Bibliography

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. *The Molecular Biology of the Cell*. Garland Publishing, New York, NY, 1989.

[2] 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 Res.*, 25:3389–3402, 1997.

[3] V. Bafna and D. Huson. The conserved exon method for gene finding. In Brunak et al. [18], pages 3–12.

[4] S. Batzoglou, L. Pachter, J. P. Mesirov, B. Berger, and E. S. Lander. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, 10:950–958, 2000.

[5] L. E. Baum. An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[6] R. Bellman. *Dynnamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.

[7] G. Bernardi. The isochore structure of the human genome. *Annual Review of Genetics*, 23:637–661, 1989.

[8] A. Bird. CpG islands as gene markers in the vertebrate nucleus. *Trends in Genetics*, 3:342–347, 1987.

[9] E. Birney and R. Durbin. Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. In Gaasterland et al. [45], pages 56–64.

[10] E. Birney and R. Durbin. Using Genewise in the Drosophila annotation experiment. *Genome Res.*, 10:547–548, 2000.

[11] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, UK, 1995.

[12] T. Blumenthal, D. Evans, C. D. Link, A. Guffanti, D. Lawson, J. Thierry-Mieg, D. Thierry-Mieg, W. L. Chui, K. Duke, M. Kirali, and S. K. Kim. A global analysis of Caenohabditis elegans operons. *Nature*, 417:851–853, 2002.

[13] T. Blumenthal and K. Steward. RNA processing and gene structure. In D. L. Riddle, T. Blumenthal, B. J. Meyer, and J. R. Priess, editors, *C.ELEGANS II*, chapter 6. Cold Spring Harbor Laboratory Press, New York, NY, 1997.

[14] T. Blumenthal, O. White, and C. Fields. The C. elegans cleavage and polyadenylation signal. *The Worm Breeder's Gazette*, 13:62–63, 1993.

[15] M. Borodovsky and J. McIninch. GENMARK: parallel gene recognition for both DNA strands. *Computers and Chemistry*, 17(2):123–133, 1993.

[16] R. P. Brent. An algorithm with guranteed convergence for finding the minimum of a function of one variable. In *Algorithms for function minimization without derivatives*, chapter 5, pages 61–80. Prentice Hall, Englewood Cliffs, NJ, 1973.

[17] S. Brunak, J. Engelbrecht, and S. Knudsen. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, 220(1):49–65, 1991.

[18] S. Brunak, F. Galison, M. Gribskov, A. Krogh, A. G. Pederson, P. Rouze, G. Stormo, and A. Tramontano, editors. *Proceedings of the Ninth International Conference on Intelligent Systems for Molecular Biology*, Oxford, UK, 2001. Oxford University Press.

[19] P. Bucher. Weight matrix descriptions of four eukaryotic RNA polymerase elements derived from 502 unrelated promoter seqeunces. *J. Mol. Biol.*, 212:563–578, 1990.

[20] C. Burge. *The identification of genes in human genomic DNA*. PhD thesis, Stanford University, 1997.

[21] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.

[22] C. Burge and S. Karlin. Finding the genes in genomic DNA. *Current Opinion in Structural Biology*, 8:346–354, 1998.

[23] M. Burset and R. Guigo. Evaluation of gene structure prediction programs. *Genomics*, 34:353–367, 1996.

[24] The chromosome 21 mapping and sequencing consortium. The DNA sequence of human chromosome 21. *Nature*, 405:311–319, 2000.

[25] J. S. Chuang and D. Roth. Gene recognition based on DAG shortest paths. In Brunak et al. [18], pages 56–64.

[26] J.-M. Claverie. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Gen.*, 6:1735–1744, 1997.

[27] R. Conrad, J. Thomas, J. Speith, and T. Blumenthal. Insertion of part of an intron into the 5' untranslated region of a Caenorhabditis elegans gene converts it into a trans-spliced gene. *Molecular Cell Biology*, 11:1921–1926, 1991.

[28] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[29] R. Davuluri, I. Grosse, and M. Zhang. Computational identification of promoters and first exons in the human genome. *Nature Genet.*, 29:412–417, 2001.

[30] P. Deloukas, L.H. Matthews, J. Ashurst, J. Burton, J.G. Gilbert, M. Jones, G. Stavrides, J.P. Almeida, A.K. Babbage, C.L. Bagguley, J. Bailey, K.F. Barlow, K.N. Bates, L.M. Beard, D.M. Beare, O.P. Beasley, C.P. Bird, S.E. Blakey, A.M. Bridgeman, A.J. Brown, et al. The DNA sequence and comparative analysis of human chromosome 20. *Nature*, 414:865–871, 2001.

[31] 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*, B 39:1–38, 1977.

[32] E. W. Dijkstra. A note on two problems in connection with graphs. *Numerische Mathematics*, 1:269–271, 1959.

[33] S. Dong and D. B. Searls. Gene structure prediction by linguistic methods. *Genomics*, 23:540–551, 1994.

[34] T. Down and T. Hubbard. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, 12:458–461, 2002.

[35] I. Dunham, N. Shimizu, B.A. Roe, S. Chissoe, A.R. Hunt, J.E. Collins, R. Bruskiewich, D.M. Beare, M. Clamp, L.J. Smink, R. Ainscough, J.P. Almeida, A. Babbage, C. Bagguley, J. Bailey, K. Barlow, K.N. Bates, O. Beasley, C.P. Bird, S. Blakey, et al. The DNA sequence of human chromosome 22. *Nature*, 402:489–495, 1999.

[36] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1998.

[37] L. Duret, D. Mourchiroud, and C. Gautier. Statistical analysis of vertebrate sequences reveals that long genes are scare in GC-rich isochores. *Journal of Molecular Evolution*, 40:308–317, 1995.

[38] S. R. Eddy. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, 2:919–929, 2001.

[39] J. W. Fickett. The gene identification problem - an overview for developers. *Computers and Chemistry*, 20:103–118, 1996.

[40] J. W. Fickett and A. G. Hatzigeorgiou. Eukaryotic promoter recognition. *Genome Res.*, 7:861–878, 1997.

[41] J. W. Fickett and C. S. Tung. Assessment of protein coding measures. *Nucleic Acids Res.*, 20:6441–6450, 1992.

[42] C. A. Fields and C. A. Soderlund. gm: a practical tool for automating DNA sequence analysis. *Comput. Applic. Biosci.*, 6:263–270, 1990.

[43] R. Fletcher and C. Reeves. Function minimisation by conjugate gradients. *Computing Journal*, pages 149–154, 1964.

[44] L. Florea, G. Hartzell, Z Zhang, G. M. Rubin, and W. Miller. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, 8:967–974, 1998.

[45] T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, editors. *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, CA, 1997. AAAI Press.

[46] M. Gelfand, L. Podolski, T. Astakhova, and A. Roytberg. Recognition of genes in human DNA sequences. *Journal of Computational Biology*, 3:223–234, 1996.

[47] M. S. Gelfand. Computer prediciton of the exon-intron structure of mammalian pre-mRNAs. *Nucleic Acids Res.*, 18:5865–5869, 1990.

[48] M. S. Gelfand, A. A. Mironov, and P. A. Pevzner. Gene recognition via spliced sequence alignment. *Proc. Natl. Acad. Sci. USA*, 93:9061–9066, 1996.

[49] M. S. Gelfand and M. A. Roytberg. Prediction of the exon-intron structure by a dynamic programming approach. *Biosystems*, 3:173–182, 1993.

[50] D. E. Goldberg. *Genetic algorithms in search, optimisation and machine learning*. Addison Wesley, Boston, MA, 1989.

[51] R. Guigo. Computational gene prediction - an open problem. *Computers and Chemistry*, 21:215–222, 1997.

[52] R. Guigo. Assembling genes from predicted exons in linear time with dynamic programming. *Journal of Computational Biology*, 5:681–702, 1998.

[53] R. Guigo, P. Agarwal, J. F. Abril, M. Burset, and J. W. Fickett. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.*, 10:1631–1642, 2000.

[54] R. Guigo and J.W. Fickett. Distinctive sequence features in protein coding, genic non-coding and intergenic human DNA. *J. Mol. Biol.*, 253:51–60, 1995.

[55] A. Hatzigeorgiou. Translation initiation start prediction in human cDNAs with high accuracy. *Bioinformatics*, 18:343–50, 2002.

[56] J. Henderson, S. Salzberg, and K. H. Fasman. Finding genes in DNA with a hidden Markov model. *Journal of Computational Biology*, 4(2):127–141, 1997.

[57] S. Henikoff, M. A. Keene, K. Fechtel, and J. W. Fristrom. Gene within a gene: nested Drosophila genes encode unrelated proteins on oppostie DNA strands. *Cell*, 44:33–42, 1986.

[58] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 1979.

[59] K. L. Howe, T. Chothia, and R. Durbin. GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. *Genome Res.*, 12:1418–1427, 2002.

[60] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and Clamp M. The Ensembl genome database project. *Nucleic Acids Res.*, 30:38–41, 2002.

[61] W. J. Kent. BLAT - the BLAST-like alignment tool. *Genome Res.*, 12:656–664, 2002.

[62] W. J. Kent and A. M. Zahler. Conservation, regulation, synteny, and introns in a large-scale C.briggsae-C.elegans genomic alignment. *Genome Res.*, 10:1115–1125, 2000.

[63] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[64] I. Korf, P. Flicek, D. Duan, and M. R. Brent. Integrating genomic homology into gene structure prediction. In T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, editors, *Proceedings of the Eigth International Conference on Intelligent Systems for Molecular Biology*, pages 140–148, Menlo Park, CA, 2002. AAAI Press.

[65] M. Kozak. Initiation of translation in prokaryotes and eukaryotes. *Gene*, 234:187–208, 1999.

[66] M. Krause and D. Hirsh. A trans-spliced leader sequence on actin mRNA in C.elegans. *Cell*, 49:753–761, 1987.

[67] A. Krogh. Hidden Markov models for labeled sequences. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, pages 140–144, Los Alamitos, CA, 1994. IEEE Computer Society Press.

[68] A. Krogh. Two methods for improving performance of a HMM and their application for gene finding. In Gaasterland et al. [45], pages 179–186.

[69] A. Krogh. Using database matches with HMMGene for automated gene detection on Drosophila. *Genome Res.*, 10:523–528, 2000.

[70] A. Krogh, I. S. Mian, and D. Haussler. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res.*, 22:4768–4778, 1994.

[71] D. Kulp, D. Haussler, M. Reese, and F. H. Eeckman. Integrating database homology in a probabilistic gene structure model. In R. Altman, A. Dunker, and T. Klein L. Hunter, editors, *Proceedings of Pacific Symposium on Biocomputing*, pages 232–244, 1997.

[72] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman. A generalized hidden Markov model for the recognition of human genes in DNA. In States et al. [106], pages 134–142.

[73] A. N. Ladd and T. A. Cooper. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.*, 3:reviews0008, 2002.

[74] A. Levine. Bioinformatics approaches to RNA splicing. Master's thesis, The Sanger Centre, 2001.

[75] T. M. Lowe and S. R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, 25:955–964, 1997.

[76] A. V. Lukashin and M. Borodovsky. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, 26:1107–1115, 1998.

[77] C. Mathe, M. Sagot, T. Schiex, and P. Rouze. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, 30:4103–4117, 2002.

[78] I. Meyer and R. Durbin. Comparative ab initio gene prediction of gene structures using pair HMMs. *Bioinformatics*, 18:1309–1318, 2002.

[79] R. Mott. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Applic. Biosci.*, pages 477–478, 1997.

[80] E. W. Myers and W. Miller. Optimal alignments in linear space. *Comput. Applic. Biosci.*, 4(1):11–17, 1988.

[81] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computing Journal*, 7:308–313, 1965.

[82] L. Pachter, M. Alexandersson, and S. Cawley. Applications of generalized pair hidden markov models to alignment and gene finding problems. *Journal of Computational Biology*, 9:389–400, 2002.

[83] G. Parra, P. Agarwal, J.F. Abril, T. Wiehe, J.W. Fickett, and R. Guigo. Comparative gene prediction in human and mouse. *Genome Res.*, 13:108–117, 2003.

[84] G. Parra, E. Blanco, and R. Guigo. GeneID in Drosophila. *Genome Res.*, 10:511–515, 2000.

[85] N. Pavy, S. Rombauts, P. Dehais, C. Mathe, D. V. V.Ramana, P. Leroy, and P. Rouze. Evaluation of gene prediction software using a genomic data set: application to Arabidopsis thaliana sequences. *Bioinformatics*, 15:887–899, 1999.

[86] M. Pertea, X. Lin, and S. Salzberg. GeneSplicer: a new computational model for splice site prediction. *Nucleic Acids Res.*, pages 1185–1190, 2001.

[87] E. Polak and G. Ribiere. Note sur la convergence de methodes de directions conjuguees. *Revue Francaise d'informatique et de recherche operationelle*, 16:35–43, 1969. In French.

[88] W. H. Press, S. A. Teukolsky, W. Vetterling T., and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK, 1992.

[89] N. Drake R. Guigo, S. Knudsen and T. Smith. Prediction of gene structure. *J. Mol. Biol.*, 226:141–157, 1992.

[90] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[91] M. G. Reese, G. Hartzell, N. L. Harris, U. Ohler, J. F. Abril, and S. E. Lewis. Genome annotation assessment in Drosophila melanogaster. *Genome Res.*, 10:483–501, 2000.

[92] M. G. Reese, D. Kulp, H. Tammana, and D. Haussler. Genie : gene-finding in Drosophila melanogaster. *Genome Res.*, 10:529–538, 2000.

[93] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8, 2001.

[94] S. Rogic, A. Mackworth, and F. Oullette. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.*, 11:817–832, 2001.

[95] A. Roytberg, T. Astakhova, and M. Gelfand. Combinatorial approaches to gene recognition. *Computers and Chemistry*, 21:229–235, 1997.

[96] A. A. Salamov and V. V. Solovyev. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.*, 10:516–522, 2000.

[97] S. L. Salzberg, M. Pertea, A. L. Delcher, M. J. Gardner, and H. Tettelin. Interpolated markov models for eukaryotic gene finding. *Genomics*, 59(1):24–31., Jul 1 1999.

[98] M. Scherf, A. Klingenhoff, and T. Werner. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector. *J. Mol. Biol.*, 297:599–606, 2000.

[99] T. Schiex, A. Moisan, and P. Rouze. EuGene: a eukaryotic gene finder that combines several sources of evidence. In O. Gascuel and M. F. Sagot, editors, *Lecture Notes in Computer Science*, volume 2006, pages 111–125. Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 2001.

[100] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

[101] 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 Res.*, 21:607–613, 1993.

[102] E. E. Snyder and G. D. Stormo. Identification of protein coding regions in genomic DNA. *J. Mol. Biol.*, 248:1–18, 1995.

[103] 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 Res.*, 22:5156–5163, 1994.

[104] V. V. Solovyev, A. A. Salamov, and C. B. Lawrence. Identification of human gene structure using linear discriminant functions and dynamic programming. In C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak, editors, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 367–375, Menlo Park, CA, 1995. AAAI Press.

[105] R. Staden. Methods to define and locate patterns of motifs in sequences. *Comput. Applic. Biosci.*, 4(1):53–60, 1988.

[106] D. J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. F. Smith, editors. *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, CA, 1996. AAAI Press.

[107] L. Stein, P. Sternberg, R. Durbin, J. Thierry-Mieg, and J. Spieth. WormBase: network access to the genome and biology of Caenorhabditis elegans. *Nucleic Acids Res.*, 29:82–86, 2001.

[108] G. D. Stormo. Gene-finding approaches for eukaryotes. *Genome Res.*, 10:394–397, 2000.

[109] G. D. Stormo and D. Haussler. Optimally parsing a sequence into different classes based on multiple types of evidence. In States et al. [106], pages 369–375.

[110] J. Tabaska, R. Davuluri, and M. Zhang. Identifying the 3'-terminal exon in human DNA. *Bioinformatics*, 17:602–607, 2001.

[111] J. Tabaska and M. Zhang. Detection of polyadenylation signals in human DNA sequences. *Gene*, 231:77–86, 1999.

[112] The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[113] 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.

[114] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, pages 260–269, 1967.

[115] J. D. Watson, N. H. Hopkins, J. W. Roberts, J. A. Steitz, and A. M. Weiner. *Molecular Biology of the Gene*. Benjamin/Cummings, Menlo Park, CA, 1987.

[116] T. Wu. A segment-based dynamic programming algorithm for predicting gene structure. *Journal of Computational Biology*, 3:375–394, 1996.

[117] Y. Xu, J. R. Einstein, M. Shah, and E. C. Uberbacher. An improved system for exon recognition and gene modelling in human DNA sequences. In R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls, editors, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 376–383, Menlo Park, CA, 1994. AAAI Press.

[118] Y. Xu, R. J. Mural, and E. C. Uberbacher. Constructing gene models from accurately predicted exons: an application of dynamic programming. *Comput. Applic. Biosci.*, 10:613–623, 1994.

[119] R. Yeh, L. Lim, and C. Burge. Computational inference of homologous gene structures in the human genome. *Genome Res.*, 11:803–816, 2001.

[120] M. Zhang. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA*, 94:565–568, 1997.

[121] M. Zhang. Statistical analysis of human exons and their flanking regions. *Hum. Mol. Gen.*, 7:919–932, 1998.

[122] M. Zhang. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.*, 3:698–709, 2002.