Bibliography

[AA01] J. O. Andersson and S. G. Andersson. Pseudogenes, junk, and the dynamics of rickettsia genomes. *Molecular Biology and Evolution*, 18(5):829–39., 2001.

[AB97] L. Arvestad and W.J. Bruno. Estimation of reversible substitution matrices from multiple pairs of sequences. *Journal of Molecular Evolution*, 45(6):696–703, 1997.

[ABW+04] R. Appweiler, A. Bairoch, C.H. Wu, Barker W.C., Boeckmann B., S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S. Yeh. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 32:D115–D119, 2004.

[ACSR03] P. Aloy, H. Ceulemans, A. Stark, and R. B. Russell. The relationship between sequence and interaction divergence in proteins. *Journal of Molecular Biology*, 332(5):989–98, 2003.

[AGT01a] G. Apic, J. Gough, and S. A. Teichmann. An insight into domain combinations. *Bioinformatics*, 17(1):S83–S89, 2001.

[AGT01b] G. Apic, J. Gough, and S.A. Teichmann. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, 310(2):311–25, 2001.

[AH96] J. Adachi and M. Hasegawa. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution*, 42(4):459–68, 1996.

[AHB+04] A. Andreeva, D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32 Database issue:D226–9., 2004.

[AHT03] G. Apic, W. Huber, and S. A. Teichmann. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *Journal of Structural and Functional Genomics*, 4(2-3):67–78, 2003.

[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.

[ASHS04] R. Aroul-Selvam, T. Hubbard, and R. Sasidharan. Domain insertions in protein structures. *Journal of Molecular Biology*, 338(4):633–41, 2004.

[AWMH00] J. Adachi, P.J. Waddell, W. Martin, and M. Hasegawa. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution*, 50(4):348–58, 2000.

[BA03] E. S. Balakirev and F. J. Ayala. Pseudogenes: are they "junk" or functional? *Annual Review of Genetics*, 37:123–51., 2003.

[BAB+04] E. Birney, D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyras, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. Hotz, V. Iyer, A. Kahari, K. Jekosch, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, C. Woodwark, M. Clamp, and T. Hubbard. Ensembl 2004. *Nucleic Acids Research*, 32(1):D468–70., 2004.

[BBI+99] N. Bobola, P. Briata, C. Ilengo, N. Rosatto, C. Craft, G. Corte, and R. Ravazzolo. Otx2 homeodomain protein binds a element necessary for interphotore-

ceptor retinoid binding protein gene expression. *Mechanisms of Development*, 82(1-2):165–169., 1999.

[BBT03] G. J. Bartlett, N. Borkakoti, and J. M. Thornton. Catalysing new reactions during evolution: economy of residues and mechanism. *Journal of Molecular Biology*, 331(4):829–60, 2003.

[BC02] M. Bashton and C. Chothia. The geometry of domain combination in proteins. *Journal of Molecular Biology*, 315(4):927–39, 2002.

[BCD+04] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The pfam protein families database. *Nucleic Acids Research*, 32(1):D138–41., 2004.

[BCH98] S. E. Brenner, C. Chothia, and T. J. Hubbard. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences of the USA*, 95(11):6073–8., 1998.

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

[BDF+04] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P.A. Futreal, M.R. Stratton, and R. Wooster. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer*, 91(2):355–8, 2004.

[BDWC04] H. P. Bogerd, B. P. Doehle, H. L. Wiegand, and B. R. Cullen. A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor. *Proceedings of the National Academy of Sciences of the USA*, 101(11):3770–4, 2004.

[BEW03] N. Bierne and A. Eyre-Walker. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates. implications for the

correlation between the synonymous substitution rate and codon usage bias. *Genetics*, 165(3):1587–97., 2003.

[BNP+02] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. Rubin, and M. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proceedings of the National Academy of Sciences of the USA*, 99(2):757–762, 2002.

[Bre00] S. E. Brenner. Target selection for structural genomics. *Nature Structural Biology*, 7 Suppl:967–9., 2000.

[Bru96] W.J. Bruno. Modeling residue usage in aligned protein sequences via maximum likelihood. *Molecular Biology and Evolution*, 13(10):1368–74, 1996.

[CAJ+94] Y. Cao, J. Adachi, A. Janke, S. Paabo, and M. Hasegawa. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *Journal of Molecular Evolution*, 39(5):519–27., 1994.

[CBD03] L. Coin, A. Bateman, and R. Durbin. Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proceedings of the National Academy of Sciences of the USA*, 100(8):4516–20, 2003.

[CBD04] L. Coin, A. Bateman, and R. Durbin. Enhanced protein domain discovery using taxonomy. *BMC Bioinformatics*, 5(1):56, 2004.

[CD04] L. Coin and R. Durbin. Improved techniques for the identification of pseudogenes. *Bioinformatics*, 20 Suppl 1:I94–I100, 2004.

[Cha93] E. Charniak. *Statistical Language Learning*. The MIT Press, Cambridge, Massachusetts, 1993.

[CHN03] S. G. Conticello, R. S. Harris, and M. S. Neuberger. The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. *Current Biology*, 13(22):2009–13, 2003.

[Cho59] N. Chomsky. On certain formal properties of grammars. *Information and Control*, 2:137–167, 1959.

[CKM+03] D. Chivian, D. E. Kim, L. Malmstrm, P. Bradley, T. Robertson, P. Murphy, C. E. M. Strauss, R. Bonneau, C. A. Rohl, and D. Baker. Automated prediction of CASP-5 structures using the Robetta server. *Proteins*, 53 Suppl 6:524–33, 2003.

[Cla03a] A. G. Clark et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, 302(5652):1960–3, 2003.

[CLRS01] T.H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2001.

[CTPMN04] S. G Conticello, C. J. F. Thomas, S. Petersen-Mahrt, and M. S. Neuberger. Evolution of the AID/APOBEC Family of Polynucleotide (Deoxy)Cytidine Deaminases. *Molecular Biology and Evolution*, 2004.

[CWC+02] J. M. Chandonia, N. S. Walker, Lo L. Conte, P. Koehl, M. Levitt, and S. E. Brenner. Astral compendium enhancements. *Nucleic Acids Research*, 30(1):260–3., 2002.

[CWN+97] S. Chen, Q. L. Wang, Z. Nie, H. Sun, G. Lennon, N. G. Copeland, D. J. Gilbert, N. A. Jenkins, and D. J. Zack. Crx, a novel otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron*, 19(5):1017–1030., 1997.

[CWX+02] S. Chen, Q. L. Wang, S. Xu, I. Liu, L. Y. Li, Y. Wang, and D. J. Zack. Functional analysis of cone-rod homeobox) mutations associated with retinal dystrophy. *Human Molecular Genetics*, 11(8):873–884., 2002.

[DCB98] A. K. Das, P. W. Cohen, and D. Barford. The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for-mediated protein-protein interactions. *EMBO Journal*, 17(5):1192–9., 1998.

[Deh] P. Dehal. Phylogenetically inferred groups. http://phigs.jgi-psf.org.

[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.

[Den95] D.C. Dennett. *Darwin's dangerous idea*. The Penguin Press, 1995.

[DLR77] 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.

[DMBB] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou. Probcons: Probabilistic consistency-based multiple alignment of amino acid sequences. *Genome Research*. In press.

[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.

[DSW01] V. G. Durner, M. Scherf, and T. Werner. Experimental data of a single promoter can be used for in silico detection of genes with related regulation in the absence of sequence similarity. *Mammalian Genome*, 12:67–72, 2001.

[Edd98] S. R. Eddy. Profile-hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[Edd03] S. Eddy. Hmmer users guide, 2003.

[Edg04] R. C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1):113, 2004.

[EHP00] M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. Wiley, 2000.

[EKO03] A. J. Enright, V. Kunin, and C. A. Ouzounis. Protein families and TRIBES in genome sequence space. *Nucleic Acids Research*, 31(15):4632–8, 2003.

[FC96] J. Felsenstein and G. A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13:93–104, 1996.

[FCM+04] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–83, 2004.

[Fel81] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[FMC97] T. Furukawa, E. M. Morrow, and C. L. Cepko. Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell*, 91(4):531–541., 1997.

[Ger98] M. Gerstein. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding and Design*, 3(6):497–512, 1998.

[GG03] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, 2003.

[GMB01] E.A. Gaucher, M.M. Miyamoto, and S.A. Benner. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. *Proceedings of the National Academy of Sciences of the USA*, 98(2):548–52, 2001.

[GS93] W. Gish and D.J. States. Identification of protein coding regions by database similarity search. *Nature Genetics*, 3(3):266–72, 1993.

[GSC94] M. Gerstein, E. L. L. Sonnhammer, and C. Chothia. Volume changes in protein evolution. *Journal of Molecular Biology*, 236:1067–1078, 1994.

[Gu01] X. Gu. Maximum-likelihood approach for gene family evolution under functional divergence. *Molecular Biology and Evolution*, 18(4):453–64, 2001.

[GW02] N. Goldman and S. Whelan. novel use of equilibrium frequencies in models of sequence evolution. *Molecular Biology and Evolution*, 19(11):1821–31., 2002.

[GY94] N. Goldman and Z. Yang. codon-based model of nucleotide substitution for protein-coding sequences. *Molecular Biology and Evolution*, 11(5):725–36., 1994.

[HB98] A.L. Halpern and W.J. Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15(7):910–7, 1998.

[HB01] I. Holmes and W.J. Bruno. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17(9):803–20, 2001.

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

[HG01] H. Hegyi and M. Gerstein. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Research*, 11(10):1632–40, 2001.

[HG02] P. M. Harrison and M. Gerstein. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *Journal of Molecular Biology*, 318(5):1155–74., 2002.

[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.

[HH96] J. G. Henikoff and S. Henikoff. Using substitution probabilities to improve position-specific scoring matrices. *Computer Applications in the Biosciences*, 12:135–143, 1996.

[HHB+02] P. M. Harrison, H. Hegyi, S. Balasubramanian, N. M. Luscombe, P. Bertone, N. Echols, T. Johnson, and M. Gerstein. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Research*, 12(2):272–80., 2002.

[HHP99] J.G. Henikoff, S. Henikoff, and S. Pietrokovski. New features of the Blocks Database servers. *Nucleic Acids Research*, 27(1):226–8, 1999.

[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.

[HKY85] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial. *Journal of Molecular Evolution*, 22(2):160–74., 1985.

[HMBC97] T. J. P. Hubbard, A. Murzin, S. Brenner, and C. Chothia. SCOP: a structural classification of proteins database. *Nucleic Acids Research*, 25:236–239, 1997.

[HMZ+03] P. M. Harrison, D. Milburn, Z. Zhang, P. Bertone, and M. Gerstein. Identification of pseudogenes in the drosophila melanogaster genome. *Nucleic Acids Research*, 31(3):1033–7., 2003.

[HN88] A.L. Hughes and M. Nei. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186):167–70, 1988.

[Hol03] I. Holmes. Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics*, 19 Suppl 1:i147–57, 2003.

[HR02] I. Holmes and G.M. Rubin. An expectation maximization algorithm for training hidden substitution models. *Journal of Molecular Biology*, 317(5):753–64, 2002.

[HSGMS02] L. Q. Ha, E. I. Sicilia-Garcia, J. Ming, and F. J. Smith. Extension of zipf's law to words and phrases. In R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls, editors, *Coling proceedings*, pages 315–320, Taipei, Taiwan, 2002.

[HW01] D. Husmeier and F. Wright. Detection of recombination in DNA multiple alignments with hidden Markov models. *Journal of Computational Biology*, 8(4):401–27, 2001.

[HYC+03] S. Hirotsune, N. Yoshida, A. Chen, L. Garrett, F. Sugiyama, S. Takahashi, K. Yagami, A. Wynshaw-Boris, and A. Yoshiki. An expressed pseudogene regulates the messenger stability of its homologous coding gene. *Nature*, 423(6935):91–6., 2003.

[Jay03] E. Jaynes. *Probability Theory. The logic of science.* Cambridge University Press, Cambridge, U.K., 2003.

[JCB+02] A. Jarmuz, A. Chester, J. Bayliss, J. Gisbourne, I. Dunham, J. Scott, and N. Navaratnam. An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics*, 79(3):285–96, 2002.

[JDH00] T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7:95–114, 2000.

[Jel97] F. Jelinek. *Statistical Methods For Speech Recognition*. The MIT Press, Cambridge, Massachusetts, 1997.

[JKA+] I. K. Jordan, F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, A. S. Kondrashov, S. Sunyaev, and E. V. Koonin. A universal trend of amino acid gain and loss in protein evolution: the modern echo of code origin. To be published.

[JTT92] 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(3):275–82., 1992.

[JY99] A. Joshi and Y.Schabes. Tree-adjoining grammars. Technical report, University of Pennsylvania, 1999. http://www.cis.upenn.edu/ joshi/.

[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, 1994.

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

[KM95] A. Krogh and G. Mitchison. Maximum entropy weighting of aligned sequences of proteins or. In 1995/01/01, editor, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 215–21., Laboratory of Molecular Biology, Cambridge, England., 1995. AAAI Press.

[KM01] B. Knudsen and M.M. Miyamoto. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proceedings of the National Academy of Sciences of the USA*, 98(25):14512–7, 2001.

[KZNL02] H. Kaessmann, S. Zllner, A. Nekrutenko, and W-H. Li. Signatures of domain shuffling in the human genome. *Genome Research*, 12(11):1642–50, 2002.

[Lay94] D. C. Lay. *Linear algebra and its applications*. Addison Wesley, 1994.

[LD43] S.E. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491–511, 1943.

[LEC+04] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–76, 2004.

[LG99] P. Lio and N. Goldman. Using protein structural information in evolutionary inference: transmembrane proteins. *Molecular Biology and Evolution*, 16(12):1696–710., 1999.

[LGTJ98] P. Lio, N. Goldman, J. L. Thorne, and D. T. Jones. : PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics*, 14(8):726–33., 1998.

[LJN+96] L. Lavaissiere, S. Jia, M. Nishiyama, de la S. Monte, A. M. Stern, J. R. Wands, and P. A. Friedman. Overexpression of human aspartyl(asparaginyl)beta-hydroxylase in hepatocellular carcinoma and cholangiocarcinoma. *Journal of Clinical Investigation*, 98(6):1313–23., 1996.

[LOV95] Y.H. Lee, T. Ota, and V.D. Vacquier. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Molecular Biology and Evolution*, 12(2):231–8, 1995.

[LP04] N. Lartillot and H. Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–109, 2004.

[LPA+93] H. Lutcke, S. Prehn, A. J. Ashford, M. Remus, R. Frank, and B. Dobberstein. Assembly of the 68- and 72-kd proteins of signal recognition particle with 7s. *Journal of Cell Biology*, 121(5):977–985., 1993.

[LPR01] A.N. Lupas, C.P. Ponting, and R.B. Russell. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of Structural Biology*, 134(2-3):191–203, 2001.

[LPSS84] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20(1):86–93, 1984.

[LQZ+02] N. M. Luscombe, J. Qian, Z. Zhang, T. Johnson, and M. Gerstein. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biology*, 3(8):RESEARCH0040, 2002.

[MCP+04] A. M. Moses, D. Y. Chiang, D. A. Pollard, V. N. Iyer, and M. B. Eisen. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biology*, 5(12):R98, 2004.

[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.

[MG94] S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–24., 1994.

[MG02] M. Madera and J. Gough. Comparison of profile hidden markov model procedures for remote homology detection. *Nucleic Acids Research*, 30(19):4321–8., 2002.

[Mig00] A.J. Mighell et al. Vertebrate pseudogenes. *FEBS Letters*, 468:109–14., 2000.

[MLH04] I. Miklós, G.A. Lunter, and I. Holmes. A "Long Indel" model for evolutionary sequence alignment. *Molecular Biology and Evolution*, 21(3):529–40, 2004.

[Mou02] A. Mounsey et al. Evidence suggesting that a fifth of annotated caenorhabditis elegans genes may be pseudogenes. *Genome Research*, 12:770–75., 2002.

[MPJ03] J.D McAuliffe, L. Pachter, and M.I. Jordan. Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. Technical report, University of California Berkley, 2003.

[MSBP02] R. Mott, J. Schultz, P. Bork, and C.P. Ponting. Predicting cellular localization using a domain projection method. *Genome Research*, 8:1168–1174, 2002.

[Mun03] A. J. Mungall et al. The sequence and analysis of human chromosome 6. *Nature*, 425(6960):805–11., 2003.

[NG86] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5):418–26., 1986.

[Nor97] J.R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[NY98] R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–36, 1998.

[PFL+01] P. Pavlidis, T. S. Furey, M. Liberto, D. Haussler, and W. N. Grundy. Promoter region based classification of genes. *Proceedings of the Pacific Symposium on Biocomputing*, pages 151–163, 2001.

[Pie96] S. Pietrokovski. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, 24(19):3836–45, 1996.

[PKB+98] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology*, 284(4):1201–10., 1998.

[PMHN02] S. K. Petersen-Mahrt, R. S. Harris, and M. S. Neuberger. AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature*, 418(6893):99–103, 2002.

[PP02] V. E. Prince and F. B. Pickett. Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews Genetics*, 3(11):827–37, 2002.

[PTTF92] W. H. Press, S. A. Teukolsky, W. Vetterling T., and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK, 1992.

[QG03] B. Qian and R. A. Goldstein. Detecting distant homologs using phylogenetic tree-based HMMs. *Proteins*, 52(3):446–53., 2003.

[QLG01] J. Qian, N.M. Luscombe, and M. Gerstein. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *Journal of Molecular Biology*, 313(4):673–81, 2001.

[RBD01] C. Rivolta, E. L. Berson, and T. P. Dryja. Dominant leber congenital amaurosis, cone-rod degeneration, and retinitis pigmentosa caused by mutant versions of the transcription factor. *Human Mutation*, 18(6):488–498., 2001.

[RCZ01] R. Rosenfeld, S.F. Chen, and X. Zhu. Whole-sentence exponential language models: a vehicle for linguistic statistical integration. *Computer Speech and Language*, 15(1), 2001.

[RJ93] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[RJK+03] D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*, 20(10):1692–704, 2003.

[Ros02] B. Rost. Did evolution leap to create the protein universe? *Current Opinion in Structural Biology*, 12(3):409–16, 2002.

[SA90] P. R. Sibbald and P. Argos. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *Journal of Molecular Biology*, 216:813–818, 1990.

[SBF+00] A. Scaloni, C. Bottiglieri, L. Ferrara, M. Corona, G. B. Gurrola, C. Batista, E. Wanke, and L. D. Possani. Disulfide bridges of ergtoxin, a member of a new sub-family of peptide blockers of the ether-a-go-go-related K+ channel. *FEBS Letters*, 479(3):156–7., 2000.

[SBG03] R. I. Sadreyev, D. Baker, and N. V. Grishin. Profile-profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Science*, 12(10):2262–72, 2003.

[SCB+00] M. Salzet, V. Chopin, J. Baert, I. Matias, and J. Malecha. Theromin, a novel leech thrombin inhibitor. *Journal of Biological Chemistry*, 275(40):30774–80., 2000.

[SCW+97] P. K. Swain, S. Chen, Q. L. Wang, L. M. Affatigato, C. L. Coats, K. D. Brady, G. A. Fishman, S. G. Jacobson, A. Swaroop, E. Stone, P. A. Sieving, and D. J. Zack. Mutations in the cone-rod homeobox gene are associated with the cone-rod dystrophy photoreceptor degeneration. *Neuron*, 19(6):1329–1336., 1997.

[SEM04] S. L. Sawyer, M. Emerman, and H. S. Malik. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biology*, 2(9):E275, 2004.

[SG99] Y. Suzuki and T. Gojobori. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, 16(10):1315–28., 1999.

[SGCM02] A. M. Sheehy, N. C. Gaddis, J. D. Choi, and M. H. Malim. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418(6898):646–50, 2002.

[SH04] A. Siepel and D. Haussler. Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology*, 11(2-3):413–28, 2004.

[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.

[SL03] J. Sding and A. N. Lupas. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*, 25(9):837–46, 2003.

[SPD+99] 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., 1999.

[SV95] W. J. Swanson and V. D. Vacquier. Extraordinary divergence and positive darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. *Proceedings of the National Academy of Sciences of the USA*, 92(11):4957–61., 1995.

[Sd04] J. Sding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 2004.

[TBD93] B. Teng, C.F. Burant, and N.O. Davidson. Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science*, 260(5115):1816–9, 1993.

[TGJ96] J. L. Thorne, N. Goldman, and D. T. Jones. Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13:666–673, 1996.

[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.

[TN93] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–26, 1993.

[TOT01] A.E. Todd, C.A. Orengo, and J.M. Thornton. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, 307(4):1113–43, 2001.

[TPC98] S.A. Teichmann, J. Park, and C. Chothia. Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proceedings of the National Academy of Sciences of the USA*, 95(25):14658–63, 1998.

[TSZB03] D. Torrents, M. Suyama, E. Zdobnov, and P. Bork. A genome-wide survey of human pseudogenes. *Genome Research*, 13(12):2559–67., 2003.

[VBB+04] C. Vogel, C. Berzuini, M. Bashton, J. Gough, and S. A. Teichmann. Supra-domains: evolutionary units larger than single protein domains. *Journal of Molecular Biology*, 336(3):809–23, 2004.

[VBK+04] C. Vogel, M. Bashton, N. D. Kerrison, C. Chothia, and S. A. Teichmann. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*, 14(2):208–16, 2004.

[Ven04] J. C. Venter et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004.

[WG01] S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18(5):691–9., 2001.

[WS04] M. Wistrand and E. L. L. Sonnhammer. Improving profile HMM discrimination by adapting transition probabilities. *Journal of Molecular Biology*, 338(4):847–54, 2004.

[WYGN04] W. S. W. Wong, Z. Yang, N. Goldman, and R. Nielsen. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, 168(2):1041–51, 2004.

[Yan93] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10:1396–1401, 1993.

[Yan95] Z. Yang. A space-time process model for the evolution of dna sequences. *Genetics*, 139:993–1005, 1995.

[Yano04] Y. Yano et al. A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *Journal of Molecular Medicine*, 82(7):414–22, 2004.

[YL02] G. Yona and M. Levitt. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *Journal of Molecular Biology*, 315(5):1257–75, 2002.

[YN98] Z. Yang and R. Nielsen. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, 46(4):409–18., 1998.

[YN00] Z. Yang and R. Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, 17(1):32–43., 2000.

[YN02] Z. Yang and R. Nielsen. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, 19(6):908–17, 2002.

[YNGP00] Z. Yang, R. Nielsen, N. Goldman, and A.M. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–49, 2000.

[YSV00] Z. Yang, W. J. Swanson, and V.D. Vacquier. Maximum-likelihood analysis of molecular adaption in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Molecular Biology and Evolution*, 2000.

[Zha04] J. Zhang. Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular Biology and Evolution*, 21(7):1332–9, 2004.

[Zip35] G.K. Zipf. *Psycho-Biology of Languages*. Houghton-Mifflin, 1935.

[ZYP+03] H. Zhang, B. Yang, R. J. Pomerantz, C. Zhang, S. C. Arunachalam, and L. Gao. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. *Nature*, 424(6944):94–8, 2003.