Uberbacher. Automated gene identification in large-scale genomic sequences. *Journal of Computational Biology*, 4(3):325–338, 1997.
914. Y. Xu and E. C. Uberbacher. Computational gene prediction using neural networks and similarity search. In *Computational Methods in Molecular Biology*, chapter 7, pages 109–128. Elsevier, 1998.
915. T. Yada, M. Ishikawa, H. Tanaka, and K. Asai. Extraction of hidden Markov model representations of signal patterns in DNA sequences. In *Proceedings of Pacific Symposium on Biocomputing*, pages 686–696, 1996.
916. F. Yang, L. G. Moss, and G. N. Phillips. The molecular structure of green fluorescent protein. *Nature Biotechnology*, 14(10):1246, 1996.
917. R. J. Yarger, G. Reese, and T. King. *MySQL & mSQL*. O'Reilly, 1999.
918. E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. William, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Reilling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.
919. G. Yona, N. Linial, and M. Linial. ProtoMap: Automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Research*, 28:49–55, 2000.
920. M. Yoshida, K. Fukuda, and T. Takagi. PNAD-CSS: A workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, 16:169–175, 2000.
921. D. Yu, S. Chatterjee, G. Sheikholeslami, and A. Zhang. Efficiently detecting arbitrary shaped clusters in very large datasets with high dimensions. Technical Report 98-08, Department of Computer Science and Engineering, State University of New York, Buffalo, New York, November 1998.
922. Z. Yuan. Prediction of protein subcellular locations using Markov chain models. *FEBS Letters*, 451:23–26, 1999.
923. Z. Yuan and R. D. Teasdale. Prediction of golgi type II membrane proteins based on their transmembrane domains. *Bioinformatics*, 18:1109–1115, 2002.
924. B. P. Yuhas, Jr. M. H. Goldstein, T. J. Sejnowski, and R. E. Jenkins. Neural network model of sensory integration for improved vowel recognition. *Proceedings of the IEEE*, 78(10):1658–1668, 1990.
925. M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *Proceedings of 3rd International Conference on Knowledge Discovery and Data Mining*, pages 283–286, 1997.

926. L. Zawel and D. Reinberg. Initiation of transcription by RNA polymerase II: A multi-step process. *Progress in Nucleic Acid Research and Molecular Biology*, 44:67–108, 1993.
927. E. M. Zdobnov and R. Apweiler. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17:847–848, 2001.
928. F. Zeng, R. Yap, and L. Wong. Using feature generation and feature selection for accurate prediction of translation initiation sites. In *Proceedings of 13th International Conference on Genome Informatics*, pages 192–200, 2002.
929. M. Zeremski, J. E. Hill, S. S. Kwek, I. A. Grigorian, K. V. Gurova, I. V. Garkevtsev, L. Diatchenko, E. V. koonin, and A. V. Gudkov. Structure and regulation of the mouse ING1 gene. Three alternative transcripts encode two PHD finger proteins that have opposite effects on p53 function. *Journal of Biological Chemistry*, 274:32172–32181, 1999.
930. M. Q. Zhang. Computational prediction of eukaryotic protein-coding genes. *Nature Reviews Genetics*, 3(9):698–709, 2002.
931. M. Q. Zhang. Identification of human gene core promoter in silico. *Genome Research*, 8:319–326, 1998.
932. R. Zhang, G. Evans, F. J. Rotella, E. M. Westbrook, D. Beno, E. Huberman, A. Joachimiak, and F. R. Collart. Characteristics and crystal structure of bacterial inosine-5'-monophosphate dehydrogenase. *Biochemistry*, 38:4691–4700, 1999.
933. T. Zhang. Association rules. In *Proceedings of 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 245–256, 2000.
934. T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of ACM-SIGMOD International Conference on Management of Data*, pages 103–114, Montreal, Canada, June 1996.
935. X. Zhang, G. Dong, and K. Ramamohanarao. Exploring constraints to efficiently mine emerging patterns from large high-dimensional datasets. In *Proceedings of 6th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 310–314, 2000.
936. Z. Y. Zhang. Protein-tyrosine phosphatases: biological function, structural characteristics, and mechanism of catalysis. *Critical Reviews in Biochemistry and Molecular Biology*, 33(1):1–52, 1998.
937. G. Zhou and J. Si. Subset-based training and pruning of sigmoid neural networks. *Neural Networks*, 12(1):80–89, 1999.
938. X. Zhou, M. Cahoon, P. Rosa, and L. Hedstrom. Expression, purification, and characterization of inosine 5'-monophosphate dehydrogenase from *Borrelia burgdorferi*. *Journal of Biological Chemistry*, 272:21977–21981, 1997.
939. H. Zhu, M. Bilgin, R. Bangham, et al. Global analysis of protein activities using proteome chips. *Science*, 293(5537):2101–2105, 2001.
940. A. Zien, G. Raatsch, S. Mika, B. Schoelkopf, C. Lemmem, A. Smola, T. Lengauer, and K.R. Mueller. Engineering support vector machine kernels that recognize translation initiation sites. In *Proceedings of German Conference on Bioinformatics*, pages 37–43, 1999.
941. A. Zien, G. Raatsch, S. Mika, B. Schoelkopf, T. Lengauer, and K.R. Mueller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, 2000.

942. M. Zilversmit, P. O'Grady, and R. Desalle. Shallow gemomics, phylogenetics, and evolution in the family Drosophilidae. In *Proceedings of Pacific Symposium on Biocomputing*, pages 512–523, 2002.
943. M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.
944. M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.