Overview of motif discovery methods in an integrated framework

The following tables show the characteristics of 119 motif discovery methods with respect to an integrated model described in a forthcoming article. More specifically, table 1 gives an overview of match models, occurrence priors and score functions on inter-motif distances. Table 2 gives an overview of models of single motif combination, gene level score and genome level score (motif signifiance). As these aspects are not always described in articles presenting new methods, and not even relevant for all methods, some fields are left blank. Please contact us if you have any corrections or other comments regarding the tables.

Table 1: Match model, occurrence prior and distance score for different methods

NR	ALGORITHM NAME	MATCH MODEL	OCC. PRIOR	DISTANCE FUNCTION
1	Pratt2[53]	reg.exp	-	-
2	MultiProfiler[55]	mismatch	-	-
3	Weeder[73]	mismatch	-	-
4	YMF[93, 96]	reg.exp	-	-
5	TEIRESIAS[82]	reg.exp	-	-
6	Splash[44]	reg.exp		-
7	Mitra[31]	mismatch	-	-
8	Mitra-dyad[31]	mismatch	-	constraint
9	Mot.Disc.Toolkit[7]	mismatch	-	
10	MERMAID[49]	PWM	-	constraint
11	DMotifs[94]	reg.exp	-	constraint
12	Dyad analysis[107]	oligos	-	constraint
13	TFBScluster[28]	PWM	strand bias	window
14	MCAST[6]	PWM	-	gap penalty
15	GCMD[86]	flexible	-	constraint
16	[63]	mismatch	-	
17	[43]	PWM	-	-
18	[116]	DM	-	-
19	REDUCE[20]	PWM	-	constraint

Table 1: Match model, occurrence prior and distance score for different methods

NR	ALGORITHM NAME	MATCH MODEL	OCC. PRIOR	DISTANCE FUNCTION
20	MDScan[65]	PWM		-
21	HMDM[112, 113]	DM	-	-
22	[8]	DM	-	-
23	Gibbs sampler[61]	PWM	-	uniform
24	MEME[5]	PWM	-	-
25	[19]	oligos	-	-
26	LOGOS[115]	DM	-	distribution
27	[18]	known sites	-	constraint
28	[58]	oligos	-	
29	MM[4]	PWM	-	-
30	Motif regressor[24]	PWM	-	-
31	SOMBERO[66]	PWM	-	-
32	MISAE[98]	mismatch	-	-
33	CENSUS[32]	mismatch	-	-
34	MScan[52]	PWM	-	
35	[92]	reg.exp	-	constraint
36	[75]		-	constraint
37	[39]		-	distribution
38	[17]	flexible	-	uniform
39	[95]		-	-
40	Oligo-analysis[105]	oligos	-	-
41	Pattern-assembly[106]		-	-
42	ModuleSearcher[2]	PWM	conservation	window
43	[1]	PWM	-	window
44	COMET[40]		-	-
45	Stubb[97]	PWM	conservation	window
46	Modulescanner[2]	PWM	conservation	window
47	MotifLocator[2]	PWM	conservation	window
48	MotifSampler[100]	PWM	-	-
49	Footprinter[14]		-	-
50	GANN[9]	flexible	DNA struct.	window
51	FrameWorker[21]	PWM	-	constraint

Table 1: Match model, occurrence prior and distance score for different methods

NR	ALGORITHM NAME	MATCH MODEL	OCC. PRIOR	DISTANCE FUNCTION
52	[25]	oligos	conservation	-
53	[26]	oligos	-	-
54	[22]	oligos	-	-
55	[27]		-	-
56	MITRA-PSSM[30]	PWM	-	-
57	Partition-PSSM[30]	PWM	-	-
58	ModelGenerator[35]	PWM	-	distribution
59	ModelInspector[35]	PWM	-	distribution
60	GLAM[38]		-	-
61	DMS[48]	PWM	-	
62	ANN-Spec[111]	PWM	-	-
63	[110]	PWM	conservation	window
64	CoBind[42]	PWM	-	window
65	[78]	DM		-
66	OrthoMEME[76]	PWM	-	
67	WINNOWER[74]	mismatch	-	-
68	[56]	PWM	-	-
69	MAPPER[67]	HMM	-	
70	[50]	oligos	-	-
71	Footprinter[13, 12]	mismatch	-	-
72	[70]	PWM	-	-
73	Cister[36]	PWM	-	distribution
74	PromoterInsp. [87]	oligos	-	constraint
75	[15]	PWM	-	
76	SeSiMCMC [33]	PWM	-	-
77	FastM[60]	PWM	-	constraint
78	SMILE[69, 68]	mismatch	-	constraint
79	[108]	flexible	-	distribution
80	BioProspector[64]	PWM	strand bias	constraint
81	[88]	PWM	-	
82	[93]	reg.exp	-	constraint
83	[104]	mismatch	-	-

Table 1: Match model, occurrence prior and distance score for different methods

NR	**ALGORITHM NAME**	**MATCH MODEL**	**OCC. PRIOR**	**DISTANCE FUNCTION**
84	ConsecID[91]	PWM	conservation	window
85	SCORE[80]		-	window
86	ClusterScan[57]	PWM	-	constraint
87	Gibbs recursive [103]	PWM	location	distribution
88	[72]	known sites	-	-
89	[47]	PWM	-	-
90	[10]	DM	-	-
91	Cis-analyst [11]	PWM	-	window
92	[46]	PWM	-	-
93	BioOptimizer[51]	PWM	-	constraint
94	[117]	DM	-	-
95	[89]	PWM	-	
96	[71]	PWM	-	-
97	[23]	oligos	-	-
98	Clover[37]	PWM	-	-
99	ProMapper[77]	DM	-	-
100	COOP[16]	reg.exp	-	-
101	CAGER[84]		-	-
102	AlignACE[83]	PWM	-	-
103	Consensus[45]	PWM	-	-
104	Improbizer[3]	PWM	-	-
105	QuickScore[81]	IUPAC	-	-
106	Motifprototyper[114]	DM		-
107	CisModule[118]	PWM	-	
108	[79]	PWM	-	-
109	NONPAR [59]	Mixture	-	-
110	[34]	alignment	-	-
111	NestedMICA[29]	PWM	-	
112	[99]	reg.exp	-	-
113	Motif sampler[101]	PWM	-	-
114	[54]	PWM	-	uniform
115	[102]	PWM	conservation	constraint

Table 1: Match model, occurrence prior and distance score for different methods

NR	ALGORITHM NAME	MATCH MODEL	OCC. PRIOR	DISTANCE FUNCTION
116	ConSite[85, 62]	PWM	conservation	-
117	PhyloCon[109]	PWM	-	-
118	[41]	PWM	-	-
119	[90]	PWM	-	uniform

Table 2: Composite motif model, gene score and significance evalution for different methods

NR	ALGORITHM NAME	MOTIF COMB.	GENE SCORE	SIGNIFI-CANCE
1	Pratt2[53]	-		
2	MultiProfiler[55]	-		
3	Weeder[73]	-	max	sum
4	YMF[93, 96]	-		
5	TEIRESIAS[82]	-		
6	Splash[44]	-	max	sum
7	Mitra[31]	-		
8	Mitra-dyad[31]	dyad		
9	Mot.Disc.Toolkit[7]			
10	MERMAID[49]	dyad		
11	DMotifs[94]	dyad		
12	Dyad analysis[107]	dyad		
13	TFBScluster[28]	intersection		-
14	MCAST[6]	sum	HMM	classification
15	GCMD[86]	intersection	max	sum
16	[63]			
17	[43]	dictionary		
18	[116]	-		
19	REDUCE[20]	dyad	sum	regression
20	MDScan[65]	-	max	MAP
21	HMDM[112, 113]	-		
22	[8]	-		

Table 2: Composite motif model, gene score and significance evalution for different methods

NR	ALGORITHM NAME	MOTIF COMB.	GENE SCORE	SIGNIFI-CANCE
23	Gibbs sampler[61]	intersection	max	p-value
24	MEME[5]	-		IC of PWM
25	[19]	dictionary	special	
26	LOGOS[115]	HMM	HMM	
27	[18]	dyad		
28	[58]	sum		
29	MM[4]	-		
30	Motif regressor[24]	-	sum	regression
31	SOMBERO[66]	SOM		
32	MISAE[98]	-		
33	CENSUS[32]	-		
34	MScan[52]	min comp. score		
35	[92]	intersection		
36	[75]	mismatch	max	
37	[39]			
38	[17]	intersection		sum
39	[95]	-		
40	Oligo-analysis[105]	-	sum	sum
41	Pattern-assembly[106]	-		
42	ModuleSearcher[2]	sum	max	sum
43	[1]	sum	max	
44	COMET[40]	-		
45	Stubb[97]	HMM	HMM	-
46	Modulescanner[2]	sum	max	sum
47	MotifLocator[2]	sum	max	
48	MotifSampler[100]	-		
49	Footprinter[14]	-		
50	GANN[9]	ANN	ANN	
51	FrameWorker[21]	intersection		min
52	[25]	-	p-value	-
53	[26]	single motif	p-value	-
54	[22]	single motif	p-value	-

Table 2: Composite motif model, gene score and significance evalution for different methods

NR	ALGORITHM NAME	MOTIF COMB.	GENE SCORE	SIGNIFI-CANCE
55	[27]	single motif		regression
56	MITRA-PSSM[30]	-	max	Discrete IC
57	Partition-PSSM[30]	-	max	Discrete IC
58	ModelGenerator[35]	sum		min
59	ModelInspector[35]	sum		min
60	GLAM[38]	-		
61	DMS[48]			
62	ANN-Spec[111]	-	max	IC of PWM
63	[110]	Logistic regr.	max	regression
64	CoBind[42]	sum	sum	sum
65	[78]	-	-	-
66	OrthoMEME[76]			
67	WINNOWER[74]	-	max	
68	[56]	-	max	sum
69	MAPPER[67]			
70	[50]	-	max	
71	Footprinter[13, 12]	-	max	
72	[70]	-		
73	Cister[36]	HMM	HMM	
74	PromoterInsp. [87]	intersection		
75	[15]	mixture model	mixture model	
76	SeSiMCMC [33]		mixture model	
77	FastM[60]	sum	max	
78	SMILE[69, 68]	intersection	max	sum
79	[108]	intersection		
80	BioProspector[64]	sum	sum	z-score
81	[88]	-	logistic func.	regression
82	[93]	dyad	sum	z-value
83	[104]	-	max	z-value
84	ConsecID[91]	intersection	sum	p-value
85	SCORE[80]	intersection	sum	p-value
86	ClusterScan[57]	sum	sum	

Table 2: Composite motif model, gene score and significance evalution for different methods

NR	ALGORITHM NAME	MOTIF COMB.	GENE SCORE	SIGNIFI-CANCE
87	Gibbs recursive [103]	mixture model	mixture model	
88	[72]	special	special	special
89	[47]	-	hyperb. tan.	classification
90	[10]	-	-	
91	Cis-analyst [11]	sum	-	
92	[46]	-		special
93	BioOptimizer[51]	dyad	sum	
94	[117]	-		
95	[89]	sum	max	
96	[71]	-		
97	[23]	-	-	special
98	Clover[37]	-	sum	special
99	ProMapper[77]	-		
100	COOP[16]	-	-	
101	CAGER[84]	-		
102	AlignACE[83]	-	mixture model	p-value
103	Consensus[45]	-		IC of PWM
104	Improbizer[3]	-	mixture model	mixture model
105	QuickScore[81]	-		
106	Motifprototyper[114]	-		
107	CisModule[118]	mixture model	mixture model	
108	[79]	-	mixture model	
109	NONPAR [59]	-	-	
110	[34]	-	max	
111	NestedMICA[29]	mixture model	mixture model	
112	[99]	-	max	p-value
113	Motif sampler[101]	-	distribution	
114	[54]	intersection	max	expr. similarity
115	[102]	Markov model		
116	ConSite[85, 62]	-	-	-
117	PhyloCon[109]	-	sum	sum
118	[41]	-	-	special

Table 2: Composite motif model, gene score and significance evalution for different methods

NR	ALGORITHM NAME	MOTIF COMB.	GENE SCORE	SIGNIFI-CANCE
119	[90]	dyad	max	p-valus

References

[1] Stein Aerts, Peter Van Loo, Yves Moreau, and Bart De Moor. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*, 20(12):1974–6, 2004.

[2] Stein Aerts, Peter Van Loo, Gert Thijs, Yves Moreau, and Bart De Moor. Computational detection of cis -regulatory modules. *Bioinformatics*, 19 Suppl 2(1367-4803):II5–II14, 2003.

[3] Wanyuan Ao, Jeb Gaudet, W James Kent, Srikanth Muttumu, and Susan E Mango. Environmentally induced foregut remodeling by pha-4/foxa and daf-12/nhr. *Science*, 305(5691):1743–6, 2004.

[4] T L Bailey and C Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2(1553-0833):28–36, 1994.

[5] T L Bailey and C Elkan. The value of prior knowledge in discovering motifs with meme. *Proc Int Conf Intell Syst Mol Biol*, 3(1553-0833):21–9, 1995.

[6] Timothy L Bailey and William Stafford Noble. Searching for statistically significant regulatory modules. *Bioinformatics*, 19 Suppl 2(1367-4803):II16–II25, 2003.

[7] N. E. Baldwin, R. L. Collins, M. A. Langston, M. R. Leuze, C. T. Symons, and B. H. Voy. High performance computational tools for motif discovery. In *18th International Parallel and Distributed Processing Symposium (IPDPS'04) - Workshop 9*, page p. 192a, 2004.

[8] Y Barash, G Elidan, N Friedman, and T Kaplan. Modeling dependencies in protein-dna binding sites. In *RECOMB '03: Proceedings of the*

seventh annual international conference on Computational molecular biology, pages 28–37, New York, NY, USA, 2003. ACM Press.

[9] Robert G Beiko and Robert L Charlebois. Gann: genetic algorithm neural networks for the detection of conserved combinations of features in dna. *BMC Bioinformatics*, 6(1):36, 2005.

[10] I Ben-Gal, A Shani, A Gohr, J Grau, S Arviv, A Shmilovici, S Posch, and I Grosse. Identification of transcription factor binding sites with variable-order bayesian networks. *Bioinformatics*, 21(11):1367–4803, 2005.

[11] Benjamin P Berman, Yutaka Nibu, Barret D Pfeiffer, Pavel Tomancak, Susan E Celniker, Michael Levine, Gerald M Rubin, and Michael B Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proc Natl Acad Sci U S A*, 99(2):757–62, 2002.

[12] M Blanchette, B Schwikowski, and M Tompa. An exact algorithm to identify motifs in orthologous sequences from multiple species. *Proc Int Conf Intell Syst Mol Biol*, 8(1553-0833):37–45, 2000.

[13] Mathieu Blanchette and Martin Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*, 12(5):739–48, 2002.

[14] Mathieu Blanchette and Martin Tompa. Footprinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res*, 31(13):3840–2, 2003.

[15] Konstantinos Blekas, Dimitrios I Fotiadis, and Aristidis Likas. Greedy mixture learning for multiple motif discovery in biological sequences. *Bioinformatics*, 19(5):607–17, 2003.

[16] S Bortoluzzi, A Coppe, A Bisognin, C Pizzi, and GA Danieli. A multistep bioinformatic approach detects putative regulatory elements in gene promoters. *BMC Bioinformatics*, 6(1):121, 2005.

[17] A Brazma, J Vilo, E Ukkonen, and K Valtonen. Data mining for regulatory elements in yeast genome. *Proc Int Conf Intell Syst Mol Biol*, 5(1553-0833):65–74, 1997.

[18] Martha L Bulyk, Abigail M McGuire, Nobuhisa Masuda, and George M Church. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in escherichia coli. *Genome Res*, 14(2):201–8, 2004.

[19] H J Bussemaker, H Li, and E D Siggia. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A*, 97(18):10096–100, 2000.

[20] H J Bussemaker, H Li, and E D Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–71, 2001.

[21] K Cartharius, K Frech, K Grote, B Klocke, M Haltmeier, A Klingenhoff, M Frisch, M Bayerlein, and T Werner. Matinspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 21(13):2933–2942, 2005.

[22] Michele Caselle, Ferdinando Di Cunto, and Paolo Provero. Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. *BMC Bioinformatics*, 3(1):7, 2002.

[23] Paul Cliften, Priya Sudarsanam, Ashwin Desikan, Lucinda Fulton, Bob Fulton, John Majors, Robert Waterston, Barak A Cohen, and Mark Johnston. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science*, 301(5629):71–6, 2003.

[24] Erin M Conlon, X Shirley Liu, Jason D Lieb, and Jun S Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A*, 100(6):3339–44, 2003.

[25] D Cora, C Herrmann, C Dieterich, F Di Cunto, P Provero, and M Caselle. Ab initio identification of putative human transcription factor binding sites by comparative genomics. *BMC Bioinformatics*, 6(1):110, 2005.

[26] Davide Cora, Ferdinando Di Cunto, Paolo Provero, Lorenzo Silengo, and Michele Caselle. Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs. *BMC Bioinformatics*, 5(1):57, 2004.

[27] Marla D. Curran, Hong Liu, Fan Long, and Nanxiang Ge. Statistical methods for joint data mining of gene expression and dna sequence database. *SIGKDD Explor Newsl*, 5(2):122–129, 2003.

[28] I. J. Donaldson, M. Chapman, and B. Gottgens. TFBScluster: a resource for the characterisation of transcriptional regulatory networks. *Bioinformatics*, 21(13):1367–4803, 2005.

[29] Thomas A Down and Tim J P Hubbard. Nestedmica: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res*, 33(5):1445–53, 2005.

[30] E Eskin, WS Noble, Y Singer, and S Snir. A unified approach for sequence prediction using sparse sequence models. Technical report, Hebrew University, 2003.

[31] Eleazar Eskin and Pavel A Pevzner. Finding composite regulatory patterns in dna sequences. *Bioinformatics*, 18 Suppl 1(1367-4803):S354–63, 2002.

[32] P. A. Evans and A. D. Smith. Toward optimal motif enumeration. In *Proceedings of Workshop on Algorithms and Data Structures (WADS 2003)*, volume 2751 of *LNCS*, pages 47–58. Springer-Verlag, 2003.

[33] 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, 2005.

[34] Gary B Fogel, Dana G Weekes, Gabor Varga, Ernst R Dow, Harry B Harlow, Jude E Onyia, and Chen Su. Discovery of sequence motifs related to coexpression of genes using evolutionary computation. *Nucleic Acids Res*, 32(13):3826–35, 2004.

[35] K Frech and T Werner. Specific modelling of regulatory units in dna sequences. In *Pac Symp Biocomput*, pages 151–62, 1997.

[36] M C Frith, U Hansen, and Z Weng. Detection of cis-element clusters in higher eukaryotic dna. *Bioinformatics*, 17(10):878–89, 2001.

[37] Martin C Frith, Yutao Fu, Liqun Yu, Jiang-Fan Chen, Ulla Hansen, and Zhiping Weng. Detection of functional dna motifs via statistical over-representation. *Nucleic Acids Res*, 32(4):1372–81, 2004.

[38] Martin C Frith, Ulla Hansen, John L Spouge, and Zhiping Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*, 32(1):189–200, 2004.

[39] Martin C Frith, Michael C Li, and Zhiping Weng. Cluster-buster: Finding dense clusters of motifs in dna sequences. *Nucleic Acids Res*, 31(13):3666–8, 2003.

[40] Martin C Frith, John L Spouge, Ulla Hansen, and Zhiping Weng. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res*, 30(14):3214–24, 2002.

[41] Naum I Gershenzon, Gary D Stormo, and Ilya P Ioshikhes. Computational technique for improvement of the position-weight matrices for the dna/protein binding sites. *Nucleic Acids Res*, 33(7):2290–301, 2005.

[42] D GuhaThakurta and G D Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17(7):608–21, 2001.

[43] M. Gupta and J. S. Liu. Discovery of conserved sequence patterns using a stochastic dictionary model. *Journal of the American Statistical Association*, 98:55–66, 2003.

[44] R K Hart, A K Royyuru, G Stolovitzky, and A Califano. Systematic and fully automated identification of protein sequence patterns. *J Comput Biol*, 7(3-4):585–600, 2000.

[45] G Z Hertz and G D Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–77, 1999.

[46] I Holmes and W J Bruno. Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc Int Conf Intell Syst Mol Biol*, 8(1553-0833):202–10, 2000.

[47] 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, 2005.

[48] Y J Hu, S Sandmeyer, C McLaughlin, and D Kibler. Combinatorial motif analysis and hypothesis generation on a genomic scale. *Bioinformatics*, 16(3):222–32, 2000.

[49] Yuh-Jyh Hu. Finding subtle motifs with variable gaps in unaligned dna sequences. *Comput Methods Programs Biomed*, 70(1):11–20, 2003.

[50] L J Jensen and S Knudsen. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, 16(4):326–33, 2000.

[51] Shane T Jensen and Jun S Liu. Biooptimizer: a bayesian scoring function approach to motif discovery. *Bioinformatics*, 20(10):1557–64, 2004.

[52] O Johansson, W Alkema, W W Wasserman, and J Lagergren. Identification of functional clusters of transcription factor binding motifs in genome sequences: the mscan algorithm. *Bioinformatics*, 19 Suppl 1(1367-4803):i169–76, 2003.

[53] I Jonassen. Efficient discovery of conserved patterns using a pattern graph. *Comput Appl Biosci*, 13(5):509–22, 1997.

[54] Je-Gun Joung, Sok June Oh, and Byoung-Tak Zhang. Searching transcriptional modules using evolutionary algorithms. In *Parallel Problem Solving from Nature - PPSN VIII*, volume 3242 of *Lecture Notes in Computer Science*, pages 532–540, Berlin, 2004. Springer-Verlag.

[55] U Keich and P A Pevzner. Finding motifs in the twilight zone. *Bioinformatics*, 18(10):1374–81, 2002.

[56] A Kel, Y Tikunov, N Voss, and E Wingender. Recognition of multiple patterns in unaligned sets of sequences: comparison of kernel clustering method with other methods. *Bioinformatics*, 20(10):1512–6, 2004.

[57] Alexander Kel, Olga Kel-Margoulis, Tatyana Ivanova, and Edgar Wingender. Clusterscan: A tool for automatic annotation of genomic regulatory sequences by searching for composite clusters. In *Proceedings of the German Conference on Bioinformatics*, pages 96–101, 2001.

[58] Sunduz Keles, Mark van der Laan, and Michael B Eisen. Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18(9):1167–75, 2002.

[59] Oliver D King and Frederick P Roth. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res*, 31(19):e116, 2003.

[60] A Klingenhoff, K Frech, K Quandt, and T Werner. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics*, 15(3):180–6, 1999.

[61] 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):208–14, 1993.

[62] Boris Lenhard, Albin Sandelin, Luis Mendoza, Par Engstrom, Niclas Jareborg, and Wyeth W Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, 2003.

[63] Han-Lin Li and Chang-Jui Fu. A linear programming approach for identifying a consensus sequence on dna sequences. *Bioinformatics*, 21(19):1838–1845, 2005.

[64] X Liu, D L Brutlag, and J S Liu. Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In *Pac Symp Biocomput*, pages 127–38, 2001.

[65] X Shirley Liu, Douglas L Brutlag, and Jun S Liu. An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, 20(8):835–9, 2002.

[66] Shaun Mahony, David Hendrix, Aaron Golden, Terry J Smith, and Daniel S Rokhsar. Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, 21(9):1807–1814, 2005.

[67] 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(1):79, 2005.

[68] 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):345–62, 2000.

[69] Laurent Marsan and Marie-France Sagot. Extracting structured motifs using a suffix treealgorithms and application to promoter consensus identification. In *RECOMB '00: Proceedings of the fourth annual international conference on Computational molecular biology*, pages 210–219, New York, NY, USA, 2000. ACM Press.

[70] 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):744–57, 2000.

[71] A A Mironov, E V Koonin, M A Roytberg, and M S Gelfand. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res*, 27(14):2981–9, 1999.

[72] Peter J Park, Atul J Butte, and Isaac S Kohane. Comparing expression profiles of genes with similar promoter regions. *Bioinformatics*, 18(12):1576–84, 2002.

[73] G Pavesi, G Mauri, and G Pesole. An algorithm for finding signals of unknown length in dna sequences. *Bioinformatics*, 17 Suppl 1(1367-4803):S207–14, 2001.

[74] P A Pevzner and S H Sze. Combinatorial approaches to finding subtle signals in dna sequences. *Proc Int Conf Intell Syst Mol Biol*, 8(1553-0833):269–78, 2000.

[75] Alberto Policriti, Nicola Vitacolonna, Michele Morgante, and Andrea Zuccolo. Structured motifs search. In *RECOMB '04: Proceedings of the eighth annual international conference on Computational molecular biology*, pages 133–139, New York, NY, USA, 2004. ACM Press.

[76] A Prakash, M Blanchette, S Sinha, and M Tompa. Motif discovery in heterogeneous sequence data. In *Pac Symp Biocomput*, pages 348–59, 2004.

[77] R. Pudimat, E. G. Schukat-Talamazzini, and R. Backofen. A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics*, 21(14):3082–3088, 2005.

[78] R. Pudimat, Ernst Günter Schukat-Talamazzini, and Rolf Backofen. Feature based representation and detection of transcription factor binding sites. In *Proceedings of the German Conference on Bioinformatics*, pages 43–52, 2004.

[79] Zhaohui S Qin, Lee Ann McCue, William Thompson, Linda Mayerhofer, Charles E Lawrence, and Jun S Liu. Identification of co-regulated genes through bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol*, 21(4):435–9, 2003.

[80] Mark Rebeiz, Nick L Reeves, and James W Posakony. Score: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. site clustering over random expectation. *Proc Natl Acad Sci U S A*, 99(15):9888–93, 2002.

[81] M. Régnier and A. Denise. Rare events and conditional events on random strings. *Discrete Math. Theor. Comput. Sci.*, 6:191–214, 2004.

[82] I Rigoutsos and A Floratos. Combinatorial pattern discovery in biological sequences: The teiresias algorithm. *Bioinformatics*, 14(1):55–67, 1998.

[83] F P Roth, J D Hughes, P W Estep, and G M Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nat Biotechnol*, 16(10):939–45, 1998.

[84] J Ruan and W Zhang. Cager: classification analysis of gene expression regulation using multiple information sources. *BMC Bioinformatics*, 6(1):114, 2005.

[85] Albin Sandelin, Wyeth W Wasserman, and Boris Lenhard. Consite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, 32(Web Server issue):W249–52, 2004.

[86] Geir Kjetil Sandve and Finn Drabløs. Generalized composite motif discovery. In *7th Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems, KES*, volume In press of *LNCS/LNAI*. Springer-Verlag, 2005.

[87] M Scherf, A Klingenhoff, and T Werner. Highly specific localization of promoter regions in large genomic sequences by promoterinspector: a novel context analysis approach. *J Mol Biol*, 297(3):599–606, 2000.

[88] E Segal, R Yelensky, and D Koller. Genome-wide discovery of transcriptional modules from dna sequence and gene expression. *Bioinformatics*, 19 Suppl 1(1367-4803):i273–82, 2003.

[89] Eran Segal, Yoseph Barash, Itamar Simon, Nir Friedman, and Daphne Koller. From promoter sequence to expression: a probabilistic framework. In *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*, pages 263–272, New York, NY, USA, 2002. ACM Press.

[90] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–76, 2003.

[91] Roded Sharan, Ivan Ovcharenko, Asa Ben-Hur, and Richard M Karp. Creme: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19 Suppl 1(1367-4803):i283–91, 2003.

[92] Daisuke Shinozaki and Osamu Maruyama. A method for the best model selection for single and paired motifs. In *Genome Informatics*, volume 13, pages 432–433. Universal Academy Press, 2002.

[93] S Sinha and M Tompa. A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol*, 8(1553-0833):344–54, 2000.

[94] Saurabh Sinha. Discriminative motifs. *J Comput Biol*, 10(3-4):599–615, 2003.

[95] Saurabh Sinha, Mathieu Blanchette, and Martin Tompa. Phyme: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5(1):170, 2004.

[96] Saurabh Sinha and Martin Tompa. Ymf: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, 31(13):3586–8, 2003.

[97] Saurabh Sinha, Erik van Nimwegen, and Eric D Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 Suppl 1(1367-4803):i292–301, 2003.

[98] Zhaohui Sun, Jingyi Yang, and Jitender S. Deogun. Misae: A new approach for regulatory motif extraction. In *CSB '04: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04)*, pages 173–181, Washington, DC, USA, 2004. IEEE Computer Society.

[99] K T Takusagawa and D K Gifford. Negative information for motif discovery. In *Pac Symp Biocomput*, pages 360–71, 2004.

[100] G Thijs, M Lescot, K Marchal, S Rombauts, B De Moor, P Rouze, and Y Moreau. A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics*, 17(12):1113–22, 2001.

[101] Gert Thijs, Kathleen Marchal, Magali Lescot, Stephane Rombauts, Bart De Moor, Pierre Rouze, and Yves Moreau. A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, 9(2):447–64, 2002.

[102] William Thompson, Michael J Palumbo, Wyeth W Wasserman, Jun S Liu, and Charles E Lawrence. Decoding human regulatory circuits. *Genome Res*, 14(10A):1967–74, 2004.

[103] William Thompson, Eric C Rouchka, and Charles E Lawrence. Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res*, 31(13):3580–5, 2003.

[104] M Tompa. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Proc Int*

Conf Intell Syst Mol Biol, pages 262–71, Heidelberg, Germany, August 1999.

[105] J van Helden, B Andre, and J Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, 281(5):827–42, 1998.

[106] J van Helden, B Andre, and J Collado-Vides. A web site for the computational analysis of yeast regulatory sequences. *Yeast*, 16(2):177–87, 2000.

[107] J van Helden, A F Rios, and J Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res*, 28(8):1808–18, 2000.

[108] A Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, 15(10):776–84, 1999.

[109] Ting Wang and Gary D Stormo. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–80, 2003.

[110] W W Wasserman and J W Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1):167–81, 1998.

[111] C T Workman and G D Stormo. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. In *Pac Symp Biocomput*, pages 467–78, 2000.

[112] E P Xing, M I Jordan, R M Karp, and S Russell. A hierarchical bayesian markovian model for motifs in biopolymer sequences. In S Becker, S Thrun, and K Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, Cambridge, MA, 2002.

[113] E P Xing, W Wu, and R M Karp. Capturing characteristic structural features for motif detection using a hierarchical bayesian markovian model. *Genome Biol*, 2003.

[114] Eric P Xing and Richard M Karp. Motifprototyper: a bayesian profile model for motif families. *Proc Natl Acad Sci U S A*, 101(29):10523–8, 2004.

[115] Eric P Xing, Wei Wu, Michael I Jordan, and Richard M Karp. Logos: a modular bayesian model for de novo motif detection. *J Bioinform Comput Biol*, 2(1):127–54, 2004.

[116] Xiaoyue Zhao, Haiyan Huang, and Terence P. Speed. Finding short dna motifs using permuted markov models. In *RECOMB '04: Proceedings of the eighth annual international conference on Computational molecular biology*, pages 68–75, New York, NY, USA, 2004. ACM Press.

[117] Qing Zhou and Jun S Liu. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–16, 2004.

[118] Qing Zhou and Wing H Wong. Cismodule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, 101(33):12114–9, 2004.