Bibliography

[AB94] T. K. Attwood and M. E. Beck. PRINTS – a protein motif fingerprint database. *Protein Engineering*, 7:841–848, 1994.

[AF94] T. K. Attwood and J. B. Findlay. Fingerprinting G-protein-coupled receptors. *Protein Engineering*, 7:195–203, 1994.

[AG96] S. F. Altschul and W. Gish. Local alignment statistics. *Methods in Enzymology*, 266:460–480, 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.

[AKW+98] N. R. Ashcroft, M. E. Kosinski, D. Wickramasinghe, P. J. Donovan, and A. Golden. The four cdc25 genes from the nematode Caenorhabditis elegans. *Gene*, 214:59–66, 1998.

[AMS+97] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[BB98] M. L. Beçak and W. Beçak. Evolution by polyploidy in Amphibia: new insights. *Cytogenetic Cell Genetics*, 80:28–33, 1998.

[BC94] P. Baldi and Y. Chauvin. Smooth on-line learning algorithms for hidden Markov models. *Neural Computation*, 6(2):307–318, 1994.

[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.

[BH96] P. Bucher and K. Hofmann. A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In States et al. [SAG+96], pages 44–51.

[BH97] L. Bloom and H. R. Horvitz. The Caenorhabditis elegans gene unc-76 and its human homologs define a new gene family involved in axonal outgrowth and fasciculation. *Proceedings of the National Academy of Sciences of the USA*, 94:3414–3419, 1997.

[BHK+93] M. Brown, R. Hughey, A. Krogh, I. S. Mian, K. Sjölander, and D. Haussler. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In L. Hunter, D. B. Searls, and J. Shavlik, editors, *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, pages 47–55, Menlo Park, CA, 1993. AAAI Press.

[BHP89] K. Banki, D. Halladay, and A. Perl. Cloning and expression of the human gene for transaldolase: A novel highly repetitive element constitutes an integral part of the coding sequence. *Journal of Biological Chemistry*, 269:2847–2851, 1989.

[Bir] Ewan Birney. Personal communication.

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

[BJVU98] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. Preprint, to appear in Genome Research, 1998.

[Bla98] M. Blaxter. Caenorhabditis elegans is a nematode. *Science*, 282:2041–2046, 1998.

[Blu98] T. Blumenthal. Gene clusters and polycistronic transcription in eukaryotes. *Bioessays*, 20:480–487, 1998.

[BPS98] C. B. Burge, R. A. Padgett, and P. A. Sharp. Evolutionary fates and origins of U12-type introns. *Molecular Cell*, 2:1–20, 1998.

[BR93] T. R. Burglin and G. Ruvkun. The Caenorhabditis elegans homeobox gene cluster. *Current Opinion in Genetics and Development*, 3:615–620, 1993.

[BS87] G. J. Barton and M. J. E. Sternberg. A strategy for the rapid multiple alignment of protein sequences. *Journal of Molecular Biology*, 198:327–337, 1987.

[BS97] T. Blumenthal and K. Steward. RNA processing and gene structure. In D. Riddle, T. Blumenthal, B. Meyer, and J. Priess, editors, *C.elegans II*, chapter 6, pages 117–145. Cold Spring Harbor Laboratory Press, Plainview, NY, 1997.

[BTR98] C. Brown, K. Todd, and R. F. Rosenzweig. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Molecular Biology and Evolution*, 15:931–942, 1998.

[Bul86] M. Bulmer. Neighbouring base effects on substitution rates in pseudogenes. *Molecular Biology and Evolution*, 3:322–329, 1986.

[Bur98] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. Available from http://svm.research.bell-labs.com/SVMdoc.html, 1998.

[But98] B. A. Butler. Sequence analysis using GCG. *Methods of Biochemical Analysis*, 39:74–97, 1998.

[CFA89] J. Collins, E. Forbes, and P. Anderson. The Tc3 family of transposable elements in Caenorhabditis elegans. *Nature*, 328:726–728, 1989.

[Cho92] C. Chothia. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992.

[CK98] M. Cline and K. Karplus. On alignment shift and its measures. Technical report, University of California Santa Cruz, 1998. Available from http://www.cse.ucsc.edu/~cline/shiftscore.html.

[CL88] H. Carrillo and D. Lipman. The multiple sequence alignment problem in biology. *SIAM Journal of Applied Mathematics*, 48:1073–1082, 1988.

[Cla98] M. Clamp. Jalview. In press, 1998.

[CLB93] R. Conrad, R. Liou, and T. Blumenthal. Functional analysis of a C.elegans trans-splice acceptor. *Nucleic Acids Research*, 21(4):913–919, 1993.

[Cra95] N. Craig. Unity in transposition reactions. *Science*, 270:253–254, 1995.

[CSC98] The C.elegans Sequencing Consortium. Genome sequence of the nematode Caenorhabditis elegans. A platform for investigating biology. *Science*, 282:2012–2018, 1998.

[CvLP94] S. D. Colloms, H. G. van Luenen, and R. H. Plasterk. DNA binding activities of the Caenorhabditis elegans Tc3 transposase. *Nucleic Acids Research*, 22:5548–5554, 1994.

[DDJH94] T. Doak, F. Doerder, C. Jahn, and G. Herrick. A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. *Proceedings of the National Academy of Sciences of the USA*, 91:942–946, 1994.

[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.

[DGPB98] L. Duret, N. Guex, M. C. Peitsch, and A. Bairoch. New insulin-like proteins with atypical disulfide bond pattern characterized in Caenorhabditis elegans by comparative sequence analysis and homology modeling. *Genome Research*, 8:348–353, 1998.

[DH97] R. E. Davis and S. Hodgson. Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in Schistosoma mansoni. *Molecular and Biochemical Parasitology*, 89:25–39, 1997.

[DHL97a] D. Drasdo, T. Hwa, and M. Lässig. DNA sequence alignment and critical phenomena. Available from http://matisse.ucsd.edu/~hwa/pub.html, 1997.

[DHL97b] D. Drasdo, T. Hwa, and M. Lässig. Scaling laws and similarity detection in sequence alignment with gaps. Preprint, available from http://matisse.ucsd.edu/~hwa/pub.html, 1997.

[DIB97] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[DSO78] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, supplement 3, pages 345–352. National Biomedical Research Foundation, Washington, DC, 1978.

[ED95] F. H. Eeckman and R. Durbin. ACeDB and Macace. *Methods in Cell Biology*, 48:583–605, 1995.

[Edd95] S. R. Eddy. Multiple alignment using hidden Markov models. In Rawlings et al. [RCA+95], pages 114–120.

[Edd96] S. R. Eddy. Hidden Markov models. *Current Opinion in Structural Biology*, 6:361–365, 1996.

[Eis98] J. A. Eisen. A phylogenomic study of the MutS family of proteins. *Nucleic Acids Research*, 26(18):4291–4300, 1998.

[EMD95] S. R. Eddy, G. J. Mitchison, and R. Durbin. Maximum discrimination hidden Markov models of sequence consensus. *Journal of Computational Biology*, 2:9–23, 1995.

[EZM+97] D. Evans, D. Zorio, M. MacMorris, C. E. Winter, K. Lea, and T. Blumenthal. Operons and SL2 trans-splicing exist in nematodes outside the genus Caenorhabditis. *Proceedings of the National Academy of Sciences of the USA*, 94:9751–9756, 1997.

[FAW+95] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269:496–512, 1995.

[FBT+91] D. Fitch, W. Bailey, D. Tagle, M. Goodman, L. Sieu, and J. Slightom. Duplication of the γ-globin gene mediated by L1 long

interspersed repetitive elements in an early ancestor of simian primates. *Proceedings of the National Academy of Sciences of the USA*, 88:7396–7400, 1991.

[FD87] D.-F. Feng and R. F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, 25:351–360, 1987.

[FLD+94] G. Franz, T. G. Loukeris, G. Dialektaki, C. R. Thompson, and C. Savakis. Mobile Minos elements from Drosophila hydei encode a two-exon transposase with similarity to the paired DNA-binding domain. *Proceedings of the National Academy of Sciences of the USA*, 91:4746–4750, 1994.

[GFF] GFF: an exchange format for gene-finding features. Webpage at http://www.sanger.ac.uk/Software/GFF/.

[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.

[GL95a] N. Grindley and A. Leschziner. DNA transposition: from a black box to a color monitor. *Cell*, 83:1063–1066, 1995.

[GL95b] X. Gu and W-H. Li. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *Journal of Molecular Evolution*, 40:464–473, 1995.

[Gol98] N. Goldman. Phylogeny with alignment uncertainty. Preprint, Isaac Newton Institute for the Mathematical Sciences, Univ. of Cambridge, UK, 1998.

[Got82] O. Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708, 1982.

[Got96] O. Gotoh. Significant improvement in accuracy of multiple protein alignments by iterative refinement as assessed by reference to structural alignments. *Journal of Molecular Biology*, 264:823–838, 1996.

[GPL] The GNU Public License. Available in full from http://www.fsf.org/copyleft/gpl.html.

[GV96] M. Gribskov and S. Veretnik. Identification of sequence patterns with profile analysis. *Methods in Enzymology*, 266:198–212, 1996.

[GW92] P. Gruss and C. Walther. Pax in development. *Cell*, 69:719–722, 1992.

[GWD98] P. Gibbs, W. Witke, and A. Dugaiczyk. The molecular clock runs at different rates among closely related members of a gene family. *Journal of Molecular Evolution*, 46:552–561, 1998.

[Hau98] D. Haussler. Computational genefinding. Preprint; available from http://www.cse.ucsc.edu/research/compbio/research.html, 1998.

[HBF92] D. G. Higgins, A. J. Bleasby, and R. Fuchs. CLUSTAL V: improved software for multiple sequence alignment. *Computer Applications in the Biosciences*, 8(2):189–191, 1992.

[HD98] I. Holmes and R. Durbin. Dynamic programming alignment accuracy. *Journal of Computational Biology*, 5(3):493–504, 1998.

[HH91] S. Henikoff and J. G. Henikoff. Automated assembly of protein blocks for database searching. *Nucleic Acids Research*, 19:6565–6572, 1991.

[HH92] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the USA*, 89:10915–10919, 1992.

[HK94] D. S. Horowitz and A. R. Krainer. Mechanisms for selecting 5' splice sites in mammalian pre-mRNA splicing. *Trends in Genetics*, 10:100–106, 1994.

[HK96] R. Hughey and A. Krogh. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Computer Applications in the Biosciences*, 12:95–107, 1996.

[HKMS93] D. Haussler, A. Krogh, I. S. Mian, and K. Sjölander. Protein modeling using hidden Markov models: analysis of globins. In T. N. Mudge, V. Milutinovic, and L. Hunter, editors, *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*, volume 1, pages 792–802, Los Alamitos, CA, 1993. IEEE Computer Society Press.

[HKY85] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

[HL96] T. Hwa and M. Lässig. Similarity-detection and localization. *Physical Review Letters*, 76:2591, 1996.

[HLL97] D. Hartl, A. Lohe, and E. Lozovskaya. Modern thoughts on an ancyent *Marinere*: function, evolution, regulation. *Annual Review of Genetics*, 31:337–358, 1997.

[HLNL97] D. Hartl, E. Lozovskaya, D. Nurminsky, and A. Lohe. What restricts the activity of *mariner*-like transposable elements? *Trends in Genetics*, 13(5):197–201, 1997.

[HS89] D. G. Higgins and P. M. Sharp. Fast and sensitive multiple sequence alignments on a microcomputer. *Computer Applications in the Biosciences*, 5:151–153, 1989.

[Hug98] A. L. Hughes. Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosomes 6, 9 and 1. *Molecular Biology and Evolution*, 15:854–870, 1998.

[Hwa96] T. Hwa. From flux pinning to DNA sequence alignment. In K. Fujikawa and Y. A. Ono, editors, *Quantum Coherence and Decoherence*, pages 109–114. Elsevier, 1996.

[Jay86] E. T. Jaynes. Bayesian methods: general background. In J. H. Justice, editor, *Maximum Entropy and Bayesian Methods in Applied Statistics*, pages 1–25. Cambridge University Press, Cambridge, UK, 1986.

[JB97] R. C. Johnsen and D. L. Baillie. Mutation. In D. Riddle, T. Blumenthal, B. Meyer, and J. Priess, editors, *C.elegans II*, chapter 4, pages 79–95. Cold Spring Harbor Laboratory Press, Plainview, NY, 1997.

[JC69] T. H. Jukes and C. Cantor. Evolution of protein molecules. In *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[JH98] T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. Preprint, Dept. of Computer Science, Univ. of California, available from http://www.cse.ucsc.edu/~haussler/pubs.html, 1998.

[Jur] J. Jurka. Personal communication.

[Jur98] J. Jurka. Repeats in genomic DNA: mining and meaning. *Current Opinion in Structural Biology*, 8:333–337, 1998.

[KA90] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences of the USA*, 87:2264–2268, 1990.

[KA93] S. Karlin and S. F. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences of the USA*, 90:5873–5877, 1993.

[Kab95] D. B. Kaback. Yeast genome structure. In A. E. Wheals, A. H. Rose, and J. S. Harrison, editors, *The Yeasts*, volume 6, pages 179–222. Academic, London, 1995.

[KB95] S. Karlin and C. Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*, 11:283–290, 1995.

[KBM+94] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. Hidden Markov models in computational biology: applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, Feb. 1994.

[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 States et al. [SAG+96], pages 134–142.

[Kim80] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[Kim83] M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK, 1983.

[KM95] A. Krogh and G. J. Mitchison. Maximum entropy weighting of aligned sequences of proteins or DNA. In Rawlings et al. [RCA+95], pages 215–221.

[KMH94] A. Krogh, I. S. Mian, and D. Haussler. A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Research*, 22:4768–4778, 1994.

[Kro94] 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.

[KT75] S. Karlin and H. Taylor. *A First Course in Stochastic Processes.* Academic Press, San Diego, CA, 1975.

[LAB+93] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993.

[LAK89] D. J. Lipman, S. F. Altschul, and J. D. Kececioglu. A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences of the USA*, 86:4412–4415, 1989.

[LC95] M. R. Leroux and E. P. Candido. Characterization of four new tcp-1-related cct genes from the nematode Caenorhabditis elegans. *DNA and Cell Biology*, 14:951–60, 1995.

[LC97] M. Labrador and V. Corces. Transposable element-host interactions: regulation of insertion and excision. *Annual Review of Genetics*, 31:381–404, 1997.

[LCR96] D. Lampe, M. Churchill, and H. Robertson. Purified *mariner* transposase is sufficient to mediate transposition in *vitro*. *EMBO Journal*, 15:5470–5479, 1996.

[LG91] W-H. Li and D. Graur. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA, 1991.

[LKW97] T. Lindahl, P. Karran, and R. D. Wood. DNA excision repair pathways. *Current Opinion in Genetics and Development*, 7:158–169, 1997.

[LL80] L. D. Landau and E. M. Lifshitz. *Statistical Physics Part I*, volume 5 of *Course of Theoretical Physics*. Pergamon Press, Oxford, UK, 1980.

[Mac92a] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

[Mac92b] D. J. C. MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1992.

[Mac96a] D. J. C. MacKay. Density networks and their application to protein modelling. In *Maximum Entropy and Bayesian Methods*, pages 259–268. Kluwer Academic Publishers, Netherlands, 1996.

[Mac96b] D. J. C. MacKay. Equivalence of Boltzmann chains and hidden Markov models. *Neural Computation*, 8(1):178–181, 1996.

[Mac97] D. J. C. MacKay. Introduction to Gaussian processes. Available from http://wol.ra.phy.cam.ac.uk/mackay/, 1997.

[MD95] G. J. Mitchison and R. Durbin. Tree-based maximal likelihood substitution matrices and hidden Markov models. *Journal of Molecular Evolution*, 41:1139–1151, 1995.

[Miy94] S. Miyazawa. A reliable sequence alignment method based on probabilities of residue correspondence. *Protein Engineering*, 8:999–1009, 1994.

[MKW91] D. G. Moerman, J. E. Kiff, and R. H. Waterston. Germline excision of the transposable element Tc1 in C.elegans. *Nucleic Acids Research*, 19:5669–5672, 1991.

[MM88] E. W. Myers and W. Miller. Optimal alignments in linear space. *Computer Applications in the Biosciences*, 4(1):11–17, 1988.

[MV96] H. T. Mevissen and M. Vingron. Quantifying the local reliability of a sequence alignment. *Protein Engineering*, 9(2):127–132, 1996.

[MVF94] M. A. McClure, T. K. Vasi, and W. M. Fitch. Comparative analysis of multiple protein-sequence alignment methods. *Journal of Molecular Evolution*, 11:571–592, 1994.

[NCC+92] G. Naclerio, G. Cangiano, A. Coulson, A. Levitt, V. Ruvolo, and A. La Volpe. Molecular and genomic organization of clusters of repetitive DNA sequences in Caenorhabditis elegans. *Journal of Molecular Biology*, 226:159–168, 1992.

[Nea] R. M. Neal. Personal communication.

[Nea96] R. M. Neal. *Bayesian Learning in Neural Networks*. Springer (Lecture Notes in Statistics), New York, Berlin, 1996.

[Nea98] R. M. Neal. Annealed importance sampling. Technical report no. 9805, Dept. of Statistics, Univ. of Toronto, available from http://www.cs.utoronto.ca/~radford/, 1998.

[NH93] R. M. Neal and G. E. Hinton. A new view of the EM algorithm that justifies incremental and other variants. Preprint, Dept. of Computer

Science, Univ. of Toronto, available from ftp://archive.cis.ohio-state.edu/pub/neuroprose/neal.em.ps.Z, 1993.

[NH96] C. Notredame and D. G. Higgins. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Research*, 24(8):1515–1524, 1996.

[NKML98] L. S. Nelson, K. Kim, J. E. Memmott, and C. Li. FMRFamide-related gene family in the nematode, Caenorhabditis elegans. *Brain Research. Molecular Brain Research*, 58:103–111, 1998.

[NW70] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[OGB95] T. Oosumi, B. Garlick, and W. Belknap. Identification and characterization of putative transposable DNA elements in solanaceous plants and Caenorhabditis elegans. *Proceedings of the National Academy of Sciences of the USA*, 92:8886–8890, 1995.

[OGB96] T. Oosumi, B. Garlick, and W. Belknap. Identification of putative nonautonomous transposable elements associated with several transposon families in Caenorhabditis elegans. *Journal of Molecular Evolution*, 43:11–18, 1996.

[Ohn70] S. Ohno. *Evolution by gene duplication.* Springer Verlag, Berlin/Heidelberg/New York, 1970.

[Oht89] T. Ohta. Role of gene duplication in evolution. *Genome*, 31:304–310, 1989.

[Oht91] T. Ohta. Multigene families and the evolution of complexity. *Journal of Molecular Evolution*, 33:34–41, 1991.

[Ols91] M. V. Olson. Genome structure and organization in Saccharomyces cerevisiae. In J. R. Broach, J. R. Pringle, and E. W. Jones, editors,

The Molecular and Cellular Biology of the Yeast Saccharomyces, volume 1, pages 1–40. Cold Spring Harbor Laboratory Press, New York, 1991.

[PBBC96] A. G. Pedersen, P. Baldi, S. Brunak, and Y. Chauvin. Characterization of prokaryotic and eukaryotic promoters using hidden Markov models. In States et al. [SAG+96], pages 182–191.

[PL88] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA*, 4:2444–2448, 1988.

[Pov98] M. Povinelli. Density networks with application to protein modelling. M.Phil thesis, Univ. of Cambridge, UK. Available from http://www.mit.edu/~mpovinel/pub.html, 1998.

[PS98] V. E. Papaioannou and L. M. Silver. The T-box gene family. *Bioessays*, 20:9–19, 1998.

[PvL97] R. Plasterk and H. van Luenen. Transposons. In D. Riddle, T. Blumenthal, B. Meyer, and J. Priess, editors, *C.elegans II*, chapter 5, pages 97–116. Cold Spring Harbor Laboratory Press, Plainview, NY, 1997.

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

[RCA+95] 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*, Menlo Park, CA, 1995. AAAI Press.

[REKH97] M. G. Reese, F. H. Eeckman, D. Kulp, and D. Haussler. Improved splice site detection in Genie. *Journal of Computational Biology*, 4(3):311–323, 1997.

[Rob98] H. M. Robertson. Two large families of chemoreceptor genes in the nematodes Caenorhabditis elegans and Caenorhabditis briggsae reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Research*, 8:449–463, 1998.

[RvLDP97] R. Rezsohazy, H. van Luenen, R. Durbin, and R. Plasterk. Tc7, a Tc1-hitch hiking transposon in Caenorhabditis elegans. *Nucleic Acids Research*, 25(20):4048–4054, 1997.

[SAG+96] 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.

[SAL+95] P. M. Sharp, M. Averof, A. T. Lloyd, G. Matassi, and J. F. Peden. DNA sequence evolution: the sounds of silence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 349:241–247, 1995.

[SD94] E. Sonnhammer and R. Durbin. A workbench for large-scale sequence homology analysis. *Computer Applications in the Biosciences*, 10(3):301–307, 1994.

[SD97] E. L. Sonnhammer and R. Durbin. Analysis of protein domain families in Caenorhabditis elegans. *Genomics*, 46:200–216, 1997.

[SDF+96] C. Savage, P. Das, A. L. Finelli, S. R. Townsend, C. Y. Sun, S. E. Baird, and R. W. Padgett. Caenorhabditis elegans genes

sma-2, sma-3, and sma-4 define a conserved family of transforming growth factor beta pathway components. *Proceedings of the National Academy of Sciences of the USA*, 93:790–794, 1996.

[SDT+92] J. Sulston, Z. Du, K. Thomas, R. Wilson, L. Hillier, R. Staden, N. Halloran, P. Green, J. Thierry-Mieg, L. Qiu, et al. The C.elegans genome sequencing project: a beginning. *Nature*, 356:37–41, 1992.

[SEB+98] E. Sonnhammer, S. Eddy, E. Birney, A. Bateman, and R. Durbin. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research*, 26(1):320–322, 1998.

[Sha97] E. I. Shakhnovich. Theoretical studies of protein folding thermodynamics and kinetics. *Current Opinion in Structural Biology*, 7:29–40, 1997.

[Sid96] A. Sidow. Gen(om)e duplications in the evolution of early vertebrates. *Current Opinion in Genetics and Development*, 6:715–722, 1996.

[SK83] D. Sankoff and J. B. Kruskal. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, London, 1983.

[SKB+96] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Computer Applications in the Biosciences*, 12(4):327–345, 1996.

[SKR89] K. Schughart, C. Kappen, and F. Ruddle. Duplication of large genomic regions during the evolution of vertebrate homeobox genes. *Proceedings of the National Academy of Sciences of the USA*, 86:7067–7071, 1989.

[SLP+96] T. A. Starich, R. Y. Lee, C. Panzarella, L. Avery, and J. E. Shaw. eat-5 and unc-7 represent a multigene family in Caenorhabditis elegans involved in cell-cell coupling. *Journal of Computational Biology*, 134:537–548, 1996.

[Smi87] M. M. Smith. Molecular evolution of the Saccharomyces cerevisiae histone gene loci. *Journal of Molecular Evolution*, 24:252–259, 1987.

[Smi96] A. Smit. The origin of interspersed repeats in the human genome. *Current Opinion in Genetics*, 6:743–748, 1996.

[SO95] A. Shinohara and T. Ogawa. Homologous recombination and the roles of double-strand breaks. *Trends in the Biochemical Sciences*, 20:387–391, 1995.

[SR96] A. Smit and A. Riggs. *Tiggers* and other DNA transposon fossils in the human genome. *Proceedings of the National Academy of Sciences of the USA*, 93:1443–1448, 1996.

[SW81] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[Tat] R. Tatusov. Personal communication.

[Tay87] W. R. Taylor. Multiple sequence alignment by a pairwise algorithm. *Computer Applications in the Biosciences*, 3:81–87, 1987.

[TCD+95] E. R. Troemel, J. H. Chou, N. D. Dwyer, H. A. Colbert, and C. I. Bargmann. Divergent seven transmembrane receptors are candidate chemosensory receptors in C.elegans. *Cell*, 83:207–218, 1995.

[THG94a] 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.

[THG94b] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Computer Applications in the Biosciences*, 10:19–29, 1994.

[TK98] H. Tachida and T. Kuboyama. Evolution of multigene families by gene duplication: a haploid model. *Genetics*, 149:2147–2158, 1998.

[TKF91] J. L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, 33:114–124, 1991.

[TKF92] J. L. Thorne, H. Kishino, and J. Felsenstein. Inching toward reality: an improved likelihood model of sequence evolution. *Methods in Enzymology*, 34:3–16, 1992.

[TRTA98] R. Tarrio, F. Rodriguez-Trelles, and F. J. Ayala. New Drosophila introns originate by duplication. *Proceedings of the National Academy of Sciences of the USA*, 95:1658–1662, 1998.

[TW96] M. G. Tomlinson and M. D. Wright. A new transmembrane 4 superfamily molecule in the nematode, Caenorhabditis elegans. *Journal of Molecular Evolution*, 43:312–314, 1996.

[VBA+98] P. Vincens, L. Buffat, C. Andre, J. P. Chevrolat, J. F. Boisvieux, and S. Hazout. A strategy for finding regions of similarity in complete genome sequences. *Bioinformatics*, 14:715–725, 1998.

[VBP96] J. Vos, I. De Baere, and R. Plasterk. Transposase is the only nematode protein required for in vitro transposition of Tc1. *Genes and Development*, 10:755–761, 1996.

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

[vLCP94] H. van Luenen, S. Colloms, and R. Plasterk. The mechanism of transposition of Tc3 in C.elegans. *Cell*, 79:293–301, 1994.

[VP94] J. C. Vos and R. H. Plasterk. Tc1 transposase of Caenorhabditis elegans is an endonuclease with a bipartite DNA binding domain. *EMBO Journal*, 13:6125–6132, 1994.

[VvLP93] J. C. Vos, H. G. van Luenen, and R. H. Plasterk. Characterization of the Caenorhabditis elegans Tc1 transposase in vivo and in vitro. *Genes and Development*, 7:1244–1253, 1993.

[VW94] M. Vingron and M. S. Waterman. Sequence alignment and penalty choice: review of concepts, case studies and implications. *Journal of Molecular Biology*, 235:1–12, 1994.

[Wat95] M. S. Waterman. *Introduction to Computational Biology.* Chapman & Hall, London, 1995.

[WE87] M. S. Waterman and M. Eggert. A new algorithm for best subsequence alignments with application to tRNA–rRNA comparisons. *Journal of Molecular Biology*, 197:723–725, 1987.

[Wes89] S. Wessler. The splicing of maize transposable elements from pre-mRNA - a minireview. *Gene*, 82:127–133, 1989.

[Woo94] J. C. Wootton. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computational Chemistry*, 18:269–85, 1994.

[WP84] M. S. Waterman and M. D. Perlwitz. Line geometries for sequence comparisons. *Bulletin of Mathematical Biology*, 46:567–577, 1984.

[WS97] K. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387:708–713, 1997.

[WS98] R. Waterston and J. E. Sulston. The Human Genome Project: reaching the finish line. *Science*, 282:53–54, 1998.

[YWB97] J. Yoder, C. Walsh, and T. Bestor. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, 13(8):335–340, 1997.

[ZB96] H. Zhang and T. Blumenthal. Functional analysis of an intron 3' splice site in Caenorhabditis elegans. *RNA*, 2(4):380–388, 1996.

[ZLL97] J. Zhu, J. Liu, and C. Lawrence. Bayesian adaptive alignment and inference. In Gaasterland et al. [GKK+97], pages 358–368.

[ZLL98] J. Zhu, J. S. Liu, and C. E. Lawrence. Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, 14:25–39, 1998.

[ZR95] M. Zetka and A. Rose. The genetics of meiosis in Caenorhabditis elegans. *Trends in Genetics*, 11:27–31, 1995.