Chapter 5

Conclusions and Perspectives

Characterizing the binding sites of transcription factors is key to understanding the mechanisms that mediate transcriptional regulation. Given the short and degenerate nature of binding sites and the huge background genomic noise, this remains a largely daunting task. In the post-genomic era, cross-species comparisons to limit the search to conserved regions provide an elegant remedy. Here researchers seek binding sites which are conserved across different species, with the hypothesis that they are more likely to be functional. Current computational methods in the field segregate the two axes of information – conservation and binding site annotation. In this multi-step setting, low sequence similarity or poor profile quality poses problems. The reliance on pre-determined optimal alignments makes the approaches susceptible to gaps in the aligned sites. Barring a few examples, most methods additionally disregard the evolutionary characteristics of binding sites.

In this work, we addressed the above issues by introducing a novel integrated approach SimAnn to detect conserved transcription factor binding sites (TFBSs). In SimAnn, the alignment and annotation steps are combined in one extended alignment model. This enables a local rearrangement of gaps in the alignment to make the conserved hits stand out more clearly. The output is an *annotated alignment*, that is an alignment with parts annotated as putative conserved TFBSs.

Clearly, in such an integrated setting the choice of parameters needed in the model is crucial. We presented a statistical framework for parameter selection based on desired type I and type II error constraints. We also demonstrated how position-specific evolutionary characteristics of a TFBS can be taken into account, leading to the extension eSimAnn. The predictions now are in real terms "conserved" binding sites – perfectly aligned binding sites that share a common evolutionary history.

We highlighted the applicability of SimAnn via a systematic comparison with other multi-step approaches on simulated data. We showed how SimAnn can predict perfectly aligned conserved hit pairs even in conditions of higher evolutionary distance or poorer profile quality. On similar lines, we compared eSimAnn against SimAnn in a controlled setting with evolutionarily related artificial binding sites. By analyzing the well-known even-skipped stripe 2 enhancer region in two Drosophila species we illustrated the potential of SimAnn in a biological setting. Finally, using a larger testset of human-mouse sequences, we demonstrated how the performance of eSimAnn compares well against the current state-of-the-art multi-step method which explicitly considers evolution of binding sites.

The simultaneous framework with a formal theoretical motivation for parameter choice has certain advantages. Usually, sufficiently large datasets of experimentally verified sites are

unavailable for most transcription factors. Those that are available are commonly biased to the particular setting of an application, for example the conservation requirements, quality of profiles, positional preferences with respect to each other or to the transcription start site, etc. The prescription for parameter selection underlying SimAnn (and eSimAnn) provides a means to handle any profile of interest. This prevents the compilation of large testsets or dealing with biases introduced during training. Furthermore, since the predictions are devoid of gaps, it precludes the necessity of manual tweaking of the aligned regions or adopting heuristic measures to limit gaps in the alignment, as is usually done.

From a pair Hidden Markov Model perspective, the approach implies considering additional profile states which emit pairs of motifs. We discussed how in the pairHMM formulation of annotated alignments, the transition probabilities between the background alignment states and the pair-profile states can be related to the cutoff in a log-likelihood ratio test. Pair Hidden Markov Models have been successfully employed in gene prediction; the application in predicting conserved binding sites and the strategy for parameter choice is a novel contribution of this thesis.

At present, the algorithm does not allow gaps while defining conserved binding sites. An imaginable step forward is to allow gaps in the aligned binding site regions. In other words, to consider a pairHMM for the binding site motifs *inside* the pairHMM for the whole sequences. Besides allowing a possibility to consider evolutionarily motivated gaps in binding sites, it may particularly enable one to search for binding sites of factors like dimers or those which consist of a core of highly ambiguous positions surrounded by specific ends, as for example in GAL4. Clearly, the task is complicated due to the increased time and space complexity, as well as the choice of parameters regarding the transitions inside the motif pair.

The proposed approach is an optimal alignment algorithm, making it infeasible for the analysis of large sequences. A possible improvement can be to build upon existing time and space efficient strategies available for standard alignments. Either more efficient optimal alignment algorithms (eg. [86]) can be considered or heuristic approaches can be adapted for the purpose. For the latter, considering a two-phase algorithm – scan individual sequences for putative TFBSs and then use them to generate the alignment – can be an option. The tool SITEBLAST [131], which works along similar lines, could be explored for its applicability as a heuristic to SimAnn.

Using putative binding sites on individual sequences as seeds for generating alignments falls in the broader field of what is usually referred to as *anchor-based* alignments. Here methods usually adopt a three-step approach to generate an alignment: identify similar substrings in the sequences (*fragments*), compute an optimal chain of colinear non-overlapping fragments and finally use this set of anchors to generate an alignment that covers intermediate regions. Sophisticated time-efficient approaches in this direction have been proposed that enable large genome-scale alignments, both pairwise as well as multiple [141, 33, 28, 2]. Exploiting these strategies could provide a way for extending the annotated alignment approach to take advantage of the immense sequence data available. A possibility could be to allow for conserved binding sites using a modified scoring scheme in the chaining step (that is, the second step). Let us conjecture how.

Usually, a fragment f is composed of a pre-computed multi-alignment of its substrings and has an associated *weight* given by the corresponding alignment score ($algmt\_score$).

Given a gap cost function $gap\_cost(f', f)$, which penalizes the jump from fragment f to f', dynamic programming can be employed to calculate the highest-scoring chain of colinear non-overlapping fragments. The optimal score of all chains ending at f' is thus given by:

$$f'.score = f'.weight + \max\{f.score - gap\_cost(f', f) : f \text{ lies before } f'\}$$

where $gap\_cost(f', f)$ depends on the length of the intermediate region and $f'.weight = algmt\_score$. Sophisticated time-efficient strategies have been proposed for this chaining step [141, 146] and hence could be exploited for our purposes. The idea, in our context, would be to allow for the possibility that a fragment is composed of putative TFBSs of a profile P. Hence, the recursion rule would be modified as following:

$$\begin{aligned} f'.score = \max\{f'.weight|_{f'\sim \mathrm{P}}, f'.weight|_{f'\sim A}\} + \\ \max\{f.score - gap\_cost(f', f) : f \text{ lies before } f'\} \end{aligned}$$

where $f'.weight|_{f'\sim \mathrm{P}}$ is the weight of a fragment under the profile and is given by $\sum \texttt{PSSM}(s_i) - pen$, $s_i \in f'$. The profile penalty pen is calculated as in the annotated alignment approach, thus reflecting the probability that the fragment is composed of conserved binding sites as opposed to being a standard alignment. The quantity $f'.weight|_{f'\sim A}$ is simply the standard alignment score of the respective substrings of f'. Intuitively, this would imply that a fragment would be annotated as a conserved TFBS hit if its profile score exceeds its alignment score. Note that the profile penalty depends only on the substrings constituting a fragment f'. The gap cost function, on the other hand, depends on the intermediate region between f and f' and hence remains same irrespective of whether $f' \sim \mathrm{P}$ or $f' \sim A$. While this may provide an efficient possibility to extend to large sequences or multiple alignments, the influence on the time complexity due to the additional comparisons for each new profile would need a more in-depth investigation.

Equally interesting but unexplored in this thesis are other open directions. Recently, Kececioglu and Kim [96] presented a linear-programming based inverse alignment approach to estimate parameters that make a given set of correct alignments optimal-scoring. In our context, this implies finding a set of "correct" alignments of cis-regulatory sequences with known transcription factor binding sites, a non-trivial task. The desired parameter set would then include profile-related parameters, besides the standard substitution scores and gap-penalties. How do the resulting profile-related parameters compare against those estimated through the proposed statistical framework?

As already mentioned in Section 4.5, using annotated alignments to study the evolutionary gain or loss of binding sites is another foreseeable possibility. Considering the alignments of binding sites over increasingly distant species may shed light on their evolutionary properties. Similarly, studying cis-regulatory modules and the competition between factors with similar binding sites would also be an interesting direction of pursuit.

Although faced with numerous challenges, the field of comparative genomics for binding site identification is gathering momentum with large scale data and sophisticated approaches. Combining the experimental knowledge with computational techniques that ease the analysis and use of such data is indispensable for further advances. The integrated approach introduced in this thesis provides a holistic view to such analyses.

Bibliography

[1] `http://web.wi.mit.edu/young/location/`.

[2] M. I. Abouelhoda and E. Ohlebusch. Chainer: Software for comparing genomes. In *Proceedings of the 12th International Conference on Intelligent Systems for Molecular Biology/3rd European Conference on Computational Biology (ISMB/ECCB 2004)*, 2004.

[3] T. Akutsu, H. Arimura, and S. Shimozono. On approximation algorithms for local multiple alignment. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, pages 1–7, New York, USA, 2000. ACM Press.

[4] M. Alexandersson, S. Cawley, and L. Pachter. SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res*, 13(3):496–502, Mar 2003.

[5] S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol*, 219(3):555–565, Jun 1991.

[6] S. F. Altschul and W. Gish. Local alignment statistics. *Methods Enzymol*, 266:460–480, 1996.

[7] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[8] W. Ao, J. Gaudet, W. J. Kent, S. Muttumu, and S. E. Mango. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science*, 305(5691):1743–1746, Sep 2004.

[9] A. Apostolico and R. Giancarlo. Sequence alignment in molecular biology. *J Comput Biol*, 5(2):173–196, 1998.

[10] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology (ISMB 1994)*, Menlo Park, California, USA, 1994. AAAI Press.

[11] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue):W369–W373, Jul 2006.

[12] A. S. Bais, S. Grossmann, and M. Vingron. Simultaneous alignment and annotation of cis-regulatory regions. *Bioinformatics*, 23(2):e44–e49, Jan 2007.

[13] Y. Barash, G. Bejerano, and N. Friedman. A simple hyper-geometric approach for discovering putative transcription factor binding sites. *Lecture Notes in Computer Science*, 2149:278–293, 2001.

[14] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-dna binding sites. In *Proceedings of the 7th International Conf. on Research in Computational Molecular Biology (RECOMB 2003)*, 2003.

[15] S. Batzoglou. The many faces of sequence alignment. *Brief Bioinform*, 6(1):6–22, Mar 2005.

[16] M. Beckstette, D. Strothmann, R. Homann, R. Giegerich, and S. Kurtz. PoSSuMsearch: Fast and sensitive matching of position specific scoring matrices using enhanced suffix arrays. In *Proceedings of the German Conference on Bioinformatics (GCB 2004)*, pages 53–64, Bielefeld, Germany, 2004.

[17] P. V. Benos, M. L. Bulyk, and G. D. Stormo. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*, 30(20):4442–4451, Oct 2002.

[18] E. Berezikov, V. Guryev, R. H. A. Plasterk, and E. Cuppen. CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res*, 14(1):170–178, Jan 2004.

[19] J. Berg, S. Willmann, and M. Lässig. Adaptive evolution of transcription factor binding sites. *BMC Evol Biol*, 4(1):42, Oct 2004.

[20] O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193(4):723–750, Feb 1987.

[21] H. Bigelow, A. Wenick, A. Wong, and O. Hobert. CisOrtho: a program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting. *BMC Bioinformatics*, 5:27, 2004.

[22] M. Blanchette and S. Sinha. Separating real motifs from their artifacts. *Bioinformatics*, 17 Suppl 1:S30–S38, 2001.

[23] E. Blanco, D. Farré, M. M. Albà, X. Messeguer, and R. Guigó. ABS: a database of Annotated regulatory Binding Sites from orthologous promoters. *Nucleic Acids Res*, 34(Database issue):D63–D67, Jan 2006.

[24] E. Blanco, X. Messeguer, T. F. Smith, and R. Guigó. Transcription factor map alignment of promoter regions. *PLoS Comput Biol*, 2(5):e49, May 2006.

[25] K. Blekas, D. I. Fotiadis, and A. Likas. Greedy mixture learning for multiple motif discovery in biological sequences. *Bioinformatics*, 19(5):607–617, Mar 2003.

[26] V. A. Bondarenko, Y. V. Liu, Y. I. Jiang, and V. M. Studitsky. Communication over a large distance: enhancers and insulators. *Biochem Cell Biol*, 81(3):241–251, Jun 2003.

[27] L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–56, Sep 2005.

[28] N. Bray, I. Dubchak, and L. Pachter. AVID: A global alignment program. *Genome Res*, 13(1):97–102, Jan 2003.

[29] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert. Approaches to the automatic discovery of patterns in biosequences. Technical report, Department of Informatics, University of Bergen, 1995.

[30] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert. Approaches to the automatic discovery of patterns in biosequences. *J Comput Biol*, 5(2):279–305, 1998.

[31] A. Brazma, I. Jonassen, J. Vilo, and E. Ukkonen. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res*, 8(11):1202–1215, Nov 1998.

[32] D. G. Brown, M. Li, and B. Ma. A tutorial of recent developments in the seeding of local alignment. *J Bioinform Comput Biol*, 2(4):819–842, Dec 2004.

[33] M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, N. I. S. C. C. S. Program, E. D. Green, A. Sidow, and S. Batzoglou. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*, 13(4):721–731, Apr 2003.

[34] M. J. Buck and J. D. Lieb. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–60, Mar 2004.

[35] J. Buhler and M. Tompa. Finding motifs using random projections. *J Comput Biol*, 9(2):225–242, 2002.

[36] M. L. Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biol*, 5(1):201, 2003.

[37] M. L. Bulyk, P. L. F. Johnson, and G. M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–1261, Mar 2002.

[38] D. R. Cavener. Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. *Nucleic Acids Res*, 15(4):1353–1361, Feb 1987.

[39] K.-M. Chao. Computing all suboptimal alignments in linear space. In *Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching (CPM 1994)*, pages 31–42, London, UK, 1994. Springer-Verlag.

[40] K.-M. Chao, R. C. Hardison, and W. Miller. Recent developments in linear-space alignment methods: a survey. *J Comput Biol*, 1(4):271–291, 1994.

[41] D. Y. Chiang, A. M. Moses, M. Kellis, E. S. Lander, and M. B. Eisen. Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. *Genome Biol*, 4(7):R43, 2003.

[42] F. Chiaromonte, V. B. Yap, and W. Miller. Scoring pairwise genomic sequence alignments. *Pac Symp Biocomput*, pages 115–126, 2002.

[43] P. Cliften, P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, 301(5629):71–76, Jul 2003.

[44] D. L. Corcoran, E. Feingold, and P. V. Benos. FOOTER: a web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting. *Nucleic Acids Res*, 33:W442–6, 2005.

[45] G. E. Crawford, I. E. Holt, J. Whittle, B. D. Webb, D. Tai, S. Davis, E. H. Margulies, Y. Chen, J. A. Bernat, D. Ginsburg, D. Zhou, S. Luo, T. J. Vasicek, M. J. Daly, T. G. Wolfsberg, and F. S. Collins. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res*, 16(1):123–31, Jan 2006.

[46] W. H. Day and F. R. McMorris. Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res*, 20(5):1093–1099, Mar 1992.

[47] M. Dayhoff, R. Schwartz, and B. Orcutt. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, 5:345–352, 1978.

[48] E. T. Dermitzakis and A. G. Clark. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol*, 19:1114–1121, 2002.

[49] P. D'haeseleer. How does DNA sequence motif discovery work? *Nat Biotechnol*, 24(8):959–961, Aug 2006.

[50] C. Dieterich, B. Cusack, H. Wang, K. Rateitschak, A. Krause, and M. Vingron. Annotating regulatory DNA based on man-mouse genomic comparison. *Bioinformatics*, 18 Suppl 2:S84–S90, 2002.

[51] C. Dieterich, S. Rahmann, and M. Vingron. Functional inference from non-random distributions of conserved predicted transcription factor binding sites. *Bioinformatics*, 20 Suppl 1:I109–I115, Aug 2004.

[52] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis.* Cambridge University Press, 1998.

[53] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[54] S. R. Eddy. What is dynamic programming? *Nat Biotechnol*, 22(7):909–910, Jul 2004.

[55] K. Ellrott, C. Yang, F. M. Sladek, and T. Jiang. Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics*, 18 Suppl 2:S100–S109, 2002.

[56] L. Elnitski, V. X. Jin, P. J. Farnham, and S. J. M. Jones. Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Res*, Oct 2006.

[57] E. Emberly, N. Rajewsky, and E. D. Siggia. Conservation of regulatory elements between two species of Drosophila. *BMC Bioinformatics*, 4:57, Nov 2003.

[58] E. Eskin and P. A. Pevzner. Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 18 Suppl 1:S354–S363, 2002.

[59] W. J. Ewens and G. R. Grant. *Statistical Methods in Bioinformatics: An Introduction.* Springer, New York, 2001.

[60] S. Faisst and S. Meyer. Compilation of vertebrate-encoded transcription factors. *Nucleic Acids Res*, 20(1):3–26, Jan 1992.

[61] A. V. Favorov, M. S. Gelfand, A. V. Gerasimova, D. A. Ravcheev, A. A. Mironov, and V. J. Makeev. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, 21(10):2240–2245, May 2005.

[62] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

[63] J. W. Fickett. Quantitative discrimination of MEF2 sites. *Mol Cell Biol*, 16(1):437–441, Jan 1996.

[64] T. Fitzwater and B. Polisky. A SELEX primer. *Methods Enzymol*, 267:275–301, 1996.

[65] M. C. Frith, U. Hansen, J. L. Spouge, and Z. Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*, 32(1):189–200, 2004.

[66] D. Galas and A. Schmitz. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res*, 5(9):3157–70, Sep 1978.

[67] M. Garner and A. Revzin. 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, Jul 1981.

[68] B. Georgi and A. Schliep. Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics*, 22(14):e166–e173, Jul 2006.

[69] U. Gerland and T. Hwa. On the selection and evolution of regulatory DNA motifs. *J Mol Evol*, 55(4):386–400, Oct 2002.

[70] O. Gotoh. An improved algorithm for matching biological sequences. *J Mol Biol*, 162(3):705–708, Dec 1982.

[71] D. S. Gross and W. T. Garrard. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem*, 57:159–197, 1988.

[72] D. Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology.* Cambridge University Press, New York, NY, USA, 1997.

[73] O. Hallikas, K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen, E. Ukkonen, and J. Taipale. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124(1):47–59, Jan 2006.

[74] A. L. Halpern and W. J. Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol*, 15(7):910–917, Jul 1998.

[75] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. MacIsaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, Sep 2004.

[76] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174, 1985.

[77] M. Häussler. Motif discovery on promoter sequences. Master's thesis, Universität Potsdam, Institut für Informatik and IRISA/INRIA Rennes, 2005.

[78] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919, Nov 1992.

[79] G. Z. Hertz and G. D. Stormo. Identification of consensus patterns in unaligned dna and protein sequences: a large-deviation statistical basis for penalizing gaps. In *Proceedings of the 3rd International Conference on Bioinformatics and Genome Research*, pages 201–216, Singapore, 1995.

[80] G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–577, 1999.

[81] D. M. Hillis, C. Moritz, and B. K. Mable. *Molecular Systematics.* Sinauer Associates Inc. Publishers, Sunderland, MA., 1996.

[82] D. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18:341–343, 1975.

[83] P. Hong, X. S. Liu, Q. Zhou, X. Lu, J. S. Liu, and W. H. Wong. A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics*, 21(11):2636–2643, Jun 2005.

[84] C. E. Horak and M. Snyder. ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol*, 350:469–483, 2002.

[85] J. Hu, B. Li, and D. Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res*, 33(15):4899–4913, 2005.

[86] X. Huang and W. Miller. A time-efficient linear-space local similarity algorithm. *Adv. Appl. Math.*, 12(3):337–357, 1991.

[87] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296(5):1205–1214, Mar 2000.

[88] H. Ji, S. A. Vokes, and W. H. Wong. A comparative analysis of genome-wide chromatin immunoprecipitation data for mammalian transcription factors. *Nucleic Acids Res*, 34(21):e146, Nov 2006.

[89] R. Johnson, R. J. Gamblin, L. Ooi, A. W. Bruce, I. J. Donaldson, D. R. Westhead, I. C. Wood, R. M. Jackson, and N. J. Buckley. Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res*, 34(14):3862–3877, 2006.

[90] T. Jukes and C. Cantor. *Evolution of Protein Molecules.*, volume 3, pages 21–132. Academic Press, New York, 1969.

[91] T. Junier and M. Pagni. Dotlet: diagonal plots in a web browser. *Bioinformatics*, 16(2):178–179, Feb 2000.

[92] J. T. Kadonaga and R. Tjian. Affinity purification of sequence-specific DNA binding proteins. *Proc Natl Acad Sci U S A*, 83(16):5889–5893, Aug 1986.

[93] P. Kapranov, S. E. Cawley, J. Drenkow, S. Bekiranov, R. L. Strausberg, S. P. A. Fodor, and T. R. Gingeras. Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, 296(5569):916–9, May 2002.

[94] S. Karlin. Statistical studies of biomolecular sequences: score-based methods. *Philos Trans R Soc Lond B Biol Sci*, 344(1310):391–402, Jun 1994.

[95] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, 87(6):2264–2268, Mar 1990.

[96] J. Kececioglu and E. Kim. Simple and fast inverse alignment. In *Proceedings of the 10th ACM Conferences on Research in Computational Molecular Biology (RECOMB 2006)*, pages 441–455, Venice, Italy., 2006.

[97] U. Keich and P. A. Pevzner. Finding motifs in the twilight zone. *Bioinformatics*, 18(10):1374–1381, Oct 2002.

[98] A. E. Kel, E. Gössling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*, 31(13):3576–3579, Jul 2003.

[99] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–54, 2003.

[100] W. J. Kent and A. M. Zahler. Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment. *Genome Res*, 10(8):1115–1125, Aug 2000.

[101] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, Dec 1980.

[102] J. King, W. Cheung, and H. H. Hoos. Neighbourhood thresholding for projection-based motif discovery. *Bioinformatics*, 2006. accepted.

[103] N. A. Kolchanov, E. A. Ananko, O. A. Podkolodnaya, E. V. Ignatieva, I. L. Stepanenko, O. V. Kel-Margoulis, A. E. Kel, T. I. Merkulova, T. N. Goryachkovskaya, T. V. Busygina, F. A. Kolpakov, N. L. Podkolodny, A. N. Naumochkin, and A. G. Romashchenko. Transcription Regulatory Regions Database (TRRD): its status in 1999. *Nucleic Acids Res*, 27(1):303–306, Jan 1999.

[104] N. A. Kolchanov, O. A. Podkolodnaya, E. A. Ananko, E. V. Ignatieva, I. L. Stepanenko, O. V. Kel-Margoulis, A. E. Kel, T. I. Merkulova, T. N. Goryachkovskaya, T. V. Busygina, F. A. Kolpakov, N. L. Podkolodny, A. N. Naumochkin, I. M. Korostishevskaya, A. G. Romashchenko, and G. C. Overton. Transcription Regulatory Regions Database (TRRD): its status in 2000. *Nucleic Acids Res*, 28(1):298–301, Jan 2000.

[105] E. A. Kotelnikova, V. J. Makeev, and M. S. Gelfand. Evolution of transcription factor DNA binding sites. *Gene*, 347(2):255–263, Mar 2005.

[106] M. Kozak. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res*, 12(2):857–872, Jan 1984.

[107] J. Krumsiek, R. Arnold, and T. Rattei. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, 23(8):1026–1028, Feb 2007.

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

[109] B. Lenhard, A. Sandelin, L. Mendoza, P. Engström, N. Jareborg, and W. W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, 2003.

[110] P. Liò and N. Goldman. Models of molecular evolution and phylogeny. *Genome Res*, 8(12):1233–1244, Dec 1998.

[111] D. J. Lipman and W. R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, Mar 1985.

[112] J. S. Liu. *Monte Carlo strategies in scientific computing.* Springer-Verlag, New York, 2001.

[113] G. G. Loots and I. Ovcharenko. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue):W217–W221, Jul 2004.

[114] Q. Lu and B. Richardson. DNaseI hypersensitivity analysis of chromatin structure. *Methods Mol Biol*, 287:77–86, 2004.

[115] M. Z. Ludwig, A. Palsson, E. Alekseeva, C. M. Bergman, J. Nathan, and M. Kreitman. Functional evolution of a cis-regulatory module. *PLoS Biol*, 3(4):e93, Apr 2005.

[116] M. Z. Ludwig, N. H. Patel, and M. Kreitman. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development*, 125:949–58, 1998.

[117] K. D. MacIsaac and E. Fraenkel. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol*, 2(4):e36, Apr 2006.

[118] K. Malde and R. Giegerich. Calculating PSSM probabilities with lazy dynamic programming. *J. Funct. Program.*, 16:75–81, 2006.

[119] V. D. Marinescu, I. S. Kohane, and A. Riva. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, 6:79, 2005.

[120] V. D. Marinescu, I. S. Kohane, and A. Riva. The MAPPER database: a multi-genome catalog of putative transcription factor binding sites. *Nucleic Acids Res*, 33(Database issue):D91–D97, Jan 2005.

[121] L. Marsan and M. F. Sagot. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J Comput Biol*, 7(3-4):345–362, 2000.

[122] G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional Regulatory Elements in the Human Genome. *Annu Rev Genomics Hum Genet*, 7:29–59, Sep 2006.

[123] V. Matys, E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31:374–8, 2003.

[124] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110, Jan 2006.

[125] M. McArthur, S. Gerum, and G. Stamatoyannopoulos. Quantification of DNaseI-sensitivity by real-time PCR: quantitative analysis of DNaseI-hypersensitivity of the mouse beta-globin LCR. *J Mol Biol*, 313(1):27–34, Oct 2001.

[126] L. McCue, W. Thompson, C. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*, 29(3):774–782, Feb 2001.

[127] L. A. McCue, W. Thompson, C. S. Carmack, and C. E. Lawrence. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res*, 12(10):1523–1532, Oct 2002.

[128] A. M. McGuire, J. D. Hughes, and G. M. Church. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res*, 10(6):744–757, Jun 2000.

[129] A. D. McLachlan. Analysis of gene duplication repeats in the myosin rod. *J Mol Biol*, 169(1):15–30, Sep 1983.

[130] I. M. Meyer and R. Durbin. Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics*, 18(10):1309–1318, Oct 2002.

[131] M. Michael, C. Dieterich, and M. Vingron. SITEBLAST–rapid and sensitive local alignment of genomic sequences employing motif anchors. *Bioinformatics*, 21(9):2093–2094, May 2005.

[132] A. M. Moses, D. Y. Chiang, and M. B. Eisen. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac Symp Biocomput*, pages 324–35, 2004.

[133] A. M. Moses, D. Y. Chiang, M. Kellis, E. S. Lander, and M. B. Eisen. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol*, 3:19, Aug 2003.

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

[135] A. M. Moses, D. A. Pollard, D. A. Nix, V. N. Iyer, X.-Y. Li, M. D. Biggin, and M. B. Eisen. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol*, 2(10):e130, Oct 2006.

[136] R. Mott. Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull Math Biol*, 54:59–75, 1992.

[137] F. Müller, D. W. Williams, J. Kobolák, L. Gauvry, G. Goldspink, L. Orbán, and N. Maclean. Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos. *Mol Reprod Dev*, 47(4):404–12, Aug 1997.

[138] V. Mustonen and M. Lässig. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proc Natl Acad Sci U S A*, 102(44):15936–15941, Nov 2005.

[139] E. W. Myers. An overview of sequence comparison algorithms in molecular biology. Technical Report TR 91-29, University of Arizona, Tucson, Department of Computer Science, 1991.

[140] E. W. Myers and W. Miller. Optimal alignments in linear space. *Comput Appl Biosci*, 4(1):11–17, Mar 1988.

[141] G. Myers and W. Miller. Chaining multiple-alignment fragments in sub-quadratic time. In *Proceedings of the 6th annual ACM-SIAM symposium on Discrete algorithms (SODA 1995)*, pages 38–47, Philadelphia, PA, USA, 1995. Society for Industrial and Applied Mathematics.

[142] D. Naor and D. L. Brutlag. On near-optimal alignments of biological sequences. *J Comput Biol*, 1(4):349–366, 1994.

[143] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, Mar 1970.

[144] A. F. Neuwald, J. S. Liu, and C. E. Lawrence. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*, 4(8):1618–1632, Aug 1995.

[145] D. T. Odom, R. D. Dowell, E. S. Jacobsen, L. Nekludova, P. A. Rolfe, T. W. Danford, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. Core transcriptional regulatory circuitry in human hepatocytes. *Mol Syst Biol*, 2:2006.0017, 2006.

[146] E. Ohlebusch and M. I. Abouelhoda. *Handbook of Computational Molecular Biology*, chapter 15, "Chaining Algorithms and Applications in Comparative Genomics". CRC Press, 2005.

[147] R. Osada, E. Zaslavsky, and M. Singh. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, 20(18):3516–3525, Dec 2004.

[148] F. Ozsolak, J. S. Song, X. S. Liu, and D. E. Fisher. High-throughput mapping of the chromatin structure of human promoters. *Nat Biotechnol*, 25(2):244–248, Feb 2007.

[149] G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17 Suppl 1:S207–S214, 2001.

[150] G. Pavesi, G. Mauri, and G. Pesole. In silico representation and discovery of transcription factor binding sites. *Brief Bioinform*, 5(3):217–236, Sep 2004.

[151] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res*, 32(Web Server issue):W199–W203, Jul 2004.

[152] W. R. Pearson. Comparison of methods for searching protein sequence databases. *Protein Sci*, 4(6):1145–1160, Jun 1995.

[153] W. R. Pearson. Empirical statistical estimates for sequence similarity searches. *J Mol Biol*, 276(1):71–84, Feb 1998.

[154] P. A. Pevzner and S.-H. Sze. Combinatorial approaches to finding subtle signals in DNA sequences. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, volume 8, pages 269–278. AAAI Press, 2000.

[155] D. A. Pollard, A. M. Moses, V. N. Iyer, and M. B. Eisen. Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. *BMC Bioinformatics*, 7:376, 2006.

[156] A. Prakash and M. Tompa. Discovery of regulatory elements in vertebrates through comparative genomics. *Nat Biotechnol*, 23(10):1249–1256, Oct 2005.

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

[158] S. Rahmann, T. Müller, and M. Vingron. On the power of profiles for transcription factor binding site detection. *Stat. Appl. Genet. Mol. Biol.*, 2:Article 7, 2003.

[159] N. Rajewsky, M. Vergassola, U. Gaul, and E. D. Siggia. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics*, 3:30, 2002.

[160] J. Reese and W. Pearson. Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics*, 18(11):1500–1507, Nov 2002.

[161] B. Ren, F. Robert, J. Wyrick, O. Aparicio, E. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. Volkert, C. Wilson, S. Bell, and R. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–9, Dec 2000.

[162] I. Rigoutsos, A. Floratos, L. Parida, Y. Gao, and D. Platt. The emergence of pattern discovery techniques in computational biology. *Metab Eng*, 2(3):159–177, Jul 2000.

[163] E. Rocke and M. Tompa. An algorithm for finding novel gapped motifs in DNA sequences. In *Proceedings of the 2nd Annual International Conference on Computational Molecular Biology (RECOMB 1998)*, pages 228–233, New York, USA, 1998. ACM Press.

[164] S. Roepcke, S. Grossmann, S. Rahmann, and M. Vingron. T-Reg Comparator: an analysis tool for the comparison of position weight matrices. *Nucleic Acids Res*, 33(Web Server issue):W438–W441, Jul 2005.

[165] H. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, Nov 2006.

[166] E. Roulet, I. Fisch, T. Junier, P. Bucher, and N. Mermod. Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. *In Silico Biol*, 1(1):21–28, 1998.

[167] M.-F. Sagot. Spelling approximate repeated or common motifs using a suffix tree. In *Proceedings of the 3rd Latin American Symposium on Theoretical Informatics (LATIN 1998)*, pages 374–390, London, UK, 1998. Springer-Verlag.

[168] A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–D94, Jan 2004.

[169] A. Sandelin and W. W. Wasserman. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol*, 338(2):207–215, Apr 2004.

[170] A. Sandelin, W. W. Wasserman, and B. Lenhard. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, 32:W249–W252, 2004.

[171] G. K. Sandve and F. Drabløs. A survey of motif discovery methods in an integrated framework. *Biol Direct*, 1:11, 2006.

[172] D. Sankoff. The early introduction of dynamic programming into computational biology. *Bioinformatics*, 16(1):41–47, Jan 2000.

[173] D. Sankoff and J. B. Kruskal. *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*. Addison-Wesley Publication, 1983.

[174] D. E. Schones, P. Sumazin, and M. Q. Zhang. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, 21(3):307–313, Feb 2005.

[175] J. Schug and G. C. Overton. TESS: Transcription Element Search Software on the WWW. Technical report, Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania, 1997.

[176] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–107, Jan 2003.

[177] J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. Boston, Mass.: PWS Publishing Company, 1997.

[178] R. Siddharthan, E. D. Siggia, and E. van Nimwegen. PhyloGibbs: A Gibbs Sampling Motif Finder that incorporates phylogeny. *PLoS Comput Biol*, 1:e67, 2005.

[179] H. Siemen, M. Nix, E. Endl, P. Koch, J. Itskovitz-Eldor, and O. Brüstle. Nucleofection of human embryonic stem cells. *Stem Cells Dev*, 14(4):378–83, Aug 2005.

[180] S. Sinha, M. Blanchette, and M. Tompa. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5:170, Oct 2004.

[181] S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*. AAAI Press, 2000.

[182] T. F. Smith and M. S. Waterman. Comparison of biosequences. *Adv. Appl. Math.*, 2:482–489, 1981.

[183] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197, Mar 1981.

[184] E. L. Sonnhammer and R. Durbin. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, 167(1-2):GC1–G10, Dec 1995.

[185] R. Staden. Methods for calculating the probabilities of finding patterns in sequences. *CABIOS*, 5:89–96, 1989.

[186] R. Staden. Methods for discovering novel motifs in nucleic acid sequences. *Comput Appl Biosci*, 5(4):293–298, Oct 1989.

[187] D. Stanojevic, S. Small, and M. Levine. Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science*, 254:1385–7, 1991.

[188] G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, Jan 2000.

[189] J. Stoye, D. Evers, and F. Meyer. Rose: generating sequence families. *Bioinformatics*, 14(2):157–163, 1998.

[190] W. Strauss. Transfection of mammalian cells via lipofection. *Methods Mol Biol*, 54:307–27, 1996.

[191] S. J. H. Sui, J. R. Mortimer, D. J. Arenillas, J. Brumm, C. J. Walsh, B. P. Kennedy, and W. W. Wasserman. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*, 33:3154–3164, 2005.

[192] R. L. Tatusov, S. F. Altschul, and E. V. Koonin. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci U S A*, 91(25):12091–12095, Dec 1994.

[193] J. L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol*, 33(2):114–124, Aug 1991.

[194] M. Tompa. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 1999)*, pages 262–271, Heidelberg, Germany, 1999. AAAI Press.

[195] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. D. Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–144, Jan 2005.

[196] F. Tronche, F. Ringeisen, M. Blumenfeld, M. Yaniv, and M. Pontoglio. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J Mol Biol*, 266(2):231–245, Feb 1997.

[197] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–510, Aug 1990.

[198] A. Ureta-Vidal, L. Ettwiller, and E. Birney. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet*, 4:251–62, 2003.

[199] J. van Helden, M. del Olmo, and J. E. Pérez-Ortín. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Res*, 28(4):1000–1010, Feb 2000.

[200] T. Vavouri and G. Elgar. Prediction of cis-regulatory elements using binding site matrices–the successes, the failures and the reasons for both. *Curr Opin Genet Dev*, 15(4):395–402, Aug 2005.

[201] M. Vettese-Dadey, P. A. Grant, T. R. Hebbes, C. Crane-Robinson, C. D. Allis, and J. L. Workman. Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA in vitro. *EMBO J*, 15(10):2508–2518, May 1996.

[202] M. Vingron. Near-optimal sequence alignment. *Curr Opin Struct Biol*, 6(3):346–352, Jun 1996.

[203] M. Vingron and M. Waterman. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J Mol Biol*, 235(1):1–12, Jan 1994.

[204] D. Vlieghe, A. Sandelin, P. J. D. Bleser, K. Vleminckx, W. W. Wasserman, F. van Roy, and B. Lenhard. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*, 34(Database issue):D95–D97, Jan 2006.

[205] I. M. Wallace, G. Blackshields, and D. G. Higgins. Multiple sequence alignments. *Curr Opin Struct Biol*, 15(3):261–266, Jun 2005.

[206] W. W. Wasserman, M. Palumbo, W. Thompson, J. W. Fickett, and C. E. Lawrence. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet*, 26(2):225–228, Oct 2000.

[207] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5:276–87, 2004.

[208] B. Wasylyk. Enhancers and transcription factors in the control of gene expression. *Biochim Biophys Acta*, 951(1):17–35, Nov 1988.

[209] M. S. Waterman. Efficient sequence alignment algorithms. *J Theor Biol*, 108(3):333–337, Jun 1984.

[210] M. S. Waterman. *Introduction to Computational Biology. Maps, Sequences and Genomes.* Chapman & Hall, London, UK, 1995.

[211] M. S. Waterman, R. Arratia, and D. J. Galas. Pattern recognition in several sequences: consensus and alignment. *Bull Math Biol*, 46(4):515–527, 1984.

[212] M. S. Waterman and M. Eggert. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J Mol Biol*, 197(4):723–728, Oct 1987.

[213] M. S. Waterman, T. F. Smith, and W. A. Beyer. Some biological sequence metrics. *Adv Math*, 20:367–387, 1976.

[214] M. S. Waterman, T. F. Smith, and H. L. Katcher. Algorithms for restriction map comparisons. *Nucleic Acids Res*, 12(1 Pt 1):237–242, Jan 1984.

[215] P. J. Wittkopp. Evolution of cis-regulatory sequence and function in Diptera. *Heredity*, 97(3):139–147, Sep 2006.

[216] C. T. Workman and G. D. Stormo. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*, pages 467–478, 2000.

[217] J. Workman and R. Kingston. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu Rev Biochem*, 67:545–79, 1998.

[218] G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20(9):1377–419, Sep 2003.

[219] J. Wu, L. T. Smith, C. Plass, and T. H.-M. Huang. ChIP-chip comes of age for genome-wide functional analysis. *Cancer Res*, 66(14):6899–902, Jul 2006.

[220] T. D. Wu, C. G. Nevill-Manning, and D. L. Brutlag. Fast probabilistic analysis of sequence function using scoring matrices. *Bioinformatics*, 16(3):233–244, Mar 2000.

[221] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, Mar 2005.

[222] G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, 309(5734):626–30, Jul 2005.

[223] F. Zhao, Z. Xuan, L. Liu, and M. Q. Zhang. TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Res*, 33(Database issue):D103–D107, Jan 2005.

[224] J. Zhu and M. Q. Zhang. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, 15(7-8):607–611, 1999.

Appendix A: Software availability

The SimAnn/eSimAnn package is available upon request. The package contains:

- C++ source code for the dynamic programming alignment algorithm
- Perl code for the estimation of the profile-related parameters.

The C++ code (gcc v3.4.4 or later) performs local alignments with affine gap costs, with or without profiles.

The Perl library (Perl v5.8.4 or later and BioPerl 1.4 or later) provides modules needed for the estimation of profile-related parameters (`PSA` and *pen*), given a user-defined position-specific count matrix. A Perl script to calculate the above (getPsa.pl) is also included. The Perl library uses parts of the GENEREGng package developed in-house for single sequence TFBS annotation. The build-up is the pair-profile related code concerning the annotated alignment approach.

Appendix B: Summary (German)

Zusammenfassung

Ansätze, die das bessere Verständnis von Mechanismen transkriptioneller Regulation zum Ziel haben, bauen oft auf der Annotation der Genomsequenz bezüglich DNA Bindestellen von Transkriptionsfaktoren (TFBS) auf. Dies ist auch das Thema der vorliegenden Arbeit. Von großem Interesse sind Bindestellen, die zwischen zwei oder mehr Spezies erhalten sind. Dem liegt die Hypothese zugrunde, dass diesen mit grösserer Wahrscheinlichkeit eine biologische Funktion zukommt. Gewöhnlich findet man solche Bindestellen mit Hilfe von Computermethoden, die einen separaten Alignment- und Annotationsschritt durchführen. Ist die Beschreibung der Bindestelle nicht sehr spezifisch, oder sind sich die zu annotierenden Sequenzen sich nicht besonders ähnlich, so bereitet die lokale Gapstruktur im Alignment Probleme beim Auffinden konservierter Bindestellen. In dieser Arbeit stellen wir neue Methoden vor, die Sequenzalignment und -annotation simultan ausführen und deren Endergebnis annotierte Alignments - paarweise Sequenzalignments mit als TFBS annotierten Teilsequenzen - sind.

Diesbezüglich wurde der Standardansatz paarweiser Alignments dahingehend erweitert, dass nun zusätzliche Zustände für TFBS beschreibende Profile möglich sind. Wir entwickeln statistische Methoden, die das Schätzten dem Profil assoziierter algorithmischer Parameter mit kontrollierten Fehlern erster oder zweiter Art erlauben. Zusammengenommen ergibt dies den Kern unseres Tools **SimAnn**. Zusätzlich zeigen wir, wie die von uns entwickelten Methoden ergänzt werden können, so dass den evolutionären Charakteristika der TFBS Rechnung getragen wird. Dies wird in dem Tool **eSimAnn** zusammengefasst.

Wir zeigen den Effekt eines simultanen Zugangs zu Alignment und TFBS Annotation im Kontrast zu Verfahren auf, die mehrere sequenzielle Schritte durchführen. Dazu führen wir Simulationsstudien durch und vergleichen Resultate auf realen Sequenzdatensätzen. Ein simultaner Zugang erlaubt es Gaps im Alignment automatisch lokal so zu positionieren, dass die Struktur perfekt alignierter TFBS hervorgehoben wird. Dies macht ein Entfernen von Alignmentfehlern, wie es bei sequenziellen Verfahren üblich ist, unnötig. Als besonders vorteilhaft stellt sich dies für Sequenzen mit nur mäßiger Konservierung und für Transkriptionsfaktoren mit mittlerer Profilqualität heraus. Unsere Analyse beinhaltet die Modellierung des Problems annotierter Alignments als ein „extended pair Hidden Markov Model“ und zeigt Verbindungen und Zusammenhänge verschiedener theoretischer Konzepte auf. Die Arbeit ist wie folgt strukturiert:

- Kapitel 1 und 2 führen in die grundlegenden Konzepte und Methoden ein, die im Weiteren benötigt werden. Kapitel 1 gibt einen Überblick über aktuelle Methoden, sowohl experimentell als auch in silico. In Kapitel 2 diskutieren wir formale Aspekte, die sowohl TFBS Profilen als auch Alignments zugrunde liegen.

- Annotierte Alignments werden in Kapitel 3 vorgestellt. Als erstes wird ein erweiterter Algorithmus aus der klasse der dynamischen Programmierung beschrieben, mit dem sich annotierte Alignments erstellen lassen. Danach stellen wir zwei Methoden zum Schätzen Profil assoziierter Parameter vor. Erst werden die DNA Bindestellen von Transkriptionsfaktoren in beiden Sequenzen unabhängig behandelt; danach wird mit Hilfe von positionsspezifischen evolutionären Modellen der Abhängigkeit zwischen Bindestellen explizit Rechnung getragen. Im Weiteren formulieren wir Annotierte Alignments als „extended pair Hidden Markov Model“ und schließen mit einer Laufzeitanalyse des vorgestellten Algorithmus.
- In Kapitel 4 untersuchen wir verschiedene Aspekte unsres Ansatzes. Wir untermauern unseren statistischen Ansatz zur Parameterwahl aus Kapitel 3 mit Simulationen. Mit Simulierten und realen Daten kontrastieren wir unseren simultanen Ansatz gegenüber sequenziellen mehrschrittigen Strategien. Wir betrachten sowohl den Einfluss evolutionärer Distanz als auch den der Qualität des TFBS Profils.

Schlussendlich wird in Kapitel 5 eine Zusammenfassung gegeben. Zusätzlich werden Perspektiven für zukünftige Forschung aufgezeigt, denen hier beschriebene Methoden zugrunde liegen.

Appendix C: Short Curriculum Vitae

Der Lebenslauf ist in der Online-Version
aus Gründen des Datenschutzes nicht enthalten

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, May 2007 Abha Singh Bais