References

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, et al.. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).

S. F. Altschul, T. L. Madden, A. S. Alejandro, 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).

S. Asakawa, K. Tsunematsu, A. Takayanagi, T. Sasaki, A. Shimizu, K. Kawasaki, A. J. Mungall, S. Beck, S. Minoshima, and N. Shimizu. The Genomic Structure and Promoter Region of the Human Parkin Gene. *Biochemical and Biophysical Research Communications* **286**, 863–868 (2001).

S. Audic and E. Béraud-Columb. Detection of eukaryotic promoters using Markov transition matrices. *Computation and Chemistry* **21**, 223-228 (1997).

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Seconding International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, 1994.

Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-DNA binding sites. In *Proceedings of RECOMB*, 2003.

A. Barberis and L. Gaudreau. Recruitment of the RNA polymerase II holoenzyme and its implications in gene regulation. *Biological Chemistry* **379**, 1397–1405 (1998).

A. Bateman, E. Birney, L.Cerutti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. The Pfam Protein Families Database. *Nucleic Acids Research* **30**, 276–280 (2002).

E. Birney. *Sequence alignment in bioinformatics.* Ph.D. thesis, Sanger Centre, 1999.

C. M. Bishop, M. E. Tipping. Variational Relevance Vector Machines. In *Proceedings on the 16th Conference in Uncertainty in Artificial Intelligence*, pages 46–53, 2000.

B. J. Blencowe. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases.. *Trends in Biochemical Science* **25**, 106–110 (2000).

M. Brandeis, M. Ariel, and H. Cedar. Dynamics of DNA methylation during development. *Bioessays* **15**, 709–713 (1993).

S. E. Brenner, C. Chothia, and T. J. P. Hubbard. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Science* **95**, 6073–6078 (1998).

J. E. Brownell, J. Zhou., T. Ranalli., R. Kobayashi, D. G. Edmondson, S. Y. Roth, C. D. Allis. *Tetrahymena* histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation.. *Cell* **84**, 843–851 (1996).

P. Bucher. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences.. *Journal of Molecular Biology* **212**, 563–578 (1990).

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

Martyn Byng. *A statistical model for locating regulatory regions in novel DNA sequences.* Ph.D. thesis, University of Reading, 2001.

E. Camon, M. Magrane, D. Barrel, D. Binns, W. Fleishmann, P. Kersey, N. Mulder, T. Oinn, and R. Apweiller. The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Research* **13**, 662-672 (2003).

M. F. Carey. Transcriptional activation. A holistic view of the complex. *Current Biology* **5**, 1003-1005 (1995).

The *C. elegans* Sequencing Consortium. Genome Sequence of the Nemetode *C. elegans*: A Platform for Investigating Biology. *Science* **282**, 2012–2018 (1998).

A. Chess, I. Simon, H. Cedar, and R. Axel. Allelic Inactivation Regulates Olfactory Receptor Gene Expression. *Cell* **78**, 823–834 (1994).

M. Clamp, D. Andrews, D. Barker, P. Bevan, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, T. Hubbard, A. Kasprzyk, D. Keefe, H. Lehvaslaiho, V. Iyer, C. Melsopp, E. Mongin, R. Pettett, 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 E. Birney. Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Research* **31**, 38–42 (2003).

J. Corden, B. Wasylyk, A. Buchwalder, P. Sassone-Corsi, C. Kedinger, P. Chambon. Promoter sequences of eukaryotic protein-coding genes. *Science* **209**, 1406–1414 (1980).

C. Crasto, M. S. Singer, and G. M. Shepherd. The olfactory receptor family album. *Genome Biology* **2**, 1027.1–1027.4 (2001).

D. di Bernado, T. A. Down, and T. J. P. Hubbard. aReNA: Detection of conserved secondary structures in multiple alignments. *Bioinformatics* (Submitted).

R. D. Dowell, R. M. Jokerst, A. Day, S. R. Eddy, L. Stein. The Distributed Annotation System. *BMC Bioinformatics* **2**, 7 (2001).

T. A. Down, and T. J. P. Hubbard. Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA. *Genome Research* **12**, 458–461 (2002).

T. A. Down and M. R. Pocock. Building a Distributed Annotation System. In *Proceedings of NETTAB*, 2001.

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, A. M. Bridgeman, D. Buck, J. Burgess, W. D. Burrill, K. P. O'Brien, *et al.*. The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).

R. Durbin, S. Eddy, A. Krogh, G. Mitchison. *Biological Sequence Analysis.* Cambridge University Press, 1998.

S. R. Eddy. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics* **2**, 919–929 (2001).

L. Elnetski, R. C. Hardison, J. Li, S. Yang, D. Kolbe, P. Eswara, M. J. O'Connor, S. Schwartz, W. Miller, and F. Chiaromonte. Distinguishing Regulatory DNA From Neutral Sites. *Genome Research* **13**, 64–72 (2003).

The ENCODE project. Encyclopedia of DNA elements (ENCODE). URL *http://www.genome.gov/Pages/Research/ENCODE/*.

A. J. Enright, S. Van Dongen, C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**, 1575–1584 (2002).

R. C. Périer, V. Praz, T. Junier, C. Bonnard, P. Bucher. The Eukaryotic Promoter Database. URL *http://www.epd.isb-sib.ch/*.

J.W. Fickett and A.C. Hatzigeorgiou. Eukaryotic Promoter Recognition. *Genome Research* **7**, 861–878 (1997).

C. Fields, M. D. Adams, O. White and J. C. Venter. How many genes in the human genome?. *Nature Genetics* **7**, 345–346 (1994).

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**, 906–914 (2000).

D. J. Galas and A. Schmitz. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research* **5**, 3157–3170 (1978).

M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *Journal of Molecular Biology* **196**, 261–282 (1987).

M. R. Green and R. G. Roeder. Definition of a Novel Promoter For the Major Adenovirus-associated Virus mRNA. *Cell* **22**, 231–242 (1980).

H. Grosshans and F. J. Slack. Micro-RNAs: small is plentiful. *Journal of Cell Biology* **156**, 17–21 (2002).

W. N. Grundy, T. L. Bailey, C. P. Elkan, M. E. Baker. Meta-MEME: Motif-based Hidden Markov Models of Biological Sequences. *Computer Applications in the Biosciences* **13**, 397–406 (1997).

G. Slater. Exonerate. URL *http://www.ebi.ac.uk/~guy/exonerate/*.

S. C. Harvey, C. Wang, S. Teltchea, and R. Lavery. Motifs in Nucleic Acids: Molecular Mechanics Restraints for Base Pairing and Base Stacking. *Journal of Compuational Chemistry* **24**, 1–9 (2003).

M. Hayashi. A DNA-RNA complex as an intermediate of in vitro genetic transcription. *Proceedings of the National Academy of Sciences* **54**, 1736–1743 (1965).

R. Holliday. The inheritance of epigenetic defects. *Science* **238**, 163–170 (1987).

W. Hoschek *et al.*. The Colt Distribution – Open Source Libraries for High Performance Scientific and Technical Computing in Java. URL *http://tilde-hoschek.home.cern.ch/~hoschek/colt/index.htm.*

The Ensembl Project. The Ensembl Trace Repository. URL *http://trace.ensembl.org/*.

T. J. P. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, V. Curwen, T. A. 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 M. Clamp. The Ensembl genome database project. *Nucleic Acids Research* **30**, 38–41 (2002).

The Genome International Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, 1999.

J. T. Kadonaga. Eukaryotic Transcription: An Interlaced Network of Transcription Factors and Chromatin-Modifying Machines. *Cell* **92**, 307–313 (1998).

J. Kawai, A. Singagawa, K.Shibata, M. Yoshino, M. Itoh, Y. Ishii, T. Arakawa, A. Hara, Y. Fukunishi, H. Konno *et al.*. Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).

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).

S. Knudsun. Promoter2.0: for the recognition of polII promoter sequences. *Bioinformatics* **15**, 356–361 (1999).

R. Kodandapani, F. Pio, C. Z. Ni, G. Piccalli, M. Klemsz, S. McKercher, R. A. Maki, and K. R. Ely. A new pattern for helix-turn-helix recogniztion revealed by the PU.1 ETS-domain-DNA complex. *Nature* **380**, 456–460 (1996).

R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Machine Learning: Proceedings of the Nineteenth Internaltional Conference*, 2002.

I. Korf, P. Flicek, D. Duan, and M. R. Brent. Integrating genomic homology into gene structure prediction. *Bioinformatics*, S140-148 (2001).

R. D. Kornberg. Eukaryotic Transcription Control. *Trends in Genetics* **15**, M46–49 (1999).

B. Krishnapuram, L. Carin, and A. Hartemink. Joint classifier and feature optimization for cancer diagnosis using gene expression data. In *Proceedings of RECOMB*, 2003.

A. N. Ladd and T. A. Cooper. Finding signals that regulate alternate splicing in the post-genomic era. *Genome Biology* **3**, reviews0008 (2002).

F. Larsen, G. Gundersen, R. Lopez, H. Prydz. CpG islands are gene markers in the human genome. *Genomics* **13**, 1095–1107 (1992).

N. N. Laurinn, S. P. Wang, and G. A. Mitchell. The hormone-sensitive lipase gene is transcribed from at least five alternative first exons in mouse adipose tissue. *Mammalian Genome* **11**, 972–978 (2000).

C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics, 1995.

R. Leplae and T. J. P. Hubbard. MaxBench: evaluation of sequence and structure comparison methods. *Bioinformatics* **18**, 494–495 (2002).

B. Lewin. *Genes VII*. Oxford University Press, 2000.

K. Luger, A. W. Mader, R. K. Richmond, D. F Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8Å resolution.. *Nature* **389**, 251–260 (1997).

D. J. C. Mackay. Bayesian non-linear modelling for the prediction competition. *ASHRAE Transactions* **100**, 1053–1062 (1994).

D. J. C. MacKay. Ensemble Learning and Evidence Maximization, 1995.

D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. To be published by

Cambridge University Press, 2003. URL *http://www.inference.phy.cam.ac.uk/mackay/itprnn/*.

K. Maruyama and S. Sugano. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**, 171–174 (1994).

C. Mathé, M. Sagot, T. Shiex, and P. Rouzé. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research* **30**, 4103–4117 (2002).

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, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* **31**, 374–378 (2003).

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1983.

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

The Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

J. W. Miskin. *Ensemble Learning for Independent Component Analysis*. Ph.D. thesis, University of Cambridge, 2000.

R. Mott. EST_GENOME: a program to align spliced DNA sequence to unspliced genomic DNA. *Computer Applications in the Biosciences* **13**, 477-478 (1997).

J. Moult, K. Fidelis, A. Zemla, T. Hubbard. Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins* **Suppl 5**, 2–7 (2001).

V. E. Myer and R. A. Young. RNA Polymerase II Holoenzymes and Subcomplexes. *Journal of Biological Chemistry* **273**, 27757–27760 (1998).

I. T. Nabney. Efficient training of RBF networks for classification. In *Proceedings of ICANN99*, pages 210–215, 1999.

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).

Z. Ning, A. J. Cox, and J. C. Mullikin. SSAHA: A fast search method for large DNA databases. *Genome Research* **11**, 1725–1729 (2001).

E. Nishida and Y. Gotoh. The MAP kinase cascade is essential for diverse signal transduction pathways. *Trends in Biochemical Sciences* **18**, 128–131 (1993).

Open-bio database access standards. URL *http://obda.open-bio.org/*.

A. B. Pardee, F. Jacob, J. Monod. The genetic control and cytoplasmic expression of “inducibility” in the synthesis of B-galactosidase by *E. coli*. *Journal of Molecular Biology* **1**, 165–178 (1959).

G. Pesole, F. Mignone, C. Gissi, G. Grillo, F. Licciulli, and S. Liuni. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* **276**, 73–81 (2001).

Y. Pilpel, P. Sudarsanam, G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* **29**, 153–159 (2001).

M. R. Pocock. *Computational Analysis of Genomes*. Ph.D. thesis, Wellcome Trust Sanger Instititute, 2001.

R. C. Périer, V. Praz, T. Junier, C. Bonnard, P. Bucher. The Eukaryotic Promoter Database (EPD). *Nucleic Acids Research* **28**, 307–309 (2000).

N. J. Proudfoot, A. Furger, and M. J. Dye. Integrating mRNA processing with transcription. *Cell* **108**, 501–512 (2002).

A. Razin. CpG methylation, chromatin structure and gene silencing – a three-way connection. *EMBO Journal* **17**, 4905–4908 (1998).

B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannet, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, R. A. Young. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–9 (2000).

P. Rice, I. Longden, and A. Bleasby. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276–277 (2000).

P. W. Rigby. Three in one and one in three: it all depends on TBP. *Cell* **72**, 7–10 (1993).

Tissue engineering: implications in the treatment of organ and tissue defects. Tissue engineering: implications in the treatment of organ and tissue defects. *Biogerontology* **2**, 118–125 (2001).

E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**, 8 (2001).

H. C. Roest, O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau, C. Fischer, C. Fizames, P. Wincker, P. Brottier, F. Quetier, W. Saurin, J. Weissenbach. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nature Genetics* **25**, 235–238 (2000).

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).

M. Scherf, A. Kingenhoff, K. Frech, K. Quandt, R. Schneider, K. Grote, M. Frisch, V. Gailus-Durner, A. Seidel, R. Brack-Werner and T. Werner. First pass annotation of promoters on human chromosome 22. *Genome Research* **11**, 333-340 (2001).

B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in Kernel Methods – Support Vector Learning*. MIT Press, 1999.

G. Schramm, I. Bruchhaus, and T. Roeder. A simple and reliable 5'-RACE approach. *Nucleic Acids Research* **28**, e96 (2000).

S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler and W. Miller. Human-Mouse Alignments with BLASTZ. *Genome Research* **13**, 103–107 (2003).

I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakola, R. A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697–708 (2001).

A. F. A. Smit and P. Green. RepeatMasker. URL *http://ftp.genome.washington.edu/RM/RepeatMasker.html*.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**, 195–197 (1981).

E. L. L. Sonnhammer and R. Durbin. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**, GC1–10 (1995).

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, E. Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research* **12**, 1611-1618 (2002).

Y. Suzuki, R. Yamashita, K. Nakai, and S. Sugano. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Research* **30**, 328–331 (2002).

M. Pocock, T. Down, M. Schreiber, K. James, D. Huen, *et al.*. BioJava: Open Source components for Biological Computation.. URL *http://www.biojava.org/*.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

M. E. Tipping. The Relevance Vector Machine. *Advances in Neural Information Processing Systems* **12**, 652–658 (2000).

T. A. Tatusova and T. L. Madden. Blast 2 seqeunces – a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters* **174**, 247–250 (1999).

A. Wright, B. Charlesworth, I. Rudan, A. Carothers, and H. Campbell. A polygenic basis for late-onset disease. *Trends in Genetics* **19**, 97–106 (2003).