

References

- Adryan, B. and Teichmann, S.A. FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. *Bioinformatics*, 2006. URL <http://www.google.com/search?client=safari&rls=en-us&q=FlyTF:+a+systematic+review+of+site-specific+transcription+factors+in+the+fruit+fly+Drosophila+melanogaster&ie=UTF-8&oe=UTF-8>. 6
- Aerts, S., Thijs, G., Coessens, B., Staes, M., et al. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*, 31(6):1753–1764, 2003. 183
- Affolter, M., Slattery, M., and Mann, R.S. A lexicon for homeodomain-DNA recognition. *Cell*, 133(7):1133–5, 2008. doi:10.1016/j.cell.2008.06.008. URL [http://linkinghub.elsevier.com/retrieve/pii/S0092-8674\(08\)00760-5](http://linkinghub.elsevier.com/retrieve/pii/S0092-8674(08)00760-5). 8
- Albrecht, G., Mösch, H., Hoffmann, B., and Reusser, U. Monitoring the Gcn4 Protein-mediated Response in the Yeast *Saccharomyces cerevisiae*. *Journal of Biological ...*, 1998. URL <http://www.jbc.org/content/273/21/12696.full>. 144
- Altschul, S.F. and Gish, W. Local alignment statistics. *Methods in Enzymology*, 266:460–80, 1996. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=8743700&dopt=abstractplus. 109
- Anthony-Cahill, S.J., Benfield, P.A., Fairman, R., Wasserman, Z.R., et al. Molec-

REFERENCES

- ular characterization of helix-loop-helix peptides. *Science*, 255(5047):979–83, 1992. URL <http://www.sciencemag.org/cgi/reprint/255/5047/979>. 50
- Ao, W., Gaudet, J., Kent, W.J., Muttumu, S., et al. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science*, 305(5691):1743–6, 2004. doi:10.1126/science.1102216. URL <http://www.sciencemag.org/cgi/content/full/305/5691/1743>. 102, 113
- Babu, M., Luscombe, N., Aravind, L., Gerstein, M., et al. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283–291, 2004. 9
- Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., et al. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–3, 2009. doi:10.1126/science.1162327. URL <http://www.sciencemag.org/cgi/content/full/324/5935/1720>. 12, 14, 18, 35, 38
- Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., et al. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell*, 32(6):878–87, 2008. doi:10.1016/j.molcel.2008.11.020. XIV, 12, 99, 114, 115, 165, 167
- Bailey, T.L. and Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994. 20
- Bailey, T.L. and Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, 3:21–29, 1995. 16, 27, 67
- Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(Web Server issue):W369–73, 2006. doi:10.1093/nar/gkl198. URL http://nar.oxfordjournals.org/cgi/content/full/34/suppl_2/W369. 27, 175, 198

REFERENCES

- Baker, J. Stochastic modeling as a means of automatic speech recognition. *oai.dtic.mil*, 1975. URL <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA013808>. 19
- Banerji, J., Rusconi, S., and Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1):299–308, 1981. URL <http://www.cell.com/retrieve/pii/009286748190413X>. 2
- Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. Modeling dependencies in protein-DNA binding sites. *Proceedings of the seventh annual international conference . . .*, 2003. URL <http://portal.acm.org/citation.cfm?id=640079>. 18
- Baum, L. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 1972. URL <http://research.microsoft.com/apps/pubs/default.aspx?id=64753>. 19
- Baum, L. and Petrie, T. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 1966. URL <http://www.jstor.org/stable/2238772>. 19
- Baum, L., Petrie, T., Soules, G., and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical . . .*, 1970. URL <http://www.jstor.org/stable/2239727>. 19
- Baum, L. and Eagon, J. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc*, 73(3):360–363, 1967. 19
- Baum, L. and Sell, G. Growth transformations for functions on manifolds. *Pac. J. Math*, 27:211–227, 1968. 19
- Beck, T. and Hall, M. The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors. *Nature*, 1999. URL <http://www.nature.com/nature/journal/v402/n6762/abs/402689a0.html>. 144

REFERENCES

- Ben-Gal, I., Shani, A., Gohr, A., Grau, J., et al. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics*, 21(11):2657–66, 2005. doi:10.1093/bioinformatics/bti410. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/21/11/2657>. 18
- Benos, P.V., Bulyk, M.L., and Stormo, G.D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*, 30(20):4442–4451, 2002a. 18, 38
- Benos, P.V., Lapedes, A.S., and Stormo, G.D. Probabilistic code for DNA recognition by proteins of the EGR family. *J Mol Biol*, 323(4):701–27, 2002b. 39
- Berger, M., Badis, G., Gehrke, A., Talukder, S., et al. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. *Cell*, 133(7):1266–1276, 2008. doi:10.1016/j.cell.2008.05.024. XI, 35, 36, 91, 92
- Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11):1429–35, 2006. doi:10.1038/nbt1246. URL <http://www.nature.com/nbt/journal/v24/n11/abs/nbt1246.html>. 18, 35, 36, 114, 115, 183
- Bergman, C. and Kreitman, M. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res*, 2001. URL <http://genome.cshlp.org/content/11/8/1335.full>. 5
- Beyer, A., Workman, C., Hollunder, J., Radke, D., et al. Integrated Assessment and Prediction of Transcription Factor Binding. *PLoS Computational Biology*, 2(6):e70, 2006. doi:10.1371/journal.pcbi.0020070.st005. 143
- Bhaskar, H., Hoyle, D.C., and Singh, S. Machine learning in bioinformatics: a brief survey and recommendations for practitioners. *Comput Biol Med*, 36(10):1104–1125, 2006. doi:10.1016/j.compbiomed.2005.09.002. URL <http://dx.doi.org/10.1016/j.compbiomed.2005.09.002>. 29

REFERENCES

- Bi, X. and Broach, J.R. Chromosomal boundaries in *S. cerevisiae*. *Curr Opin Genet Dev*, 11(2):199–204, 2001. 2
- Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., et al. An overview of Ensembl. *Genome Res*, 14(5):925–8, 2004. doi:10.1101/gr.1860604. URL <http://genome.cshlp.org/cgi/content/full/14/5/925>. 31, 109
- Blanchette, M. and Sinha, S. Separating real motifs from their artifacts. *Bioinformatics*, 17 Suppl 1:S30–8, 2001. URL http://bioinformatics.oxfordjournals.org/cgi/reprint/17/suppl_1/S30?view=long&pmid=11472990. 199
- Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., et al. Divergence of transcription factor binding sites across related yeast species. *Science*, 317(5839):815–9, 2007. doi:10.1126/science.1140748. URL <http://www.sciencemag.org/cgi/content/full/317/5839/815>. 5, 112
- Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res*, 8(11):1202–15, 1998. 100
- Breiman, L. Bagging predictors. *Machine Learning*, 1996. URL <http://www.springerlink.com/index/L4780124W2874025.pdf>. 29
- Breiman, L. Random Forests. *Machine Learning*, 2001a. URL <http://www.springerlink.com/index/U0P06167N6173512.pdf>. 31
- Breiman, L. Random Forests. *Machine Learning*, 45(1):5–32, 2001b. URL citeseer.ist.psu.edu/breiman01random.html. 29, 30, 84
- Brenowitz, M., Senear, D.F., Shea, M.A., and Ackers, G.K. Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Meth Enzymol*, 130:132–81, 1986. 33, 119
- Brown, M., Hughey, R., Krogh, A., Mian, I.S., et al. Using Dirichlet mixture priors to derive hidden Markov models for protein families. *Proc Int Conf Intell Syst Mol Biol*, 1:47–55, 1993. 19, 68

REFERENCES

- Bryne, J.C., Valen, E., Tang, M.H.E., Marstrand, T., et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, 36(Database issue):D102–6, 2008. doi: 10.1093/nar/gkm955. URL http://nar.oxfordjournals.org/cgi/content/full/36/suppl_1/D102. 34, 183
- Buchman, A.R., Kimmerly, W.J., Rine, J., and Kornberg, R.D. Two DNA-binding factors recognize specific sequences at silencers, upstream activating sequences, autonomously replicating sequences, and telomeres in *Saccharomyces cerevisiae*. *Mol Cell Biol*, 8(1):210–25, 1988. 2
- Bulyk, M.L., Johnson, P.L.F., and Church, G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–1261, 2002. 18
- Burge, C., Campbell, A.M., and Karlin, S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA*, 89(4):1358–62, 1992. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=1741388&dopt=abstractplus. 28
- Burge, C. and Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268(1):78–94, 1997. doi:10.1006/jmbi.1997.0951. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WK7-45VGF7T-9&_user=10&_coverDate=04rch&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=9a6fd068b3be7c44c65d6d208d7d58dc. 106
- Burset, M. and Guigó, R. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–67, 1996. doi:10.1006/geno.1996.0298. 105, 106
- Bussemaker, H.J., Li, H., and Siggia, E.D. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–71, 2001. doi:10.1038/84792. URL http://www.nature.com/ng/journal/v27/n2/abs/ng0201_167.html;jsessionid=CC8DF0DE6E3E8EB39B739F0A6F599051. 15, 101

REFERENCES

- Bussemaker, H., Li, H., and Siggia, E. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proceedings of the National Academy of Sciences*, 97(18):10096, 2000. 101
- Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–635, 2006. doi:10.1038/ng1789. URL <http://dx.doi.org/10.1038/ng1789>. 5, 8
- Carroll, S., Grenier, J., and Weatherbee, S. From DNA to diversity: The primacy of regulatory evolution. 2000. URL <http://www.google.com/search?client=safari&rls=en-us&q=From+DNA+to+diversity:+The+primacy+of+regulatory+evolution&ie=UTF-8&oe=UTF-8>. 5
- Chan, T., Li, G., Leung, K., and Lee, K. Discovering multiple realistic TFBS motifs based on a generalized model. *BMC bioinformatics*, 10(1):321, 2009. 108
- Chen, C., Liaw, A., and Breiman, L. Using Random Forest to Learn Imbalanced Data. *Unpublished manuscript*, p. 12, 2004. 173
- Chen, X., Xu, H., Yuan, P., Fang, F., et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, 2008. 67
- Chenna, R., Sugawara, H., Koike, T., Lopez, R., et al. Multiple sequence alignment with the Clustal series of programs. *Nucleic acids research*, 31(13):3497–500, 2003. 48
- Chi, Y., Huddleston, M.J., Zhang, X., Young, R.A., et al. Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. *Genes & Development*, 15(9):1078–92, 2001. doi:10.1101/gad.867501. URL <http://genesdev.cshlp.org/content/15/9/1078.long>. 144
- Choi, J.K. and Kim, Y.J. Epigenetic regulation and the variability of gene expression. *Nat Genet*, 40(2):141–7, 2008. doi:10.1038/ng.2007.58.

REFERENCES

- URL <http://www.nature.com/ng/journal/v40/n2/abs/ng.2007.58.html;jsessionid=1E050424C5A8D78C23141658C4087ECD>. 7
- Clements, M., van Someren, E.P., Knijnenburg, T.A., and Reinders, M.J.T. Integration of known transcription factor binding site information and gene expression data to advance from co-expression to co-regulation. *Genomics Proteomics Bioinformatics*, 5(2):86–101, 2007. doi:10.1016/S1672-0229(07)60019-9. CBF1 and PHO4 have closely similar motifs. 141
- Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., et al. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, 301(5629):71–6, 2003. doi:10.1126/science.1084337. URL <http://www.sciencemag.org/cgi/content/full/301/5629/71>. 167
- Comon, P. Independent component analysis, a new concept? *Signal processing*, 1994. URL <http://linkinghub.elsevier.com/retrieve/pii/0165168494900299>. 26
- Conlon, E.M., Liu, X.S., Lieb, J.D., and Liu, J.S. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci USA*, 100(6):3339–44, 2003. doi:10.1073/pnas.0630591100. URL <http://www.pnas.org/content/100/6/3339>. 15
- Consortium, F., Suzuki, H., Forrest, A.R.R., van Nimwegen, E., et al. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet*, 41(5):553–62, 2009. doi:10.1038/ng.375. 5
- Cremer, T. and Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, 2(4):292–301, 2001. doi:10.1038/35066075. URL http://www.nature.com/nrg/journal/v2/n4/full/nrg0401_292a.html. 3
- Cunningham, T.S. and Cooper, T.G. The *Saccharomyces cerevisiae* DAL80 repressor protein binds to multiple copies of GATAA-containing sequences (URSGATA). *Journal of Bacteriology*, 175(18):5851–61, 1993. URL <http://jb.asm.org/cgi/reprint/175/18/5851?view=long&pmid=8376332>. 6

REFERENCES

- Das, M.K. and Dai, H.K. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8 Suppl 7:S21, 2007. doi:10.1186/1471-2105-8-S7-S21. URL <http://www.biomedcentral.com/1471-2105/8/S7/S21>. 13, 186
- Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. 27
- Dietterich, T. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 1998. URL <http://www.mitpressjournals.org/doi/abs/10.1162/089976698300017197>. 29
- Dogruel, M. Motif based computational identification of protein subcellular localisation. pp. 1–255, 2008. 19, 56
- Dogruel, M., Down, T., and Hubbard, T. NestedMICA as an ab initio protein motif discovery tool. *BMC Bioinformatics*, 9(1):19, 2008. doi:10.1186/1471-2105-9-19. URL <http://dx.doi.org/10.1186/1471-2105-9-19>. 23, 25, 53, 56
- Donahue, T.F., Daves, R.S., Lucchini, G., and Fink, G.R. A short nucleotide sequence required for regulation of HIS4 by the general control system of yeast. *Cell*, 32(1):89–98, 1983. 6
- Down, T.A., Bergman, C.M., Su, J., and Hubbard, T.J.P. Large-Scale Discovery of Promoter Motifs in *Drosophila melanogaster*. *PLoS Comput Biol*, 3(1):e7, 2007. doi:10.1371/journal.pcbi.0030007. URL <http://dx.doi.org/10.1371/journal.pcbi.0030007>. XIII, 48, 61, 63, 66, 72, 75, 84, 116, 117, 119, 123, 145, 148, 164, 185, 186
- Down, T.A. and Hubbard, T.J.P. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res*, 33(5):1445–1453, 2005. doi:10.1093/nar/gki282. URL <http://dx.doi.org/10.1093/nar/gki282>. 17, 21, 25, 26, 66, 71, 74, 77, 103, 106, 116, 174, 185, 190
- Dreier, B., Beerli, R.R., Segal, D.J., Flippin, J.D., et al. Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and

REFERENCES

- their use in the construction of artificial transcription factors. *J Biol Chem*, 276(31):29466–78, 2001. doi:10.1074/jbc.M102604200. 90
- Dreier, B., Segal, D.J., and Barbas, C.F. Insights into the molecular recognition of the 5'-GNN-3' family of DNA sequences by zinc finger domains. *J Mol Biol*, 303(4):489–502, 2000. doi:10.1006/jmbi.2000.4133. URL <http://dx.doi.org/10.1006/jmbi.2000.4133>. 90
- Eddy, S.R. Profile hidden Markov models. *Bioinformatics*, 14(9):755–63, 1998. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/14/9/755?view=long&pmid=9918945>. 19
- Egnier, M.R. Rare Events and Conditional Events on Random Strings. 2004. URL <http://citeseer.ist.psu.edu/637853>. 113
- Elemento, O., Slonim, N., and Tavazoie, S. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. *Molecular Cell*, 28(2):337–350, 2007. doi:10.1016/j.molcel.2007.09.027. 181
- Elemento, O. and Tavazoie, S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol*, 6(2):R18, 2005. doi:10.1186/gb-2005-6-2-r18. 101, 111
- Ellingsen, S., Laplante, M.A., König, M., Kikuta, H., et al. Large-scale enhancer detection in the zebrafish genome. *Development*, 132(17):3799–811, 2005. doi:10.1242/dev.01951. 2
- Ernst, J., Plasterer, H.L., Simon, I., and Bar-Joseph, Z. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res*, 20(4):526–36, 2010. doi:10.1101/gr.096305.109. URL <http://genome.cshlp.org/content/20/4/526.long>. 107
- Eskin, E. and Pevzner, P.A. Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 18 Suppl 1:S354–63, 2002. 113
- Ettwiller. *Genome Res*, 2005. 101

REFERENCES

- Fauteux, F., Blanchette, M., and Strömviik, M.V. Seeder: discriminative seeding DNA motif discovery. *Bioinformatics*, 24(20):2303–7, 2008. doi:10.1093/bioinformatics/btn444. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/24/20/2303>. 108
- Favorov, A.V., Gelfand, M.S., Gerasimova, A.V., Ravcheev, D.A., et al. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, 21(10):2240–5, 2005. doi:10.1093/bioinformatics/bti336. 113
- Fejes, A.P., Robertson, G., Bilenky, M., Varhol, R., et al. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, 2008. doi:10.1093/bioinformatics/btn305. 185, 189
- Finn, R.D., Mistry, J., Tate, J., Coghill, P., et al. The Pfam protein families database. *Nucleic Acids Res*, 38(Database):D211–D222, 2010. doi:10.1093/nar/gkp985. 6, 19
- FitzGerald, P.C., Sturgill, D., Shyakhtenko, A., Oliver, B., et al. Comparative genomics of *Drosophila* and human core promoters. *Genome Biol*, 7(7):R53, 2006. doi:10.1186/gb-2006-7-7-r53. URL <http://dx.doi.org/10.1186/gb-2006-7-7-r53>. 27
- Flicek, P., Aken, B.L., Beal, K., Ballester, B., et al. Ensembl 2008. *Nucleic Acids Res*, 36(Database issue):D707–14, 2008. doi:10.1093/nar/gkm988. URL http://nar.oxfordjournals.org/cgi/content/full/36/suppl_1/D707. 71
- Foat, B.C., Houshmandi, S.S., Olivas, W.M., and Bussemaker, H.J. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc Natl Acad Sci USA*, 102(49):17675–80, 2005. doi:10.1073/pnas.0503803102. URL <http://www.pnas.org/content/102/49/17675.long>. 15
- Freund, Y. and Schapire, R. Experiments with a new boosting algorithm. *MACHINE LEARNING-INTERNATIONAL ...*, 1996. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.4143&rep=rep1&type=pdf>. 30

REFERENCES

- Fried, M. and Crothers, D. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res*, 1981a. URL <http://nar.oxfordjournals.org/cgi/content/abstract/9/23/6505>. 119
- Fried, M. and Crothers, D.M. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res*, 9(23):6505–25, 1981b. 33
- Frith, M.C., Valen, E., Krogh, A., Hayashizaki, Y., et al. A code for transcription initiation in mammalian genomes. *Genome Res*, 18(1):1–12, 2008. doi:10.1101/gr.6831208. URL <http://dx.doi.org/10.1101/gr.6831208>. 3
- Fuda, N.J., Ardehali, M.B., and Lis, J.T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261):186–192, 2009. doi:10.1038/nature08449. 1, 4
- Fulton, D.L., Sundararajan, S., Badis, G., Hughes, T.R., et al. TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol*, 10(3):R29, 2009. doi:10.1186/gb-2009-10-3-r29. 6
- Galant, R. and Carroll, S. Evolution of transcriptional repression domain in an insect Hox protein. *Nature*, 2002. URL <http://www.google.com/search?client=safari&rls=en-us&q=Evolution+of+transcriptional+repression+domain+in+an+insect+Hox+protein.&ie=UTF-8&oe=UTF-8>. 5
- Garcia, F., Lopez, F.J., Cano, C., and Blanco, A. FISim: a new similarity measure between transcription factor binding sites based on the fuzzy integral. *BMC bioinformatics*, 10:224, 2009. doi:10.1186/1471-2105-10-224. URL <http://www.biomedcentral.com/1471-2105/10/224>. 41
- Garner, M.M. and Revzin, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res*, 9(13):3047–60, 1981. 33

REFERENCES

- Gasser, S.M. Visualizing chromatin dynamics in interphase nuclei. *Science*, 296(5572):1412–6, 2002. doi:10.1126/science.1067703. URL <http://www.sciencemag.org/cgi/content/full/296/5572/1412>. 3
- Gelfand, M. and Mirny, L. Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res*, 2002. URL <http://www.ingentaconnect.com/content/oup/nar/2002/00000030/00000007/art01704>. 14
- Georges, A.B., Benayoun, B.A., Caburet, S., and Veitia, R.A. Generic binding sites, generic DNA-binding domains: where does specific promoter recognition come from? *The FASEB Journal*, 24(2):346–356, 2010. doi:10.1096/fj.09-142117. 8
- Gertz, J., Siggia, E.D., and Cohen, B.A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–218, 2009. doi:10.1038/nature07521. 14, 107
- Goffeau, A., Barrell, B., Bussey, H., Davis, R., et al. Life with 6000 genes. *Science*, 274(5287):546, 1996. 99, 121, 198
- Gordan, R., Narlikar, L., and Hartemink, A.J. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Research*, 38(6):e90–e90, 2010. doi:10.1093/nar/gkp1166. 5, 15
- Gordân, R. and Hartemink, A.J. Using DNA duplex stability information for transcription factor binding site discovery. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pp. 453–64, 2008. URL <http://www.ncbi.nlm.nih.gov/pubmed/18229707?dopt=abstract>. 15
- Gotea, V., Visel, A., Westlund, J.M., Nobrega, M.A., et al. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Research*, 20(5):565–577, 2010. doi:10.1101/gr.104471.109. 5
- Gough, J., Karplus, K., Hughey, R., and Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent

REFERENCES

- all proteins of known structure1. *Journal of molecular biology*, 313(4):903–919, 2001. 6
- Gregory, P. Bayesian logical data analysis for the physical sciences: a comparative approach with Mathematica support. *books.google.com*, 2005. URL http://books.google.com/books?hl=en&lr=&id=yJ_5VFo0zGMC&oi=fnd&pg=PR13&dq=+Sciences.&ots=V6MKQvT1Ds&sig=g5r6eIqk_m4J6fiHZRF07AgaNqk. 28
- Griffith, O., Montgomery, S., and Bernier, B. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic acids ...*, 2008. URL http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D107. 33
- Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455(7217):1193–1197, 2008. doi:10.1038/nature07415. 7
- Grosschedl, R., Giese, K., and Pagel, J. HMG domain proteins: architectural elements in the assembly of nucleoprotein structures. *Trends Genet*, 10(3):94–100, 1994. 167
- Grove, C.A., Masi, F.D., Barrasa, M.I., Newburger, D.E., et al. A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell*, 138(2):314–27, 2009. doi:10.1016/j.cell.2009.04.058. URL <http://www.cell.com/retrieve/pii/S0092867409005194>. 12, 35
- Gunewardena, S. and Zhang, Z. A hybrid model for robust detection of transcription factor binding sites. *Bioinformatics*, 24(4):484–91, 2008. doi:10.1093/bioinformatics/btm629. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/24/4/484>. 108
- Habib, N., Kaplan, T., Margalit, H., Friedman, N., et al. A Novel Bayesian DNA Motif Comparison Method for Clustering and Retrieval. *PLoS Computational Biology*, 4(2):e1000010, 2008. doi:10.1371/journal.pcbi.1000010. 41

REFERENCES

- Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004. doi: 10.1038/nature02800. URL <http://dx.doi.org/10.1038/nature02800>. 11, 100, 118, 119, 120, 143, 155, 172
- Hardison, R.C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet*, 16(9):369–72, 2000. 5, 111
- Harrison, S.C. A structural taxonomy of DNA-binding domains. *Nature*, 353(6346):715–9, 1991. doi:10.1038/353715a0. 34
- Hastings, W. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. 177
- Hecht, A. and Grunstein, M. Mapping DNA interaction sites of chromosomal proteins using immunoprecipitation and polymerase chain reaction. *Methods in Enzymology*, 304:399–414, 1999. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B7CV2-4B5PGMJ-48&_user=10&_coverDate=12arch&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=b77cb9d67f8cc5fbc57550afaac012d8&searchtype=a. 33
- Helden, J.V., André, B., and Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, 281(5):827–42, 1998. doi:10.1006/jmbi.1998.1947. 13
- Hertz, G.Z. and Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–77, 1999. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/15/7/563?view=long&pmid=10487864>. 113
- Ho, T. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis . . .*, 1998. URL <http://machine-learning.martinsewell.com/ensembles/rsm/Ho1998.pdf>. 29

REFERENCES

- Holland, R.C.G., Down, T.A., Pocock, M., Prlić, A., et al. BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–7, 2008. doi:10.1093/bioinformatics/btn397. 109
- Hu, J., Yang, Y., and Kihara, D. EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. *BMC bioinformatics*, 7(1):342, 2006. 108
- Hu, Z., Killion, P.J., and Iyer, V.R. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet*, 39(5):683–7, 2007. doi:10.1038/ng2012. URL <http://www.nature.com/ng/journal/v39/n5/abs/ng2012.html>. 100, 118, 120
- Hubbard, T.J.P., Aken, B.L., Ayling, S., Ballester, B., et al. Ensembl 2009. *Nucleic Acids Research*, 37(Database):D690–D697, 2009. doi:10.1093/nar/gkn828. 11, 31, 109, 121, 185
- Huber, W., von Heydebreck, A., Sülthmann, H., Poustka, A., et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002. URL http://bioinformatics.oxfordjournals.org/cgi/reprint/18/suppl_1/S96?view=long&pmid=12169536. 120
- Huerta, A.M., Salgado, H., Thieffry, D., and Collado-Vides, J. RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res*, 26(1):55–59, 1998. 6
- Ishii, T., Yoshida, K., Terai, G., Fujita, Y., et al. DBTBS: a database of Bacillus subtilis promoters and transcription factors. *Nucleic Acids Research*, 29(1):278–80, 2001. URL <http://nar.oxfordjournals.org/cgi/content/full/29/1/278?view=long&pmid=11125112>. 6
- Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409(6819):533–8, 2001. doi:10.1038/35054095. URL <http://www.nature.com/nature/journal/v409/n6819/full/409533a0.html>. 118, 120

REFERENCES

- Jacquier, A. Applications of next-generation sequencing: The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Reviews Genetics*, 10(12):833–844, 2009. doi:10.1038/nrg2683. URL <http://dx.doi.org/10.1038/nrg2683>. 8
- Jaenisch, R. and Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*, 33 Suppl:245–54, 2003. doi:10.1038/ng1089. 8
- Janga, S., Collado-Vides, J., and Babu, M. Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proceedings of the National Academy of Sciences*, 105(41):15761, 2008. 3
- Johnson, S. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967. XIII, 84, 116, 154
- Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*, pp. 1–14, 2010. doi:10.1101/gr.100552.109. 12
- Jones and Pevzner. *Bioinformatics*, 2006. 101
- Jones, S., van Heyningen, P., Berman, H.M., and Thornton, J.M. Protein-DNA interactions: A structural analysis. *J Mol Biol*, 287(5):877–96, 1999. doi:10.1006/jmbi.1999.2659. 38
- Kafri, R., Bar-Even, A., and Pilpel, Y. Transcription control reprogramming in genetic backup circuits. *Nat Genet*, 37(3):295–9, 2005. doi:10.1038/ng1523. 90
- Kanamori, M., Konno, H., Osato, N., Kawai, J., et al. A genome-wide and nonredundant mouse transcription factor database. *Biochem Biophys Res Commun*, 322(3):787–93, 2004. doi:10.1016/j.bbrc.2004.07.179. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WBK-4D4PPRP-J&_user=10&_coverDate=09rch&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=247cd1528a22107a65007eaa408a4b93. 6

REFERENCES

- Kaplan, T., Friedman, N., and Margalit, H. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol*, 1(1):e1, 2005. doi:10.1371/journal.pcbi.0010001. URL <http://dx.doi.org/10.1371/journal.pcbi.0010001>. 39
- Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., et al. Variation in Transcription Factor Binding Among Humans. *Science*, 328(5975):232–235, 2010. doi:10.1126/science.1183621. 112
- Kechris, K. and Li, H. c-REDUCE: incorporating sequence conservation to detect motifs that correlate with expression. *BMC bioinformatics*, 9:506, 2008. doi:10.1186/1471-2105-9-506. URL <http://www.biomedcentral.com/1471-2105/9/506>. 15
- Keleş, S., van der Laan, M., and Eisen, M.B. Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18(9):1167–75, 2002. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/18/9/1167>. 15
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B., et al. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–54, 2003. doi:10.1038/nature01644. 101, 111
- Kielbasa, S.M., Gonze, D., and Herzel, H. Measuring similarities between transcription factor binding sites. *BMC Bioinformatics*, 6:237, 2005. doi:10.1186/1471-2105-6-237. URL <http://dx.doi.org/10.1186/1471-2105-6-237>. 41
- Kim. Crystal structure of a yeast TBP/TATA-box complex. *Nature*, 1995. 39
- Kim, J.L. and Burley, S.K. 1.9 Å resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nat Struct Biol*, 1(9):638–53, 1994. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=7634103&dopt=abstractplus. 165
- Kim, J., Cunningham, R., James, B., Wyder, S., et al. Functional characterization of transcription factor motifs using cross-species comparison across

REFERENCES

- large evolutionary distances. *PLoS Comput Biol*, 6(1):e1000652, 2010. doi:10.1371/journal.pcbi.1000652. 5
- Kirkpatrick, C.R. and Schimmel, P. Detection of leucine-independent DNA site occupancy of the yeast Leu3p transcriptional activator in vivo. *Mol Cell Biol*, 15(8):4021–30, 1995. URL <http://mcb.asm.org/cgi/reprint/15/8/4021?view=long&pmid=7623798>. 144
- Klepper, K., Sandve, G., Abul, O., Johansen, J., et al. Assessment of composite motif discovery methods. *BMC Bioinformatics*, 9(1):123, 2008. doi:10.1186/1471-2105-9-123. URL <http://www.biomedcentral.com/1471-2105/9/123>. 108
- Kohonen, T. and Somervuo, P. How to make large self-organizing maps for nonvectorial data. *Neural Netw*, 15(8-9):945–52, 2002. 41
- Kono, H. and Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, 35(1):114–31, 1999. 38
- Korn, L.J., Queen, C.L., and Wegman, M.N. Computer analysis of nucleic acid regulatory sequences. *Proc Natl Acad Sci USA*, 74(10):4401–5, 1977. 13
- Kroeger, P.E. and Morimoto, R.I. Selection of new HSF1 and HSF2 DNA-binding sites reveals difference in trimer cooperativity. *Mol Cell Biol*, 14(11):7592–603, 1994. URL <http://mcb.asm.org/cgi/reprint/14/11/7592?view=long&pmid=7935474>. 50
- Krogh, A., Brown, M., Mian, I.S., Sjölander, K., et al. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*, 235(5):1501–31, 1994. doi:10.1006/jmbi.1994.1104. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WK7-45NSKPC-N4&_user=10&_coverDate=02arch&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=dfe80e989968625b45adca07c6681ed4. 19, 68
- Kulkarni, M.M. and Arnosti, D.N. cis-regulatory logic of short-range transcriptional repression in *Drosophila melanogaster*. *Molecular and Cellular Biology*,

REFERENCES

- 25(9):3411–20, 2005. doi:10.1128/MCB.25.9.3411-3420.2005. URL <http://mcb.asm.org/cgi/content/full/25/9/3411?view=long&pmid=15831448>. 5
- Kummerfeld, S.K. and Teichmann, S.A. DBD: a transcription factor prediction database. *Nucleic Acids Res*, 34(Database issue):D74–D81, 2006. doi:10.1093/nar/gkj131. URL <http://dx.doi.org/10.1093/nar/gkj131>. 6
- Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet*, 42(7):631–4, 2010. doi:10.1038/ng.600. URL <http://www.nature.com/ng/journal/v42/n7/full/ng.600.html>. 112
- Kwong, C., Adryan, B., Bell, I., Meadows, L., et al. Stability and dynamics of polycomb target sites in *Drosophila* development. *PLoS Genetics*, 4(9), 2008. 67
- Lagrange, T., Kapanidis, A.N., Tang, H., Reinberg, D., et al. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev*, 12(1):34–44, 1998. 123
- Lähdesmäki, H., Rust, A.G., and Shmulevich, I. Probabilistic inference of transcription factor binding from multiple data sources. *PLoS ONE*, 3(3):e1820, 2008. doi:10.1371/journal.pone.0001820. URL <http://www.plosone.org/article/infojournal.pone.0001820>. 107
- Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, 1993. URL <http://www.sciencemag.org/cgi/reprint/262/5131/208>. 27
- Lebrecht, D., Foehr, M., Smith, E., Lopes, F.J.P., et al. Bicoid cooperative DNA binding is critical for embryonic patterning in *Drosophila*. *Proc Natl Acad Sci USA*, 102(37):13176–81, 2005. doi:10.1073/pnas.0506462102. URL <http://www.pnas.org/content/102/37/13176.long>. 5
- LeClerc, S., Palaniswami, R., Xie, B.X., and Govindan, M.V. Molecular cloning and characterization of a factor that binds the human glucocorticoid

REFERENCES

- receptor gene and represses its expression. *J Biol Chem*, 266(26):17333–17340, 1991. URL <http://www.google.com/search?client=safari&rls=en-us&q=Molecular+cloning+and+characterization+of+a+factor+that+binds+the+human+glucocorticoid+receptor+gene+and+represses+its+expression.&ie=UTF-8&oe=UTF-8>. 14
- Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, 2002. doi:10.1126/science.1075090. URL <http://dx.doi.org/10.1126/science.1075090>. 100, 112
- Lefrançois, P., Euskirchen, G.M., Auerbach, R.K., Rozowsky, J., et al. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics*, 10:37, 2009. doi:10.1186/1471-2164-10-37. URL <http://www.biomedcentral.com/1471-2164/10/37>. 11
- Legras, J., Merdinoglu, D., and Cornuet, J. Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Molecular ...*, 2007. URL <http://www3.interscience.wiley.com/journal/117989623/abstract>. 121
- Levine, M. and Davidson, E.H. Gene regulatory networks for development. *Proc Natl Acad Sci U S A*, 102(14):4936–4942, 2005. doi:10.1073/pnas.0408031102. URL <http://dx.doi.org/10.1073/pnas.0408031102>. 176
- Lewis, M.A., Quint, E., Glazier, A.M., Fuchs, H., et al. An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. *Nat Genet*, 41(5):614–8, 2009. doi:10.1038/ng.369. URL <http://www.nature.com/ng/journal/v41/n5/abs/ng.369.html>. 32, 67, 99
- Li, B., Carey, M., and Workman, J. The role of chromatin during transcription. *Cell*, 128(4):707–719, 2007. 2
- Li, N. and Tompa, M. Analysis of computational approaches for motif discovery. *Algorithms Mol Biol*, 1:8, 2006. doi:10.1186/1748-7188-1-8. URL <http://dx.doi.org/10.1186/1748-7188-1-8>. 16, 108

REFERENCES

- Li, X., Zhong, S., and Wong, W.H. Reliable prediction of transcription factor binding sites by phylogenetic verification. *Proc Natl Acad Sci USA*, 102(47):16945–50, 2005. doi:10.1073/pnas.0504201102. 112, 165
- Liaw, A. and Wiener, M. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002. URL <http://CRAN.R-project.org/doc/Rnews/>. 31, 84
- Lieb, J.D., Liu, X., Botstein, D., and Brown, P.O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet*, 28(4):327–34, 2001. doi:10.1038/ng569. URL <http://www.nature.com/doifinder/10.1038/ng569>. 118, 120
- Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, 2009. doi:10.1126/science.1181369. URL <http://www.sciencemag.org/cgi/pmidlookup?view=short&pmid=19815776>. 3
- Liti, G., Carter, D.M., Moses, A.M., Warringer, J., et al. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–41, 2009. doi:10.1038/nature07743. 121
- Liu, X., Brutlag, D.L., and Liu, J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pp. 127–38, 2001. URL <http://www.ncbi.nlm.nih.gov/pubmed/11262934?dopt=abstract>. 113
- Liu, X., Noll, D., Lieb, J., and Clarke, N. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res*, 15(3):421, 2005. 12, 114
- Liu, X., Brutlag, D., and Liu, J. An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature biotechnology*, 20(8):835–839, 2002. 15, 102, 113

REFERENCES

- Loh, Y., Wu, Q., Chew, J., Vega, V., et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet*, 38(4):431–440, 2006. [67](#)
- Loots, G., Locksley, R., Blankespoor, C.M., Wang, Z., et al. Identification of a coordinate regulator of interleukins 4, 13 and 5 by cross-species sequence comparisons. *Science*, 2000. URL <http://www.google.com/search?client=safari&rls=en-us&q=Identification+of+a+coordinate+regulator+of+interleukins+4,+13+and+5+by+cross-species+sequence+comparisons.&ie=UTF-8&oe=UTF-8>. [5](#)
- Lu, C.C., Yuan, W.H., and Chen, T.M. Extracting transcription factor binding sites from unaligned gene sequences with statistical models. *BMC Bioinformatics*, 9(Suppl 12):S7, 2008. doi:10.1186/1471-2105-9-S12-S7. [108](#)
- Ludwig, M. Functional evolution of noncoding DNA. *Current opinion in genetics & development*, 12(6):634–639, 2002. [5](#)
- Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. An overview of the structures of protein-DNA complexes. *Genome Biol*, 1(1):REVIEWS001, 2000. [34](#), [123](#)
- Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*, 29(13):2860–2874, 2001. [40](#)
- Luscombe, N.M. and Thornton, J.M. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *Journal of molecular biology*, 320(5):991–1009, 2002. [38](#), [39](#), [88](#)
- Lusk, R.W. and Eisen, M.B. Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. *PLoS Genet*, 6(1):e1000829, 2010. doi:10.1371/journal.pgen.1000829. [6](#)
- MacIsaac, K., Lo, K., Gordon, W., Motola, S., et al. A Quantitative Model of Transcriptional Regulation Reveals the Influence of Binding Location on Expression. 2010. [123](#)

REFERENCES

- MacIsaac, K.D. and Fraenkel, E. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Computational Biology*, 2(4):e36, 2006. doi:10.1371/journal.pcbi.0020036. 13
- MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., et al. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 7:113, 2006. doi:10.1186/1471-2105-7-113. URL <http://dx.doi.org/10.1186/1471-2105-7-113>. 11, 83, 115, 120
- Macpherson, S., Larochele, M., and Turcotte, B. A Fungal Family of Transcriptional Regulators: the Zinc Cluster Proteins. *Microbiology and Molecular Biology Reviews*, 70(3):583–604, 2006. doi:10.1128/MMBR.00015-06. 35, 124
- Maerkl, S.J. and Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007a. doi:10.1126/science.1131007. URL <http://dx.doi.org/10.1126/science.1131007>. 12
- Maerkl, S.J. and Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–7, 2007b. doi:10.1126/science.1131007. 18
- Maerkl, S.J. and Quake, S.R. Experimental determination of the evolvability of a transcription factor. *Proc Natl Acad Sci USA*, 106(44):18650–5, 2009. doi:10.1073/pnas.0907688106. 12, 90
- Mahony, S., Auron, P.E., and Benos, P.V. DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies. *PLoS Comput Biol*, 3(3):e61, 2007. doi:10.1371/journal.pcbi.0030061. URL <http://dx.doi.org/10.1371/journal.pcbi.0030061>. 41
- Mahony, S. and Benos, P.V. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res*, 35(Web Server issue):W253–W258, 2007. doi:10.1093/nar/gkm272. URL <http://dx.doi.org/10.1093/nar/gkm272>. 58, 183

REFERENCES

- Mahony, S., Golden, A., Smith, T.J., and Benos, P.V. Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Bioinformatics*, 21 Suppl 1:i283–i291, 2005a. doi:10.1093/bioinformatics/bti1025. URL <http://dx.doi.org/10.1093/bioinformatics/bti1025>. 41, 68, 77
- Mahony, S., Hendrix, D., Golden, A., Smith, T.J., et al. Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, 21(9):1807–14, 2005b. doi:10.1093/bioinformatics/bti256. 77, 111, 200
- Makeev, V.J., Lifanov, A.P., Nazina, A.G., and Papatsenko, D.A. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res*, 31(20):6016–26, 2003. URL <http://nar.oxfordjournals.org/cgi/content/full/31/20/6016?view=long&pmid=14530449>. 8
- Mandel-Gutfreund, Y., Baron, A., and Margalit, H. A structure-based approach for prediction of protein binding sites in gene upstream regions. *Pac Symp Biocomput*, pp. 139–50, 2001. 39
- Markstein, M. and Levine, M. Decoding cis-regulatory DNAs in the Drosophila genome. *Current opinion in genetics & development*, 12(5):601–606, 2002. 5
- Marschall, T. and Rahmann, S. Efficient exact motif discovery. *Bioinformatics*, 25(12):i356–64, 2009. doi:10.1093/bioinformatics/btp188. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/25/12/i356>. 13
- Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110, 2006. doi:10.1093/nar/gkj143. URL <http://dx.doi.org/10.1093/nar/gkj143>. X, XI, 33, 50, 51, 63, 64, 84, 87, 104, 183, 194
- McDaniell, R., Lee, B.K., Song, L., Liu, Z., et al. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, 328(5975):235–9, 2010. doi:10.1126/science.1184655. URL <http://www.sciencemag.org/cgi/content/abstract/328/5975/235>. 112

REFERENCES

- Megraw, M., Pereira, F., Jensen, S.T., Ohler, U., et al. A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res*, 19(4):644–56, 2009. doi:10.1101/gr.085449.108. URL <http://genome.cshlp.org/content/19/4/644.long>. 1
- Meng, X., Brodsky, M.H., and Wolfe, S.A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat Biotechnol*, 23(8):988–994, 2005. doi:10.1038/nbt1120. 183
- Meng, X. and Wolfe, S.A. Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nat Methods*, 1(1):30–45, 2006. doi:10.1038/nprot.2006.6. 12
- Mewes, H., Albermann, K., Bähr, M., Frishman, D., et al. Overview of the yeast genome. *Nature*, 387(6632):7–8, 1997. 121
- Meyer, D., Leisch, F., and Hornik, K. Benchmarking. 2003. URL <http://citeseer.ist.psu.edu/619009>. 30
- Minichiello, M.J. and Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet*, 79(5):910–22, 2006. doi:10.1086/508901. URL <http://www.ncbi.nlm.nih.gov/pubmed/17033967?dopt=abstract>. 122
- Minka, T.P. Estimating a Dirichlet distribution, 2003. URL <http://research.microsoft.com/users/Cambridge/minka/papers/dirichlet/minka-dirichlet.ps>. 48, 49, 51, 52, 58, 85
- Mintseris, J. and Eisen, M.B. Design of a combinatorial DNA microarray for protein-DNA interaction studies. *BMC bioinformatics*, 7:429, 2006. doi:10.1186/1471-2105-7-429. URL <http://www.biomedcentral.com/1471-2105/7/429>. 114
- Mitchell, P.J. and Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245(4916):371–8, 1989. 1

REFERENCES

- Montgomery, S.B., Astakhova, T., Bilenky, M., Birney, E., et al. Sockeye: a 3D environment for comparative genomics. *Genome Res*, 14(5):956–62, 2004. doi:10.1101/gr.1890304. URL <http://genome.cshlp.org/content/14/5/956.long>. 183
- Morley, R.H., Lachani, K., Keefe, D., Gilchrist, M.J., et al. A gene regulatory network directed by zebrafish No tail accounts for its roles in mesoderm formation. *Proc Natl Acad Sci USA*, 106(10):3829–34, 2009. doi:10.1073/pnas.0808382106. 67
- Moses, A., Chiang, D., and Kellis, M. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evolutionary ...*, 2003. URL <http://www.biomedcentral.com/1471-2148/3/19>. 14
- Mukherjee, P. and Parkinson, D. A nested sampling algorithm for cosmological model selection. *The Astrophysical Journal ...*, 2006. URL <http://iopscience.iop.org/1538-4357/638/2/L51>. 25
- Mukherjee, S., Berger, M., Jona, G., and Wang, X. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 2004. URL <http://www.nature.com/ng/journal/vaop/ncurrent/full/ng1473.html>. 12
- Mustonen, V., Kinney, J., Callan, C.G., and Lässig, M. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci USA*, 105(34):12376–81, 2008. doi:10.1073/pnas.0805909105. 18
- Nadassy, K., Wodak, S.J., and Janin, J. Structural features of protein-nucleic acid recognition sites. *Biochemistry*, 38(7):1999–2017, 1999. doi:10.1021/bi982362d. 38
- Narlikar, L., Gordân, R., and Hartemink, A.J. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput Biol*, 3(11):e215, 2007. doi:10.1371/journal.pcbi.0030215. URL <http://dx.doi.org/10.1371/journal.pcbi.0030215>. 15

REFERENCES

- Narlikar, L., Gordân, R., Ohler, U., and Hartemink, A.J. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, 22(14):e384–e392, 2006. doi:10.1093/bioinformatics/btl251. URL <http://dx.doi.org/10.1093/bioinformatics/btl251>. 15, 39, 42, 67, 69, 82, 111, 113
- Narlikar, L. and Hartemink, A.J. Sequence features of DNA binding sites reveal structural class of associated transcription factor. *Bioinformatics*, 22(2):157–163, 2006. doi:10.1093/bioinformatics/bti731. URL <http://dx.doi.org/10.1093/bioinformatics/bti731>. 34, 40, 41, 82, 83, 84, 86, 88, 91
- Needleman, S. and Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970. 40
- Newburger, D.E. and Bulyk, M.L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 37(Database):D77–D82, 2009. doi:10.1093/nar/gkn660. 50, 183
- Nguyen, T.T. and Androulakis, I.P. Recent Advances in the Computational Discovery of Transcription Factor Binding Sites. *Algorithms*, 2(1):582–605, 2009. doi:10.3390/a2010582. URL <http://www.mdpi.com/1999-4893/2/1/582>. 13
- Nielsen, R. Statistical methods in molecular evolution. p. 504, 2005. URL <http://books.google.com/books?id=nJipT3toWFAC&printsec=frontcover>. 121
- Nieto, M.A. The snail superfamily of zinc-finger transcription factors. *Nat Rev Mol Cell Biol*, 3(3):155–66, 2002. doi:10.1038/nrm757. 90
- Notredame, C., Higgins, D.G., and Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17, 2000. doi:10.1006/jmbi.2000.4042. 48
- Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., et al. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, 133(7):1277–89, 2008a. doi:10.1016/j.

REFERENCES

- cell.2008.05.023. URL [http://linkinghub.elsevier.com/retrieve/pii/S0092-8674\(08\)00682-X](http://linkinghub.elsevier.com/retrieve/pii/S0092-8674(08)00682-X). XI, 12, 36, 91, 92, 93
- Noyes, M.B., Meng, X., Wakabayashi, A., Sinha, S., et al. A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Research*, 36(8):2547–60, 2008b. doi:10.1093/nar/gkn048. URL <http://nar.oxfordjournals.org/cgi/content/full/36/8/2547>. 12
- Oliphant, A.R., Brandl, C.J., and Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Molecular and Cellular Biology*, 9(7):2944–9, 1989. URL <http://mcb.asm.org/cgi/reprint/9/7/2944?view=long&pmid=2674675>. 33
- Orengo, C.A. and Taylor, W.R. SSAP: sequential structure alignment program for protein structure comparison. *Meth Enzymol*, 266:617–35, 1996. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=8743709&dopt=abstractplus. 34, 123
- Osada, R., Zaslavsky, E., and Singh, M. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, 20(18):3516–25, 2004. doi:10.1093/bioinformatics/bth438. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/20/18/3516>. 17, 18
- Pabo, C.O., Peisach, E., and Grant, R.A. Design and selection of novel Cys2His2 zinc finger proteins. *Annu Rev Biochem*, 70:313–40, 2001. doi:10.1146/annurev.biochem.70.1.313. 39
- Pachkov, M., Erb, I., Molina, N., and van Nimwegen, E. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res*, 35(Database issue):D127–D131, 2007. doi:7. URL <http://dx.doi.org/7>. 114
- Papatsenko. Organization of developmental enhancers in the *Drosophila* embryo. *Nucleic Acids Res*, 2009. URL <http://www.google.com/search?client=safari&rls=en-us&q=Organization+of+developmental+enhancers+in+the+Drosophila+embryo.&ie=UTF-8&oe=UTF-8>. 8

REFERENCES

- Pape, U.J., Rahmann, S., and Vingron, M. Natural Similarity Measures between Position Frequency Matrices with an Application to Clustering. *Bioinformatics*, 2008. doi:10.1093/bioinformatics/btm610. URL <http://dx.doi.org/10.1093/bioinformatics/btm610>. 41
- Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., et al. ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37(Database issue):D868–72, 2009. doi: 10.1093/nar/gkn889. URL http://nar.oxfordjournals.org/cgi/content/full/37/suppl_1/D868?view=long&pmid=19015125. 120
- Pavesi, G., Mauri, G., and Pesole, G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17 Suppl 1:S207–14, 2001. URL http://bioinformatics.oxfordjournals.org/cgi/reprint/17/suppl_1/S207?view=long&pmid=11473011. 13, 197
- Peng, C., Hsu, J., Chung, Y., Lin, Y., et al. Identification of degenerate motifs using position restricted selection and hybrid ranking combination. *Nucleic acids research*, 2006. 108
- Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, 2006. doi:10.1038/nature05295. URL <http://www.nature.com/nature/journal/v444/n7118/abs/nature05295.html>. 2
- Percipalle, P., Simoncsits, A., Zakhariyev, S., Guarnaccia, C., et al. Rationally designed helix-turn-helix proteins and their conformational changes upon DNA binding. *EMBO J*, 14(13):3200–5, 1995. 39
- Persikov, A.V., Osada, R., and Singh, M. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, 25(1):22–29, 2008. doi: 10.1093/bioinformatics/btn580. 39
- Pevzner, P.A. and Sze, S.H. Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol*, 8:269–78, 2000. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=10977088&dopt=abstractplus. 16, 103, 105, 174

REFERENCES

- Pfreundt, U., James, D.P., Tweedie, S., Wilson, D., et al. FlyTF: improved annotation and enhanced functionality of the Drosophila transcription factor database. *Nucleic Acids Res*, 38(Database issue):D443–7, 2010. doi:10.1093/nar/gkp910. URL http://nar.oxfordjournals.org/cgi/content/full/38/suppl_1/D443?view=long&pmid=19884132. 6
- Pic, A., Lim, F.L., Ross, S.J., Veal, E.A., et al. The forkhead protein Fkh2 is a component of the yeast cell cycle transcription factor SFF. *The EMBO Journal*, 19(14):3750–61, 2000. doi:10.1093/emboj/19.14.3750. URL <http://www.nature.com/emboj/journal/v19/n14/abs/7593192a.html>. 167
- Pietrokovski, S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, 24(19):3836, 1996. 39
- Piipari, M., Down, T., and Hubbard, T. Large-scale gene regulatory motif discovery with NestedMICA. *Advances in Genomic Sequence Analysis and pattern Discovery*, 2011. 32, 188
- Piipari, M., Down, T.A., and Hubbard, T.J. Metamotifs - a generative model for building families of nucleotide position weight matrices. *BMC Bioinformatics*, 11(348):1–24, 2010a. doi:10.1186/1471-2105-11-348. 25, 32, 38, 173, 174
- Piipari, M., Down, T.A., Saini, H., Enright, A., et al. iMotifs: an integrated sequence motif visualization and analysis environment. *Bioinformatics*, 26(6):843–4, 2010b. doi:10.1093/bioinformatics/btq026. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/26/6/843?view=long&pmid=20106815>. 32, 46, 48, 74, 77, 78, 109, 111, 166, 174, 182
- Pompeani, A.J., Irgon, J.J., Berger, M.F., Bulyk, M.L., et al. The *Vibrio harveyi* master quorum-sensing regulator, LuxR, a TetR-type protein is both an activator and a repressor: DNA recognition and binding specificity at target promoters. *Mol Microbiol*, 70(1):76–88, 2008. doi:10.1111/j.1365-2958.2008.06389.x. URL <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2008.06389.x/abstract>. 35

REFERENCES

- Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 38(Database):D105–D110, 2010. doi:10.1093/nar/gkp950. [12](#), [34](#), [40](#), [50](#), [114](#), [115](#)
- Qin, Z., McCue, L., Thompson, W., Mayerhofer, L., et al. Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol*, 21(4):435–439, 2003. [77](#)
- Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE DOI - 10.1109/5.18626*, 77(2):257–286, 1989. URL [10.1109/5.18626](#). [19](#)
- Rahmann, S., Müller, T., and Vingron, M. On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol*, 2:Article7, 2003. doi:10.2202/1544-6115.1032. URL <http://dx.doi.org/10.2202/1544-6115.1032>. [17](#)
- Ramsey, S.A., Knijnenburg, T.A., Kennedy, K.A., Zak, D.E., et al. Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, 26(17):2071–2075, 2010. doi:10.1093/bioinformatics/btq405. [107](#)
- Rastegar, S., Hess, I., Dickmeis, T., Nicod, J.C., et al. The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Dev Biol*, 318(2):366–77, 2008. doi:10.1016/j.ydbio.2008.03.034. [67](#)
- Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., et al. An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. *Cell*, 140(5):744–752, 2010. doi:10.1016/j.cell.2010.01.044. [8](#), [176](#)
- Reddy, T.E., Delisi, C., and Shakhnovich, B.E. Binding Site Graphs: A New Graph Theoretical Framework for Prediction of Transcription Factor Binding Sites. *PLoS Comput Biol*, 3(5):e90, 2007. doi:10.1371/journal.pcbi.0030090. URL <http://dx.doi.org/10.1371/journal.pcbi.0030090>. [108](#)

REFERENCES

- Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., et al. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res*, 10(4):483–501, 2000. URL <http://genome.cshlp.org/content/10/4/483.long>. 106
- Reimand, J., Vaquerizas, J.M., Todd, A.E., Vilo, J., et al. Comprehensive re-analysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Research*, pp. 1–10, 2010. doi:10.1093/nar/gkq232. 100, 118, 120, 143, 144, 145
- Reiner, A., Yekutieli, D., and Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–75, 2003. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/19/3/368?view=long&pmid=12584122>. 120
- Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651–7, 2007. doi:10.1038/nmeth1068. 8, 12, 188
- Robinson, M., Sun, Y., Boekhorst, R.T., Kaye, P., et al. Improving computational predictions of cis-regulatory binding sites. *Biocomputing-Proceedings of the Pacific Symposium*, 2006. URL <http://homepages.feis.herts.ac.uk/~nngroup/pubs/papers/Robinson-PSB05.pdf>. 108
- Rolfes, R.J., Zhang, F., and Hinnebusch, A.G. The transcriptional activators BAS1, BAS2, and ABF1 bind positive regulatory sites as the critical elements for adenine regulation of ADE5,7. *J Biol Chem*, 272(20):13343–54, 1997. URL <http://www.jbc.org/content/272/20/13343.long>. 167
- Ross, E.D., Keating, A.M., and 3RD, M.L. DNA constraints on transcription activation in vitro. *J Mol Biol*, 297(2):321–34, 2000. doi:10.1006/jmbi.2000.3562. 123
- Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature biotechnology*, 16(10):939–45, 1998. doi:10.1038/nbt1098-939. 15, 101, 102, 198

- Sagot, M. Spelling approximate repeated or common motifs using a suffix tree. *Latin'98: Theoretical Informatics*, 1998. URL <http://www.springerlink.com/index/1469887m40070445.pdf>. 14
- Salgado, H., Gama-Castro, S., Peralta-Gil, M., Díaz-Peredo, E., et al. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*, 34(Database issue):D394–D397, 2006. doi:10.1093/nar/gkj156. URL <http://dx.doi.org/10.1093/nar/gkj156>. 6
- Saltzman, A.G. and Weinmann, R. Promoter specificity and modulation of RNA polymerase II transcription. *FASEB J*, 3(6):1723–33, 1989. URL <http://www.fasebj.org/cgi/reprint/3/6/1723>. 1
- Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–D94, 2004. doi:10.1093/nar/gkh012. URL <http://dx.doi.org/10.1093/nar/gkh012>. 34
- Sandelin, A. and Wasserman, W.W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol*, 338(2):207–215, 2004. doi:10.1016/j.jmb.2004.02.048. URL <http://dx.doi.org/10.1016/j.jmb.2004.02.048>. 39, 40, 42, 68, 77, 82, 91
- Sandve, G.K., Abul, O., and Drablos, F. Compo: composite motif discovery using discrete models. *BMC Bioinformatics*, 9(1):527, 2008. doi:10.1186/1471-2105-9-527. 108
- Sandve, G.K., Abul, O., Walseng, V., and Drabløs, F. Improved benchmarks for computational motif discovery. *BMC bioinformatics*, 8:193, 2007. doi:10.1186/1471-2105-8-193. 108
- Sandve, G.K. and Drabløs, F. A survey of motif discovery methods in an integrated framework. *Biol Direct*, 1:11, 2006. doi:10.1186/1745-6150-1-11. URL <http://www.biology-direct.com/content/1/1/11>. 13

REFERENCES

- Saxonov, S., Berg, P., and Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci USA*, 103(5):1412–7, 2006. doi:10.1073/pnas.0510310103. URL <http://www.pnas.org/cgi/content/full/103/5/1412>. 28
- Scharer, C.D., McCabe, C.D., Ali-Seyed, M., Berger, M.F., et al. Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells. *Cancer Res*, 69(2):709–17, 2009. doi:10.1158/0008-5472.CAN-08-3415. URL <http://cancerres.aacrjournals.org/content/69/2/709.long>. 35
- Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., et al. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science*, 2010. doi:10.1126/science.1186176. 5, 112
- Schneider, T.D. and Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–100, 1990. 14, 18, 45, 46
- Sharon, E., Lubliner, S., Segal, E., and Stormo, G. A Feature-Based Approach to Modeling Protein–DNA Interactions. *PLoS Comput Biol*, 4(8):e1000154, 2008. doi:10.1371/journal.pcbi.1000154. 18
- Shaw, J.R., Bridges, M., and Hobson, M.P. Efficient Bayesian inference for multimodal problems in cosmology. *arXiv*, astro-ph, 2007. doi:10.1111/j.1365-2966.2007.11871.x. URL <http://arxiv.org/abs/astro-ph/0701867v2>. 25
- Sherman, D., Durrens, P., Beyne, E., Nikolski, M., et al. Genolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts. *Nucleic Acids Res*, 32(Database Issue):D315, 2004. 109
- Siddharthan, R. PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. *PLoS Computational Biology*, 4(8):e1000156, 2008. doi:10.1371/journal.pcbi.1000156. URL <http://www.ploscompbiol.org/article/info252Fjournal.pcbi.1000156>. 16, 69, 77
- Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*,

REFERENCES

- 15(8):1034–50, 2005. doi:10.1101/gr.3715005. URL <http://genome.cshlp.org/content/15/8/1034.long>. 112, 121
- Sierro, N., Makita, Y., de Hoon, M., and Nakai, K. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Research*, 36(Database issue):D93–6, 2008. doi:10.1093/nar/gkm910. URL http://nar.oxfordjournals.org/cgi/content/full/36/suppl_1/D93?view=long&pmid=17962296. 6
- Siggia, E.D. Computational methods for transcriptional regulation. *Current opinion in genetics & development*, 15(2):214–21, 2005. doi:10.1016/j.gde.2005.02.004. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VSO-4FK3PDM-2&_user=10&_coverDate=04rch&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=c765281eea85c83bfed2db3fa697914e. 5
- Sikder, D. and Kodadek, T. Genomic studies of transcription factor-DNA interactions. *Curr Opin Chem Biol*, 9(1):38–45, 2005. doi:10.1016/j.cbpa.2004.12.008. URL <http://dx.doi.org/10.1016/j.cbpa.2004.12.008>. 11
- Silva, E.K.D., Gehrke, A.R., Olszewski, K., León, I., et al. Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc Natl Acad Sci USA*, 105(24):8393–8, 2008. doi:10.1073/pnas.0801993105. URL <http://www.pnas.org/content/105/24/8393.long>. 35
- Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. ROCRC: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, 2005. doi:10.1093/bioinformatics/bti623. URL <http://dx.doi.org/10.1093/bioinformatics/bti623>. 124
- Sinha, S., Blanchette, M., and Tompa, M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5:170, 2004. doi:10.1186/1471-2105-5-170. URL <http://dx.doi.org/10.1186/1471-2105-5-170>. 16

REFERENCES

- Sinha, S. and Tompa, M. Performance Comparison of Algorithms for Finding Transcription Factor Binding Sites. 2003a. URL <http://citeseer.ist.psu.edu/592637>. Fetch Buhler & Tompa. 16
- Sinha, S. and Tompa, M. YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 31(13):3586–8, 2003b. URL <http://nar.oxfordjournals.org/cgi/content/full/31/13/3586?view=long&pmid=12824371>. 199
- Skilling, J. Nested Sampling for General Bayesian Computation. 2004. 23, 52
- Smith. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. 1987. 26
- Smith, T. Secret code. *Nature Structural Biology*, 5(2):100, 1998. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=9461070&dopt=abstractplus. 38
- Sonnhammer, E., Eddy, S., and Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Bioinformatics*, 28(3):405–420, 1997. 6, 19
- Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., et al. The Ensembl core software libraries. *Genome Res*, 14(5):929–33, 2004. doi:10.1101/gr.1857204. 32
- Stanke, M. and Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(Suppl 2), 2003. 19
- Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., et al. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, 450(7167):219–32, 2007. doi:10.1038/nature06340. URL <http://www.nature.com/nature/journal/v450/n7167/full/nature06340.html>. 5
- Starr, D.B. and Hawley, D.K. TFIID binds in the minor groove of the TATA box. *Cell*, 67(6):1231–40, 1991. URL <http://www.cell.com/retrieve/pii/009286749190299E>. 165

REFERENCES

- Steinfeld, I., Shamir, R., and Kupiec, M. A genome-wide analysis in *Saccharomyces cerevisiae* demonstrates the influence of chromatin modifiers on transcription. *Nat Genet*, 39(3):303–309, 2007. doi:10.1038/ng1965. 2
- Stormo, G.D., Schneider, T.D., Gold, L., and Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Research*, 10(9):2997–3011, 1982. 39
- Sudarsanam, P., Pilpel, Y., and Church, G.M. Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res*, 12(11):1723–1731, 2002. doi:10.1101/gr.301202. URL <http://dx.doi.org/10.1101/gr.301202>. 165
- Tamada, Y., Kim, S., Bannai, H., Imoto, S., et al. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, 19 Suppl 2:ii227–36, 2003. 167
- Tanaka, N., Kawakami, T., and Taniguchi, T. Recognition DNA sequences of interferon regulatory factor 1 (IRF-1) and IRF-2, regulators of cell growth and the interferon system. *Molecular and Cellular Biology*, 13(8):4531–8, 1993. URL <http://mcb.asm.org/cgi/reprint/13/8/4531?view=long&pmid=7687740>. 90
- Tang, M.H.E., Krogh, A., and Winther, O. BayesMD: Flexible Biological Modeling for Motif Discovery. *Journal of Computational Biology*, 15(10):1347–1363, 2008. doi:10.1089/cmb.2008.15.issue-10. 28, 56, 71, 111, 113, 174
- Team, R.D.C. *R: A Language and Environment for Statistical Computing*, 2007. URL <http://www.R-project.org>. ISBN 3-900051-07-0. 84
- Teixeira, M.C., Monteiro, P., Jain, P., Tenreiro, S., et al. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 34(Database issue):D446–51, 2006. doi:10.1093/nar/gkj013. URL http://nar.oxfordjournals.org/cgi/content/full/34/suppl_1/D446?view=long&pmid=16381908. 100, 118, 119, 143

REFERENCES

- Thijs, G., Lescot, M., Marchal, K., Rombauts, S., et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–22, 2001. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/17/12/1113?view=long&pmid=11751219>. 117, 198
- Thijs, G., Marchal, K., Lescot, M., Rombauts, S., et al. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, 9(2):447–464, 2002. doi:10.1089/10665270252935566. URL <http://dx.doi.org/10.1089/10665270252935566>. 102
- Thomas-Chollier, M., Sand, O., Turatsinze, J.V., Janky, R., et al. RSAT: regulatory sequence analysis tools. *Nucleic Acids Research*, 36(Web Server issue):W119–27, 2008. doi:10.1093/nar/gkn304. URL http://nar.oxfordjournals.org/cgi/content/full/36/suppl_2/W119. 183, 201
- Thompson, W. and Rouchka, E. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic acids ...*, 2003. URL <http://nar.oxfordjournals.org/cgi/content/abstract/31/13/3580>. 27, 68
- Tice-Baldwin, K., Fink, G.R., and Arndt, K.T. BAS1 has a Myb motif and activates HIS4 transcription only in combination with BAS2. *Science*, 246(4932):931–5, 1989. 167
- Tompa, M., Li, N., Bailey, T.L., Church, G.M., et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–144, 2005. doi:10.1038/nbt1053. URL <http://dx.doi.org/10.1038/nbt1053>. 13, 16, 102, 103, 104, 105, 107, 108, 111, 112, 153, 172, 174
- Valen, E., Pascarella, G., Chalk, A., Maeda, N., et al. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res*, 19(2):255–265, 2008. doi:10.1101/gr.084541.108. 5
- Valen, E., Sandelin, A., Winther, O., and Krogh, A. Discovery of regulatory elements is improved by a discriminatory approach. *PLoS Comput Biol*, 5(11):e1000562, 2009. doi:10.1371/journal.pcbi.1000562. URL [http:](http://)

REFERENCES

- [//www.ploscompbiol.org/article/info252Fjournal.pcbi.1000562](http://www.ploscompbiol.org/article/info252Fjournal.pcbi.1000562). 111, 113
- van Dongen, S., Abreu-Goodger, C., and Enright, A.J. Detecting microRNA binding and siRNA off-target effects from expression data. *Nat Methods*, 5(12):1023–5, 2008. doi:10.1038/nmeth.1267. URL <http://www.nature.com/nmeth/journal/v5/n12/abs/nmeth.1267.html>. 13
- van Nimwegen, E. Scaling laws in the functional content of genomes. *Trends Genet*, 19(9):479–84, 2003. 7
- Vegetti, S. and Koopmans, L. Bayesian strong gravitational-lens modelling on adaptive grids: objective detection of mass substructure in Galaxies. *Monthly Notices of the Royal Astronomical ...*, 2009. URL <http://arxiv.org/pdf/0805.0201>. 25
- Venters, B.J. and Pugh, B.F. A canonical promoter organization of the transcription machinery and its regulators in the *Saccharomyces* genome. *Genome Res*, 19(3):360–371, 2008. doi:10.1101/gr.084970.108. 2, 3, 123
- Vilo, J. Discovering Frequent Patterns from Strings. *egeeninc.com*, 1998. URL <http://www.egeeninc.com/u/vilo/Publications/CS-Report-1998-9.ps>. 101
- Vilo, J., Brazma, A., Jonassen, I., Robinson, A., et al. Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc Int Conf Intell Syst Mol Biol*, 8:384–94, 2000. 101
- Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet*, 40(2):158–60, 2008. doi:10.1038/ng.2007.55. 2
- Visel, A., Rubin, E.M., and Pennacchio, L.A. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205, 2009. doi:10.1038/nature08451. 2, 4
- von Neumann, J. Various techniques used in connection with random digits. Monte Carlo methods. *National Bureau of Standards AMS*, 12:36–38, 1951. 23

REFERENCES

- Walter, J., Dever, C., and Biggin, M. Two homeo domain proteins bind with similar specificity to a wide range of DNA sites in *Drosophila* embryos. *Genes & Development*, 1994. URL <http://genesdev.cshlp.org/content/8/14/1678.short>. 5, 8
- Wang, G. and Zhang, W. A steganalysis-based approach to comprehensive identification and characterization of functional regulatory elements. *Genome Biol*, 7(6):R49, 2006. doi:10.1186/gb-2006-7-6-r49. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Abstract&list_uids=16787547. 108
- Wang, T. and Stormo, G.D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–2380, 2003. 11, 16
- Warren, C.L., Kratochvil, N.C.S., Hauschild, K.E., Foister, S., et al. Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci USA*, 103(4):867–72, 2006. doi:10.1073/pnas.0509843102. URL <http://www.pnas.org/content/103/4/867.long>. 12, 114
- Wijaya, E., Yiu, S., Son, N., Kanagasabai, R., et al. MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders. *Bioinformatics*, 2008. doi:10.1093/bioinformatics/btn420. 108
- Wilson, D., Charoensawan, V., Kummerfeld, S.K., and Teichmann, S.A. DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Research*, 36(Database issue):D88–92, 2008a. doi: 10.1093/nar/gkm964. 6, 7, 8, 9, 10, 35, 136, 155
- Wilson, D., Pethica, R., Zhou, Y., Talbot, C., et al. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*, 37(Database issue):D380–6, 2009. doi:10.1093/nar/gkn762. URL http://nar.oxfordjournals.org/cgi/content/full/37/suppl_1/D380?view=long&pmid=19036790. 6, 19
- Wilson, R.J., Goodman, J.L., Strelets, V.B., and Consortium, F. FlyBase: integration and improvements to query tools. *Nucleic*

REFERENCES

- Acids Research*, 36(Database issue):D588–93, 2008b. doi:10.1093/nar/gkm930. URL http://nar.oxfordjournals.org/cgi/content/full/36/suppl_1/D588?view=long&pmid=18160408. 6
- Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics*, 9(4):326–332, 2008. doi:10.1093/bib/bbn016. 34
- Wingender, E., Chen, X., Fricke, E., Geffers, R., et al. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, 29(1):281–283, 2001. 33
- Wolfe, S.A., Neklodova, L., and Pabo, C.O. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct*, 29:183–212, 2000. doi:10.1146/annurev.biophys.29.1.183. URL <http://dx.doi.org/10.1146/annurev.biophys.29.1.183>. 14, 39
- Workman, C. and Stormo, G. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*, 5:464–475, 2000. 68, 77, 103, 113
- Wunderlich, Z. and Mirny, L. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics*, 25(10):434–440, 2009. 8
- Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005. doi:10.1038/nature03441. URL <http://dx.doi.org/10.1038/nature03441>. 13, 102, 111, 112
- Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., et al. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci USA*, 104(17):7145–50, 2007. doi:10.1073/pnas.0701811104. URL <http://www.pnas.org/cgi/content/full/104/17/7145>. 13, 102
- Xie, X., Rigor, P., and Baldi, P. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics*, 25(2):167–74,

REFERENCES

2009. doi:10.1093/bioinformatics/btn605. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/25/2/167>. 5
- Xing, E., Jordan, M., Karp, R., and Russell, S. A hierarchical Bayesian Markovian model for motifs in biopolymer sequences. *Advances in Neural Information Processing Systems*, pp. 1513–1520, 2003a. 68
- Xing, E.P. and Karp, R.M. MotifPrototyper: a Bayesian profile model for motif families. *Proc Natl Acad Sci U S A*, 101(29):10523–10528, 2004. doi:10.1073/pnas.0403564101. URL <http://dx.doi.org/10.1073/pnas.0403564101>. 34, 39, 42, 67, 68, 81, 82, 83, 84, 85, 86, 87, 91
- Xing, E.P., Wu, W., Jordan, M.I., and Karp, R.M. LOGOS: a modular Bayesian model for de novo motif detection. *Proc IEEE Comput Soc Bioinform Conf*, 2:266–76, 2003b. 68
- yong Li, X., MacArthur, S., Bourgon, R., Nix, D., et al. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol*, 6(2):e27, 2008. doi:10.1371/journal.pbio.0060027. URL <http://www.plosbiology.org/article/info52Fjournal.pbio.0060027>. 8
- Zare-Mirakabad, F., Ahrabian, H., Sadeghi, M., Nowzari-Dalini, A., et al. New scoring schema for finding motifs in DNA Sequences. *BMC Bioinformatics*, 10(1):93, 2009. doi:10.1186/1471-2105-10-93. 108
- Zeitlinger, J., Simon, I., Harbison, C., and Hannett, N. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell*, 2003. URL <http://linkinghub.elsevier.com/retrieve/pii/S0092867403003015>. 144
- Zhang, J., Jiang, B., Li, M., Tromp, J., et al. Computing exact P-values for DNA motifs. *Bioinformatics*, 23(5):531, 2007. 117
- Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9):R137, 2008a. doi:10.1186/gb-2008-9-9-r137. 185

REFERENCES

- Zhang, Z.D., Rozowsky, J., Snyder, M., Chang, J., et al. Modeling ChIP Sequencing In Silico with Applications. *PLoS Comput Biol*, 4(8):e1000158, 2008b. doi:10.1371/journal.pcbi.1000158. [189](#)
- Zheng, W., Zhao, H., Mancera, E., Steinmetz, L.M., et al. Genetic analysis of variation in transcription factor binding in yeast. *Nature*, 464(7292):1187–91, 2010. doi:10.1038/nature08934. [112](#)
- Zhu, C., Byers, K.J., Mccord, R.P., Shi, Z., et al. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res*, 19(4):556–566, 2009. doi:10.1101/gr.090233.108. [XII](#), [12](#), [35](#), [36](#), [99](#), [114](#), [115](#), [128](#), [129](#), [130](#), [131](#), [132](#), [134](#), [139](#), [140](#)
- Zhu, J. and Zhang, M.Q. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, 15(7-8):607–611, 1999. [101](#), [114](#)
- Zhu, Q. and Halfon, M.S. Complex organizational structure of the genome revealed by genome-wide analysis of single and alternative promoters in *Drosophila melanogaster*. *BMC Genomics*, 10(1):9, 2009. doi:10.1186/1471-2164-10-9. [3](#)