CS 260-1: "Pattern Discovery in Biosequences"
Reading list

Stefano Lonardi (stelo@cs.ucr.edu)

January 16, 2003

General References

Books

• Richard Durbin, A. Krogh, G. Mitchison, and S. Eddy, *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids,* Cambridge University Press, 1999.

• Benjamin Lewin, "Genes VII", Oxford University Press, 2000.

• Feller, W. An introduction to Probability Theory and its Applications, Wiley, New York, 1968

• Developing Bioinformatics Computer Skills by Cynthia Gibas, Per Jambeck; O'Reilly, 2001.

• Dan Gusfield, *Algorithms on Strings, Trees and Sequences - Computer Science and Computational Biology,* Cambridge University Press, 1997.

• Pavel A. Pevzner, *Computational Molecular Biology: An Algorithmic Approach,* MIT Press, 2000.

• Joao Setubal and Joao Carlos Meidanis, *Introduction to Computational Molecular Biology,* PWS Publishing Co., 1997.

• Jason Wang, Bruce A. Shapiro, and Dennis Shasha, *Pattern Discovery in Biomolecular Data Tools, Techniques, and Applications,* Oxford University Press, 1999.

• Pierre Baldi and Soren Brunak, *Bioinformatics: the Machine Learning Approach,* MIT Press, 2nd edition, 2001.

Papers

• A. Brāzma and I. Jonassen and I. Eidhammer and D. Gilbert, "Approaches to the automatic Discovery of patterns in Biosequences", J. Comp. Biology, 5(2), 277–304, 1998.

• A.Apostolico, M.E.Bock, S.Lonardi, X.Xu, "Efficient Detection of Unusual Words", Journal of Computational Biology, vol.7, no.1/2, pp.71-94, 2000 (available on the class website)

• Gesine Reinert, Sophie Schbath, Michael S. Waterman, "Probabilistic and Statistical Properties of Words: An Overview", Journal of Computational Biology, vol.7, no.1/2, 2000 (available on the class website)

• Brona Brejova, Chrysanne DiMarco, Tomas Vinar, Sandra Romero Hidalgo, Gina Holguin, Cheryl Patten. "Finding Patterns in Biological Sequences". Unpublished TR. University of Waterloo, 2000 (available on the class website)

• A.Krogh, "An introduction to hidden Markov models for biological sequences" In S. L. Salzberg, D. B. Searls, and S. Kasif, editors, Computational Methods in Molecular Biology, chapter 4, pages 45-63. Elsevier, Amsterdam, 1998 (available on the class website)

1 Motifs discovery

• Emotif `http://motif.stanford.edu/emotif/`

• Verbumculus `http://www.cs.ucr.edu/~stelo/Verbumculus/`

• Pratt `http://www.ii.uib.no/~inge/Pratt.html`

• Chapter 4 of Jason Wang, Bruce A. Shapiro, and Dennis Shasha, *Pattern Discovery in Biomolecular Data Tools, Techniques, and Applications,* Oxford University Press, 1999.

• Giulio Pavesi, Giancarlo Mauri, and Graziano Pesole, "An algorithm for finding signals of unknown length in DNA sequences", Bioinformatics 17: 207-214, 2001.

• Mathieu Blanchette and Saurabh Sinha, "Separating real motifs from their artifacts", Bioinformatics 17: S30-S38, 2001.

• Pavel A. Pevzner and Sing-Hoi Sze, "Combinatorial Approaches to Finding Subtle Signals in DNA Sequences", Proc. of the International Conference on Intelligent Systems for Molecular Biology, 269–278, 2000.

• Martin Tompa and Jeremy Buhler, "Finding Motifs Using Random Projections", Proc. RECOMB, 67-74, 2001.

• Mathieu Blanchette and Benno Schwikowski and Martin Tompa, "An Exact Algorithm to Identify Motifs in Orthologous Sequences from Multiple Species", Proc Intelligent Systems for Molecular Biology, 37–45, 2000.

• A. Apostolico and M. E. Bock and S. Lonardi and X. Xu, "Efficient Detection of Unusual Words", J. Comp. Biology, 7(1/2), 71–94, 2000.

• Yuh-Jyh Hu and Suzanne Sandmeyer and Calvin McLaughlin and Dennis Kibler, "Combinatorial motif analysis and hypothesis generation on a genomic scale", Bioinformatics, 16(3), 222–232, 2000.

• Andrea Califano, "SPLASH: Structural Pattern Localization Analysis by Sequential Histograming", Bioinformatics 15, 341–357, 2000.

• Laxmi Parida and Yuan Gao and Dan Platt and Aris Floratos and Isidore Rigoutsos, "Pattern Discovery on Character Sets and Real-valued Data: Linear Bound on Irredundant Motifs and an Efficient Polynomial Time Algorithm", SODA, 297–308, 2000.

• Isidore Rigoutsos and Aris Floratos, "Combinatorial pattern discovery in biological sequences: The Teiresias algorithm", Bioinformatics, 14(1), 55–67, 1998.

• Laxmi Parida and Isidore Rigoutsos and Dan Platt, "An Output-Sensitive Flexible Pattern Discovery Algorithm", Combinatorial Pattern Matching, LNCS 2089, 131–142, 2001.

• I. Rigoutsos and A. Floratos and L. Parida and Y. Gao and D. E. Platt, "The Emergence of Pattern Discovery Techniques in Computational Biology", Metabolic Engineering, 159-177, 2(3), 2000.

• Yada, T. and Totoki, Y. and Ishikawa, M. and Asai, K. and Nakai, K.,"Automatic extraction of motifs represented in the hidden Markov model from a number of DNA sequences", Bioinformatics 14, 317–325, 1998.

• Leung, M. Y. and Marsh, G. M. and Speed, T. P.,"Over and underrepresentation of short DNA words in Herpesvirus genomes", J. Comp. Biology, 3, 345–360, 1996.

• A. Brāzma and I. Jonassen and I. Eidhammer and D. Gilbert, "Approaches to the automatic Discovery of patterns in Biosequences", J. Comp. Biology, 5(2), 277–304, 1998.

• J. Hudak and M.A. McClure, "A Comparative Analysis of Computational Motif-Detection Methods", Pac. Symp. Biocomputing, 138-149, 1999.

1.1 Motif discovery by Gibbs sampling

• Gibbs sampler `http://bayesweb.wadsworth.org/gibbs/gibbs.html`

• Jun S. Liu and Andrew F. Neuwald and Charles E. Lawrence, "Bayesian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies", Journal of the American Statistical Association", 90(432), 1156–1170, 1995.

• A.F. Neuwald and J.S. Liu and C.E. Lawrence, "Gibbs motif sampling: Detecting bacterial outer membrane protein repeats", Protein Science, 4, 1618-1632, 1995.

• C. E. Lawrence and S. F. Altschul and M. S. Boguski and J. S. Liu and A. F. Neuwald and J. C. Wootton, "Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment", Science, 262, 208–214, 1993.

• G Thijs, K Marchal, M Lescot, S Rombauts, B de Moor, P Rouze, Y Moreau, "A Gibbs sampling method to detect over-represented motifs in the upstream regions of co-expressed genes", in Proceedings of the Fifth Annual International Conference on Computational Biology, 305-312 (ACM Press), 2001.

• Liu, J.S. and Lawrence, C.E., "Bayesian Inference on biopolymer models", Bioinformatics 15, 38-52, 1999.

1.2 Motif discovery by expectation maximization

• MEME `http://meme.sdsc.edu/meme/website/`

• Meta MEME `http://metameme.sdsc.edu/`

• Chapter 3 of Jason Wang, Bruce A. Shapiro, and Dennis Shasha, *Pattern Discovery in Biomolecular Data Tools, Techniques, and Applications,* Oxford University Press, 1999.

• Timothy L. Bailey and Michael Gribskov, "The Megaprior Heuristic for Discovering Protein Sequence Patterns", Proc. of the Fourth International Conference on Intelligent Systems for Molecular Biology, 15–24, 1996.

• Timothy L. Bailey and Charles Elkan, "Unsupervised learning of multiple motifs in biopolymers using expectation maximization", Machine Learning, 21 (1/2), 51–80, 1995.

2 DNA segmentation

• Vsevolod Makeev and Vasily Ramensky and Mikhail Gelfand and Mikhail Roytberg and Vladimir Tumanyan, "Bayesian Approach to DNA Segmentation into Regions with Different Average Nucleotide Composition", First International Conference on Biology, Informatics, and Mathematics, Lecture notes in CS 2066, 57-73, 2001.

• Bernaola-Galvan, I Grosse, P Carpena, JL Oliver, R Roman-Roldan, HE Stanley, "Finding borders between coding and noncoding DNA regions by an entropic segmentation method", Physical Review Letters 85(6):1342-1345, 2000.

• G. Stormo and D. Haussler, "Optimally parsing a sequence into different classes based on multiple types of evidence", Proc Int Conf Intell Syst Mol Biol, 369-75, 1994.

• Harmen J. Bussemaker and Hao Li and Eric D. Siggia, "Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis", Proc Natl Academy Science 97, "10096–10100", 2000.

• Harmen J. Bussemaker and Hao Li and Eric D. Siggia, "Regulatory Element Detection Using a Probabilistic Segmentation Model", Proc Eighth International Conference on Intelligent Systems for Molecular Biology, 344–354, 2000.

3 Finding genes

• Glimmer (TIGR) `http://www.tigr.org/softlab/glimmer/glimmer.html`

• GenScan `http://genes.mit.edu/GENSCAN.html`

• HMMGene `http://www.cbs.dtu.dk/services/HMMgene/`

• Grail `http://compbio.ornl.gov/Grail-1.3/`

• GenMark.hmm `http://dixie.biology.gatech.edu/GeneMark/gmhmm2_prok.cgi`

• VEIL `http://www.tigr.org/~salzberg/veil.html`

• John S. Chuang and Dan Roth, "Gene recognition based on DAG shortest paths", Bioinformatics 2001 17: S56-S64

• Roderic Guig, Pankaj Agarwal, Josep F. Abril, Moiss Burset, and James W. Fickett, "An Assessment of Gene Prediction Accuracy in Large DNA Sequences", Genome Res. 2000 10: 1631-1642.

• Stormo GD, "Gene-finding approaches for eukaryotes", Genome Research, 10(4):394-7, 2000.

• Burge, C. and Karlin, S., "Prediction of complete gene structures in human genomic DNA". J. Mol. Biol. 268, 78-94, 1997.

• Burge, C. B. and Karlin, S., "Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 8, 346-354, 1998.

• V Bafna, DH Huson, "The conserved exon method for gene finding", ISMB, 2000.

• A Krogh, "Using database matches with HMMGene for automated gene detection in Drosophila", Genome Research, 10(4):523-528, 2000.

• M Pertea, SL Salzberg, MJ Gardner, "Finding genes in Plasmodium falciparum chromosome 3", Nature, 404:34, 2000.

• A Salamov, VV Solovyev, "Ab initio gene finding in Drosophila genomic DNA", Genome Research, 10(4);516-522, 2000.

• Lukashin A. and Borodovsky M., "GeneMark.hmm: new solutions for gene finding", Nucleic Acid Research, Vol. 26, No. 4, pp. 1107-1115, 1998.

4 Finding regulatory elements

• PromoterInspector `http://genomatix.gsf.de/cgi-bin/promoterinspector/promoterinspector.pl`

• Promoter `http://www.cbs.dtu.dk/services/promoter/`

• Matthias Scherf, Andreas Klingenhoff, Thomas Werner, "Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach", Journal of Molecular Biology, 297:599-606, 2000.

• Michael B Eisen, Derek Y Chiang, Patrick O Brown, "Discovery of regulatory elements from microarray data using genome mean expression profiles", ISMB 2001.

• Sridhar Hannenhalli and Samuel Levy, "Promoter prediction in the human genome", Bioinformatics 17: 90-96, 2001.

• Uwe Ohler, Heinrich Niemann, Guo-chun Liao, and Gerald M. Rubin, "Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition", Bioinformatics 17: 199-206, 2001.

• Uwe Ohler, Heinrich Niemann, "Identification and analysis of eukaryotic promoters: recent computational approaches", Trends in Genetics, 17(2):56-60, 2001.

• Uwe Ohler, "Promoter prediction on a genomic scale - the Adh experience", Genome Research, 10(4):539-542, 2000.

• MQ Zhang, "Computational methods for promoter recognition", Chapter 10 in Current Topics in Computational Molecular Biology, eds. T Jiang, Y Xu, MQ Zhang, 2001.

• RC Hardison, "Conserved noncoding sequences are reliable guides to regulatory elements", Trends in Genetics, 16(9):369-372, 2000.

• Yoseph Barash, Gill Bejerano, Nir Friedman, "A Simple Hyper-Geometric Approach for Discovering Putative Transcription Factor Binding Sites", First International Workshop on Algorithms in Bioinformatics, Lecture notes CS 2149, 2001.

• WW Wasserman, M Palumbo, W Thompson, JW Fickett, CE Lawrence, "Human-mouse genome comparisons to locate regulatory sites", Nature Genetics, 26:225-227, 2000.

• Saurabh Sinha and Martin Tompa, "A Statistical Method for Finding Transcription Factor Binding Sites", Eighth International Conference on Intelligent Systems for Molecular Biology, 344-354, 2000.

• J. W. Fickett and W. W. Wasserman, "Discovery and modeling of transcriptional regulatory regions", Curr. Opin. Biotechnol., 11(1), 19-24, 2000.

• Harmen J. Bussemaker and Hao Li and Eric D. Siggia, "Regulatory element detection using correlation with expression", Nature Genetics, 27(2), 167-171, 2001.

• M. S. Gelfand and E. V. Koonin and A. A. Mironov, "Prediction of transcription regulatory sites in Archaea by a comparative genomic approach", Nucleic Acid Research, 28(3), 695–705, 2000.

• A. Brāzma and I. Jonassen and E. Ukkonen and J. Vilo, "Predicting Gene Regulatory Elements in Silico on a Genomic Scale", Genome Research, 8(11), 1202–1215, 1998.

• A. Brāzma and J. Vilo and E. Ukkonen and K. Valtonen, "Data Mining for Regulatory Elements in Yeast Genome", Proc. of the 5th International Conference on Intelligent Systems for Molecular Biology", 65–74, 1997.

• Jaak Vilo and A. Brāzma and Inge Jonassen and Alan Robinson and Esko Ukkonen", "Mining for Putative Regulatory Elements in the Yeast Genome using Gene Expression Data", Proc. of the International Conference on Intelligent Systems for Molecular Biology, "384–394", 2000.

• J. van Helden and B. André and J. Collado-Vides, "Extracting regulatory sites from the upstream region of the yeast genes by computational analysis of oligonucleotides", J. Mol. Biol., 281, 827–842, 1998,

• J. van Helden and Alma. F. Rios and Julio Collado-Vides, "Discovering regulatory elements in non-coding sequences by analysis of spaced dyads", Nucl. Acid Research, 28(8), 1808-1818, 2000.

5 Finding splicing sites

• Splice Predition using ANN `http://www.fruitfly.org/seq_tools/splice.html`

• ISIS: `http://isis.bit.uq.edu.au/a_splicers.html`

• Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. Mar 1;29(5):1185-90, 2001.

• Burge, C. B., Modeling dependencies in pre-mRNA splicing signals. In Salzberg, S., Searls, D. and Kasif, S., eds. Computational Methods in Molecular Biology, Elsevier Science, Amsterdam, pp. 127-163, 1998.

• Michael Bruno, Mikhail S Gelfand, Sylvia Spengler, Manfred Zorn, Inna Dubchak, and John G. Conboy, "Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing", Nucleic Acids Research, 29:2338-2348, 2001.

• Brenton R Graveley, "Alternative splicing: increasing diversity in the proteomic world", Trends in Genetics, 17(2):100-107, 2001.

• TA Thanaraj, F Clark, "GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions", Nucleic Acids Research, 29(12):2581-2593, 2001.

• D Cai, A Delcher, B Kao, S Kasif, "Modeling splice sites with Bayes networks", Bioinformatics, 16(2):152-158, 2000.

• Larry Croft, Soeren Schandorff, Francis Clark, Kevin Burrage, Peter Arctander, John S Mattick, "ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome", Nature Genetics, 24:340-341, 2000.

6 Statistical analysis of gene expression data

• `http://linkage.rockefeller.edu/wli/microarray/` is an excellent starting point for resources on micro array literature, groups, companies, etc.

• P Baldi, AD Long (2001), "A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes", Bioinformatics, 17:509-519.

• V Filkov, S Skiena, J Zhi (2001), "Analysis techniques for microarray time-series data", in RECOMB 2001: Proceedings of the Fifth Annual International Conference on Computational Biology , pp. (ACM Press).

• Trevor Hastie, Robert Tibshirani, David Botstein, Patrick Brown (2001), "Supervised harvesting of expression trees", Genome Biology, 2(1): research0003.1-0003.12.

• Cheng Li, Wing Hung Wong (2001), "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection", Proceedings of the National Academy of Sciences, 98:31-36.

• Jason C Mills, Jeffrey I Gordon (2001), "A new approach for filtering noise from high-density oligonucleotide microarray datasets", Nucleic Acids Research, 29(15):e72.

• Jeffrey G Thomas, James M Olson, Stephen J Tapscott, Lue Ping Zhao (2001), "An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles", Genome Research, 11:1227-1236.

• O Troyanskaya, M Cantor, G Sherlock, P Brown, T Hastie, R Tibshirani, D Botstein, R B Altman (2001), "Missing value estimation methods for DNA microarrays", Bioinformatics, 17(6):520-525.

6.1 Clustering

• John Aach, George M. Church, "Aligning gene expression time series with time warping algorithms ", Bioinformatics, 17:495-508, 2001.

• A. Ben-Dor, N. Friedman, Z. Yakhini (2001), "Class discovery in gene expression data", in RECOMB 2001: Proceedings of the Fifth Annual International Conference on Computational Biology , pp.31-38 (ACM Press), 2001.

• Y Barash, N Friedman (2001), "Context-specific Bayesian clustering for gene expression data", in RECOMB 2001: Proceedings of the Fifth Annual International Conference on Computational Biology , pp.12-20 (ACM Press).

• M Kathleen Kerr, Gary A Churchill, "Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments", Proceedings of the National Academy of Sciences, 98:8961-8965, 2001.

• AV Lukashin, R Fuchs (2001), "Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters", Bioinformatics, 17:405-414.

• Michael E Wall, Patricia A Dyck, Thomas S Brettin (2001), "SVDMAN - singular value decomposition analysis of microarray data", Bioinformatics, 17:566-568.

• KY Yeung, DR Haynor, WL Ruzzo (2001), "Validating clustering for gene expression data", Bioinformatics, 17:309-318.

• Rob M Ewing, J Michael Cherry (2001), "Visualization of expression clusters using Sammon's non-linear mapping", Bioinformatics, 17:658-659.

7 Protein classification

• SCOP `http://scop.mrc-lmb.cam.ac.uk/scop/`

• CATH `http://www.biochem.ucl.ac.uk/bsm/cath_new/`

• Blocks `http://blocks.fhcrc.org/`

• Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997) CATH- A Hierarchic Classification of Protein Domain Structures. Structure. Vol 5. No 8. p.1093-1108.

• Pearl, F.M.G, Lee, D., Bray, J.E, Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M. and Orengo, C.A. (2000) Assigning genomic sequences to CATH Nucleic Acids Research. Vol 28. No 1. 277-282

• JT Wang, TG Marr, D Shasha, BA Shapiro, and GW Chirn Discovering active motifs in sets of related protein sequences and using them for classification Nucleic Acids Res. 1994 22: 2769-2775.

• Shmuel Pietrokovski, Steven Henikoff, and Jorja G. Henikoff. The BLOCKS database - a system for protein classification. Nucleic Acids Research, 24:197–200, 1996.

• Dorohonceanu, B. and Nevill-Manning, C. G. (2000). Accelerating protein classification using suffix trees. In Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB), pages 128–133.

8 Finding MARs

• MarsFinder `http://www.futuresoft.org/MAR-Wiz/`

• T. Boulikas, "Chromatin domains and prediction of MAR sequences", Int. Rev. Cytol., 162A, 279–388, 1995.

• S.M. Gasser, "Nuclear scaffold and high-order folding of eukaryotic DNA", in *Architecture of Eukaryotic Genes*, 461–471, "VCH Verlagsgeselschaft, 1988, G. Kahl ed.

• B. Amati and S.M. Gasser, "Drosophila scaffold-attached regions bind nuclear scaffolds and can function as MARS elements in both budding and fission yeast", Mol. Cell. Biol., 10, 5442–5454, 1990.

9 Compression & sequence analysis

• Chapter 1 of Jason Wang, Bruce A. Shapiro, and Dennis Shasha, *Pattern Discovery in Biomolecular Data Tools, Techniques, and Applications,* Oxford University Press, 1999.

• L. Allison and T. Edgoose and T. I. Dix, "Compression of strings with approximate repeats", Proc. Intell. Sys. in Mol. Biol., 8–16, 1998.

• D.R. Powell and D.L. Dowe and L. Allison and T.I. Dix, "Discovering Simple DNA Sequences by Compression", Pacif. Symp. Biocomputing, 595-606, 1998.

• O. Delgrange and M. Dauchet and E. Rivals, Location of Repetitive Regions in Sequences By Optimizing A Compression Method, Pacif. Symp. Biocomputing, 254-265, 1999.

• D. M. Loewenstern and H. M. Berman and H. Hirsch, Maximum a posteriori classification of DNA structure from sequence information", Pacif. Symp. Biocomputing, 1998.

• S. Grumbach and F. Tahi, A new challenge for compression algorithms: Genetic sequences, Inf. Proc. and Mngm., 30(6), 875–886, 1994.

10 Mining PubMed

• T -K Jenssen, A Lgreid, J Komorowski and E Hovig, A literature network of human genes for high-throughput analysis of gene expression, Nature Genetics, v28, no 1, 21–28, May 2001.

• Methods for Large-Scale Mining of Networks of Human Genes Tor-Kristian Jenssen, Lisa M.J. berg, Magnus L. Anderson, and Jan Komorowski, Proc. SIAM International Conference on Data Mining, 2001.