Bibliography

[ABK96] F. J. Ayala, E. Barrio, and J. Kwiatowski. Molecular clock or erratic evolution? A tale of two genes. *Proceedings of the National Academy of Sciences of the USA*, 93:11729–11734, 1996.

[AGM$^+$90] 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.

[BD97] E. Birney and R. Durbin. Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. In Gaasterland et al. [GKK$^+$97], pages 56–64.

[BD00] E. Birney and R. Durbin. Using genewise in the drosophila annotation experiment. *Genome Research*, 10:547–548, 2000.

[BEK91] S. Brunak, J. Engelbrecht, and S. Knudsen. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology*, 220(1):49–65, 1991.

[Ber89] G. Bernardi. The isochore organization of the human genome. *Annual Review of Genetics*, 23:637–661, 1989.

[BFV$^+$97] J. G. Baldwin, L. M. Frisse, J. T. Vida, C. D. Eddleman, and W. K. Thomas. An evolutionary framework for the study of developmental evolution in a set of nematodes related to *Caenorhabditis elegans*. *Molecular Phylogenetics and Evolution*, 8:249–259, 1997.

[BG96] M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34:353–367, 1996.

[BH00] V. Bafna and D. H. Huson. The conserved exon method for gene finding. In R. Altman et al., editor, *Proceedings of the Eights International Conference on Intelligent Systems for Molecular Biology*, pages 3–12, Menlo Park, CA, 2000. AAAI Press.

[Bir87] A. Bird. CpG islands as gene markers in the vertebrate nucleus. *Trends in Genetics*, 3:342–347, 1987.

[Bis95] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, UK, 1995.

[BK97] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94, 1997.

[BP66] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[BPM$^+$00] S. Batzoglou, L. Pachter, J.P. Mesirov, B. Berger, and E.S. Lander. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Research*, 10:950–958, 2000.

[BSS00] M. Burset, I. A. Seledtsov, and V. V. Solovyev. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*, 28:4364–4375, 2000.

[Bur97] C. Burge. *Identification of genes in human genomic DNA.* PhD thesis, Stanford University, USA, 1997.

[CA96] J.-M. Claverie and S. Audic. The statistical significance of nucleotide position-weigth matrix matches. *Computer Applications in the Biosciences*, 12:431–439, 1996.

[CB86] J.-M. Claverie and K. Bougueleret. Heuristic informational analysis of sequences. *Nucleic Acids Research*, 14:179–196, 1986.

[Cla97] J.-M. Claverie. Computational methods for the identification of genes in vertebrate genomic sequences. *Human Molecular Genetics*, 6:1735–1744, 1997.

[Con01] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[Con02] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome, accepted for publication. *Nature*, 2002.

[DEKM98] 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.

[DMG95] L. Duret, D. Mouchiroud, and C. Gautier. Statistical analysis of vertebrate sequences reveals that long genes are scarce in CG-rich isochores. *Journal of Molecular Evolution*, 40:308–317, 1995.

[DS94] S. Dong and D. B. Searls. Gene structure prediction by linguistic methods. *Genomics*, 23:540–551, 1994.

[ea00] M. D. Adams et. al. The genome sequence of *Drosophila melanogaster*. *Science*, 24:2185–2195, 2000.

[Ens] Ensembl Webpage at http://www.ensembl.org/.

[eSC98] The *C. elegans* Sequencing Consortium. Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating. *Science*, 11:2012–2018, 1998.

[FLS92] R. Farber, A. Lapedes, and K. Sirotkin. Determintation of eukaryotic protein coding regions using neural networks and information theory. *Journal of Molecular Biology*, 226:471–479, 1992.

[FT92] J. W. Fickett and C. S. Tung. Assessment of protein coding measures. *Nucleic Acids Research*, 20:6441–6450, 1992.

[GAA+00a] R. Guigó, P. Agarwal, J. F. Abril, M. Burset, and J. W. Fickett. An assessment of gene prediction accuracy in large DNA sequences. *Genome Research*, 10:1631–1642, 2000.

[GAA+00b] R. Guigó, P. Agarwal, J. F. Abril, M. Burset, and J. W. Fickett. An assessment of gene prediction accuracy in large DNA sequences. *Genome Research*, 10:1631–1642, 2000.

[GF95] R. Guigó and J. W. Fickett. Distinctive sequence features in protein coding genic non-coding, and intergenic human DNA. *Journal of Molecular Biology*, 13:51–60, 1995.

[GKDS92] R. Guigó, S. Knudsen, N. Drake, and T. Smith. Prediction of gene structure. *Journal of Molecular Biology*, 226:141–157, 1992.

[GKK$^+$97] 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.

[GMP96] M. S. Gelfand, A. A. Mironov, and P. A. Pevzner. Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences of the USA*, 93:9061–9066, 1996.

[HHM90] X. Huang, R. C. Hardison, and W. Miller. A space-efficient algorithm for local similarities. *Computer Applications in the Biosciences*, 6:373–381, 1990.

[Hir75] D. S. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18:341–343, 1975.

[How71] R. A. Howard. *Dynamic Probabilistic Systems Volume II: Semi-Markov and Decision Processes.* John Wiley & Sons, New York, 1971.

[HRS$^+$87] N. H. Hopkins, J. W. Roberts, J. A. Steitz, J. D. Watson, and A. M. Weiner. *Molecular Biology of the Gene.* Benjamin Cummings, 4th edition, 1987.

[Ini00] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 14:796–815, 2000.

[JBD99] N. Jareborg, E. Birney, and R. Durbin. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Research*, 9:815–824, 1999.

[JMCB90] I. Sauvaget J.-M. Claverie and K. Bougueleret. K-tuple frequency analysis: from intron/exon discrimination to t-cell epitope mapping. *Methods in Enzymology*, 183:237–252, 1990.

[KAA$^+$93] B. P. Kennedy, E. J. Aamodt, F. L. Allen, M. A. Chung, M. F.P. Heschl, and J. D. McGhee. The gut esterase gene (ges-1) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Journal of Molecular Biology*, 229:890–908, 1993.

[KFDB01] I. Korf, P. Flicek, D. Duan, and M. R. Brent. Integrating genomic homology into gene structure prediction. *Bioinformatics*, 1:1–9, 2001.

[KHRE96] 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 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*, pages 134–142, Menlo Park, CA, 1996. AAAI Press.

[Koz81] M. Kozak. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Research*, 9:5233–5252, 1981.

[Kro97] A. Krogh. Two methods for improving performance of a HMM and their application for gene finding. In Gaasterland et al. [GKK$^+$97], pages 179–186.

[KZ00] W.J. Kent and A.M. Zahler. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae–C. elegans* genomic alignment. *Genome Research*, 10:1115–1125, 2000.

[LD01] A. Levine and R. Durbin. (unpublished data). 2001.

[MBD97] S. D. Martinelli, C. G. Brown, and R. Durbin. Gene expression and development databases for *C. elegans*. *Seminars in Cell and Developmental Biology*, 8:459–467, 1997.

[McL92] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, USA, 1992.

[MD02] I. M. Meyer and R. Durbin. Comparative *ab initio* prediction of gene structures using pair HMMs. *Bioinformatics*, 18:1309–1318, 2002.

[NGM01] P. S. Novichkov, M. S. Gelfand, and A. A. Mironov. Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics*, 17:1011–1018, 2001.

[OMM+97] J. C. Oeltjen, T. M. Malley, D. M. Muzny, W. Miller, R. A. Gibbs, and J. W. Belmont. Large-scale comparative sequence analysis of the human and the murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Research*, 7:315–329, 1997.

[Pac99] L. Pachter. *Domino Tiling, Gene Recognition, and Mice.* PhD thesis, Massachusetts Institute of Technology, USA, 1999.

[PAC01] L. Pachter, M. Alexandersson, and S. Cawley. Applications of generalized pair hidden markov models to alignment and gene finding problems. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology RECOMB 2001*, 2001.

[Rab89] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[RHH+00] M. G. Reese, G. Hartzell, N. L. Harris, U. Ohler, J. F. Abril, and S. E. Lewis. Genome annotation assessment. *Genome Research*, 10:483–501, 2000.

[SD94] E. Sonnhammer and R. Durbin. A workbench for large scale sequence homology analysis. *Computer Applications in the Biosciences*, 10:301–307, 1994.

[SD96] E. Sonnhammer and R. Durbin. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, 167:GC1–10, 1996.

[SDFH97] S. Salzberg, A. Delcher, K. Fasman, and J. Henderson. A decision tree system for finding genes in DNA. *Technical report John Hopkins University*, 1997.

[SH94] G. D. Stormo and D. Haussler. Optimally parsing a sequence into different classes based on multiple types of evidence. 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 369–375, Menlo Park, CA, 1994. AAAI Press.

[SMS+98] G. Stoesser, M. Moseley, J. Sleep, M. McGowran, M. Garcia-Pator, and P. Sterk. The EMBL nucleotide sequence database. *Nucleic Acids Research*, 26:8–15, 1998.

[SS93] 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.

[SS00] A. A. Salamov and V. V. Solovyev. Ab Initio Gene Finding in *Drosophila* Genomic DNA. *Bioinformatics*, 10:516–522, 2000.

[SSD+01] 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 Research*, 29:82–86, 2001.

[SSL96] 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, 1996.

[UM91] E. C. Uberbacher and R. J. Mural. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proceedings of the National Academy of Sciences of the USA*, 88:11261–11265, 1991.

[Vit67] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, pages 260–269, 1967.

[VPS98] D. A. Voronov, Y. V. Panchin, and S. E. Spiridonov. Nematode phylogeny and embryology. *Nature*, 395:28, 1998.

[WGJMOG01] T. Wiehe, S. Gebauer-Jung, T. Mitchell-Olds, and R. Guigó. SGP-1: Prediction and validation of homologous genes bases on sequence alignments. *Genome Research*, 11:1574–1583, 2001.

[WGM00] T. Wiehe, R. Guigó, and W. Miller. Genome sequence comparisons: Hurdles in the fast lane to functional genomics. *Briefings in Bioinformatics*, 1:381–388, 2000.

[Wor] Wormbase Webpage at http://www.wormbase.org/.

[YLB01] R. Yeh, L. P. Lim, and C. B. Burge. Computational inference of homologous gene structures in the human genome. *Genome Research*, 11:803–816, 2001.

[Zha97] M. Q. Zhang. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proceedings of the National Academy of Sciences of the USA*, 94:565–568, 1997.