



FREIE UNIVERSITÄT BERLIN  
FACHBEREICH MATHEMATIK UND INFORMATIK

DISSERTATION ZUR ERLANGUNG DES AKADEMISCHEN GRADES  
DES DOKTORS DER NATURWISSENSCHAFTEN (DR. RER. NAT.)

---

# **Discriminative learning for probabilistic sequence analysis**

---

Jonas MAASKOLA

June 2014



*Erstgutachter:*

**Prof. Dr. Martin VINGRON**

Transcriptional Regulation Group  
Computational Molecular Biology Department  
Max Planck Institute for Molecular Genetics

Honorary Professor of Bioinformatics  
Department of Mathematics and Computer Science  
Freie Universität Berlin

*Zweitgutachter:*

**Prof. Dr. Nikolaus RAJEWSKY**

Laboratory for Systems Biology of Gene Regulatory Elements  
Berlin Institute for Medical Systems Biology  
Max-Delbrück-Centrum für Molekulare Medizin, Berlin-Buch

Professor of Systems Biology  
Charité Medical School  
Humboldt Universität Berlin

Disputation: 16. April 2015



# Contents

<b>Abstract</b>	<b>1</b>
<b>Summary</b>	<b>3</b>
Methodology . . . . .	3
Results and contributions . . . . .	6
<b>I Biological Background</b>	<b>15</b>
<b>1 Systems biology of gene regulation</b>	<b>17</b>
1.1 State of life science research . . . . .	17
1.2 Gene regulatory processes . . . . .	18
1.3 Sequencing technologies . . . . .	22
<b>2 Published computational biology research</b>	<b>25</b>
2.1 List of publications . . . . .	25
2.2 Computational analysis of deep-sequencing data . . . . .	26
2.3 Population genomics of drosophilid miRNAs . . . . .	29
2.4 Small RNAs in <i>C. elegans</i> embryogenesis . . . . .	34
2.5 Small RNAs in <i>S. purpuratus</i> early development . . . . .	42
2.6 ChIP-Sequencing a cell-cycle regulator and a helicase . . . . .	45
2.7 Data curation for post-transcriptional research . . . . .	51
2.8 Splicing regulation . . . . .	51
<b>II Probabilistic Sequence Analysis</b>	<b>53</b>
<b>3 Motif and binding site models</b>	<b>57</b>
3.1 Binding motif models . . . . .	57
3.2 Binding site models . . . . .	64
<b>4 Hidden Markov models</b>	<b>67</b>
4.1 Formal definition . . . . .	67
4.2 Fundamental problems and basic inference algorithms . . . . .	68
4.3 Viterbi algorithm . . . . .	69
4.4 Forward-backward algorithm . . . . .	71

4.5	Scaled forward-backward algorithm . . . . .	73
4.6	Likelihood gradient . . . . .	75
<b>5</b>	<b>Binding site hidden Markov model</b>	<b>81</b>
5.1	An HMM for binding sites . . . . .	81
5.2	Posterior motif occurrence probability . . . . .	82
<b>6</b>	<b>Learning algorithms</b>	<b>85</b>
6.1	Baum-Welch algorithm . . . . .	85
6.2	Gradient learning . . . . .	88
6.3	Complexity . . . . .	89
<b>III</b>	<b>Discriminative Learning for Probabilistic Sequence Analysis</b>	<b>91</b>
<b>7</b>	<b>Statistics for discriminative learning</b>	<b>95</b>
7.1	Contrasts . . . . .	95
7.2	Contingency tables of features across contrasts . . . . .	96
<b>8</b>	<b>Table-based discriminative objective functions</b>	<b>99</b>
8.1	Difference of occurrence frequency . . . . .	99
8.2	Matthews correlation coefficient . . . . .	100
8.3	Fisher's exact test . . . . .	101
8.4	Normalized enrichment score . . . . .	102
8.5	Pearson's $\chi^2$ test for independence . . . . .	102
8.6	Mutual information of condition and occurrence . . . . .	102
<b>9</b>	<b>Probabilistic discriminative objective functions</b>	<b>105</b>
9.1	Difference of log likelihood . . . . .	105
9.2	Multiple model classification . . . . .	106
9.3	Probability of correct classification . . . . .	106
<b>10</b>	<b>Significance of association</b>	<b>109</b>
10.1	Mutual information, likelihood ratio, and $\chi^2$ test . . . . .	109
10.2	Multiple testing correction for motif discovery problems . . . . .	110
10.3	Significance of association significance filtering . . . . .	111
<b>11</b>	<b>Discrete optimization of discriminative objectives</b>	<b>113</b>
11.1	Enumerating residually most discriminative words . . . . .	113
11.2	Identifying discriminative words with degeneracy . . . . .	114
<b>12</b>	<b>Hybrid learning</b>	<b>119</b>
12.1	Signal and context parameters . . . . .	119
12.2	Learning scheme . . . . .	119
12.3	Multi-objective learning . . . . .	120

<b>13 Discovering multiple motifs</b>	<b>123</b>
13.1 Measures of conditional association . . . . .	123
13.2 Discovering multiple motifs . . . . .	124
<b>14 Related work</b>	<b>129</b>
14.1 Overview . . . . .	129
14.2 DREME: Discriminative Regular Expression Motif Elicitation . . . . .	129
14.3 YMF: Yeast Motif Finder . . . . .	130
14.4 CMF: Contrast Motif Finder . . . . .	131
14.5 DME: Discriminating Matrix Enumerator . . . . .	132
14.6 DIPS: Discriminative PWM Search . . . . .	134
14.7 DECOD: Deconvolved Discriminative Motif Discovery . . . . .	136
14.8 MoAn: Motif Annealer . . . . .	137
14.9 DEME: Discriminatively Enhanced Motif Elicitation . . . . .	139
14.10 Dispom . . . . .	140
14.11 FIRE: Finding Informative Regulatory Elements . . . . .	141
14.12 Further motif discovery tools . . . . .	142
<b>IV Empirical Study of Motif Discovery Methodology</b>	<b>143</b>
<b>15 Synthetic test data</b>	<b>147</b>
15.1 Generation of synthetic data . . . . .	147
15.2 Recognizability . . . . .	148
<b>16 Performance metrics</b>	<b>151</b>
16.1 Supervised performance metrics . . . . .	151
16.2 Summarization . . . . .	153
<b>17 Results on synthetic data</b>	<b>155</b>
17.1 Evaluated motif discovery tools . . . . .	155
17.2 Summary performance . . . . .	156
17.3 Comparing signal-only and discriminative learning . . . . .	159
17.4 Discriminative filtering . . . . .	162
<b>18 PUF RNA-binding protein family</b>	<b>163</b>
18.1 Materials . . . . .	164
18.2 Methods . . . . .	165
18.3 Discriminative motifs in PUF RBP data . . . . .	166
18.4 Comparing MICO, MMIE, and ML learning . . . . .	168
18.5 Dilution and word-based analyses . . . . .	168
<b>19 Alternative splicing regulator RBM10</b>	<b>173</b>
19.1 Materials . . . . .	174
19.2 Motif discovery for RBM10 reveals splicing-relevant motifs . . . . .	174
19.3 Previously reported RBM10 motifs not corroborated . . . . .	176

<b>20 Mouse embryonic stem cell transcription factors</b>	<b>179</b>
20.1 Materials and methods . . . . .	179
20.2 Discriminative motifs in ChIP-Seq data . . . . .	180
20.3 Spatial distribution of motif occurrences . . . . .	185
20.4 Contrasting Nanog and Tcf3 against other ChIP-Seq data . . . . .	186
20.5 Comparing results of DREME, FIRE, and Discoverer . . . . .	186
<b>V Discussion</b>	<b>191</b>
<b>21 Supervised motif discovery performance experiments</b>	<b>193</b>
21.1 Influence of parameters varied in synthetic data . . . . .	193
21.2 Generative, signal-only learning . . . . .	195
21.3 Discriminative learning . . . . .	196
21.4 Robustness of hybrid learning . . . . .	198
21.5 Comparison to published motif discovery methods . . . . .	199
<b>22 Sequence motifs in biological data</b>	<b>201</b>
22.1 RIP-Chip and PAR-CLIP data of PUF family RBPs . . . . .	201
22.2 Alternative splicing regulator RBM10 . . . . .	202
22.3 ChIP-Seq data of mouse ESC TFs . . . . .	204
<b>23 Outlook</b>	<b>207</b>
23.1 Global optimization . . . . .	207
23.2 Rank-based learning - Average rank information . . . . .	208
23.3 Faster learning . . . . .	208
23.4 Additional information sources . . . . .	209
23.5 Other applications . . . . .	210
<b>24 Conclusions</b>	<b>211</b>
<b>Appendices</b>	<b>213</b>
<b>A Proof of correctness of the scaling procedure</b>	<b>215</b>
A.1 Correctness for the forward matrix recursion . . . . .	215
A.2 Correctness for the backward matrix recursion . . . . .	216
<b>B Runtime of HMM inference algorithms</b>	<b>217</b>
<b>C Information theory</b>	<b>219</b>
C.1 Communication systems . . . . .	219
C.2 Fundamental quantities of information theory . . . . .	219
<b>D Likelihood ratio, mutual information and <math>\chi^2</math> statistic</b>	<b>223</b>
D.1 The likelihood ratio statistic for goodness of fit . . . . .	223
D.2 G-test . . . . .	225



---

D.3 Likelihood ratio and mutual information . . . . .	225
<b>E Limits of Matthews correlation coefficient</b>	<b>227</b>
<b>F Gradient calculus</b>	<b>229</b>
<b>G Running parameters</b>	<b>231</b>
<b>H Synthetic data experiments</b>	<b>235</b>
<b>I PUF RBP family data</b>	<b>245</b>
<b>J Alternative splicing regulator RBM10</b>	<b>253</b>
<b>K Mouse ESC ChIP-Seq data</b>	<b>261</b>
<b>L Research anecdotes</b>	<b>265</b>
L.1 Motif discovery anecdotes . . . . .	265
L.2 A fishy smell in ChIP-Seq data . . . . .	267
<b>Bibliography</b>	<b>269</b>
<b>Statutory Declaration</b>	<b>299</b>
<b>Zusammenfassung</b>	<b>301</b>

This is revision 66a1c19 of the manuscript, generated July 15, 2015.

# List of Figures

1.1	Biogenesis of an mRNA . . . . .	21
1.2	Structure of 4-thiouridine and 6-thioguanosine . . . . .	23
1.3	ChIP-Seq and PAR-CLIP methodologies . . . . .	24
2.1	Small RNA expression in early development of <i>C. elegans</i> and <i>S. purpuratus</i> . . . . .	37
2.2	E2F3 and HELLS regulate many common targets, most notably MLL1 . . . . .	46
3.1	Graphical probabilistic models for bindings motifs and sites . . . . .	58
3.2	Simple binding site model topologies . . . . .	64
5.1	Default topology of a binding site HMM . . . . .	82
6.1	Conditions enforced by the Moré-Thuente line searching algorithm . . . . .	89
7.1	Contrasts for discriminative sequence analysis . . . . .	95
9.1	The graphical model of MMIE . . . . .	107
11.1	IUPAC generalization and specialization . . . . .	115
11.2	Example run of Plasma on PUM2 data of Hafner et al. (2010) . . . . .	117
13.1	Conditionally independent motifs that are marginally dependent . . . . .	124
13.2	Conditionally dependent motifs that are marginally independent . . . . .	124
13.3	Flow chart of the first part of multiple motif discovery . . . . .	125
13.4	Flow chart of the second part of multiple motif discovery . . . . .	127
14.1	Discrete column types in DME . . . . .	132
16.1	Classification of nucleotides . . . . .	152
16.2	Classification of binding sites . . . . .	153
17.1	Summarized motif discovery performance . . . . .	157
17.2	Motif recognizability and discovery performance on synthetic data: nCC . . . . .	160
18.1	Discriminative word analysis of human PUM1 and PUM2 data . . . . .	171
19.1	Motif occurrences in RBM10 sequences . . . . .	175

---

20.1	Positional distribution of motif occurrences in Tcfcp2l1 ChIP-Seq regions	185
B.1	Two illustrative HMM topologies	218
B.2	Trellis structure of smoothing algorithms on the HMM of figure B.1a	218
B.3	Trellis structure of smoothing algorithms on the HMM of figure B.1b	218
C.1	Schematic diagram of a general communication system	220
C.2	Relationship of information theoretic quantities	221
D.1	Probability density function of the $\chi^2$ distribution	224
H.1	Additional metrics for motif discovery performance	237
H.2	Performance of additional motif discovery methods	238
H.3	Runtime of motif discovery methods on synthetic datasets	239
H.4	Motif recognizability and discovery performance on synthetic data: sAP	240
H.5	Motif recognizability and discovery performance on synthetic data: sSn	241
H.6	Motif recognizability and discovery performance on synthetic data: sPPV	242
H.7	Effect of significance filtering on motif discovery performance	243
I.1	Boxplot of sequence lengths for PUF RBP family data	246
I.2	Number of sequences with UGUAHAUA motif in PUF RBP family data	247
I.3	8mers in PUM1 data of Morris, Mukherjee, and Keene (2008)	248
I.4	8mers in PUM1 data of Galgano et al. (2008)	249
I.5	8mers in PUM2 data of Galgano et al. (2008)	250
I.6	8mers in PUM2 data of Hafner et al. (2010)	251
L.1	Bug in random number generation of MoAn	266

# List of Tables

3.1	The IUPAC code for nucleic acids . . . . .	59
4.1	Runtime complexity and inter-dependence of inference algorithms. . . . .	79
7.1	$2 \times 2$ contingency table of sequences with and without a feature . . . . .	97
7.2	$k \times 2$ contingency table of sequences with and without a feature . . . . .	97
8.1	Examples of discriminative measures . . . . .	100
14.1	Related methods for discriminative motif discovery . . . . .	129
14.2	Default parameters of DME . . . . .	134
15.1	Parameters for the generation of synthetic sequence data . . . . .	148
18.1	PUF RBP family discriminative motif analysis . . . . .	166
18.2	Motif discovery results for datasets of PUF RBP family members . . . . .	167
18.3	Dilution analysis of human PUM1 and PUM2 data for MICO and MMIE . . . . .	169
19.1	Motifs discovered in RBM10 PAR-CLIP data . . . . .	174
19.2	RBM10 motifs of Bechara et al. in PAR-CLIP sequences of Y. Wang et al. . . . .	176
20.1	Discriminative motif analysis of mouse ChIP-Seq data . . . . .	180
20.2	Longer motif variants of Tcfcp2l1 . . . . .	183
20.3	Motifs discriminating Nanog and Tcf3 data from other ChIP-Seq data . . . . .	187
20.4	Discover, DREME, and FIRE applied to Oct4 ChIP-Seq data . . . . .	188
H.1	Runtime of excluded motif discovery methods . . . . .	236
H.2	Motif discovery performance . . . . .	236
J.1	RBM10 motifs of Bechara et al. in exonic sequences of Y. Wang et al. . . . .	253
J.2	RBM10 motifs of Bechara et al. in intronic sequences of Y. Wang et al. . . . .	256
J.3	RBM10 motifs of Inoue et al. in exonic sequences of Y. Wang et al. . . . .	258
J.4	RBM10 motifs of Inoue et al. in intronic sequences of Y. Wang et al. . . . .	259
K.1	Positional distribution of motif occurrences in ChIP-Seq regions . . . . .	261

# List of Algorithms

1	Baum-Welch expectation step . . . . .	87
2	Baum-Welch maximization step . . . . .	87
3	Baum-Welch . . . . .	88
4	Core-nmer analysis . . . . .	114
5	Most discriminative IUPAC motif . . . . .	115
6	Synthetic sequence data generation . . . . .	148

# List of Acronyms

- ANOPS** arbitrary number of occurrences per sequence. 64–66
- BW** Baum-Welch. 85–90, 156–160, 196
- ChIP** chromatin IP. x, 22, 46, 267
- ChIP-Chip** ChIP followed by microarray quantification. 22, 184, 267
- ChIP-Seq** ChIP followed by sequencing. 5, 11, 12, 22, 24, 45, 46, 49, 50, 142, 145, 179, 180, 182–186, 201, 204, 205, 208, 211, 261, 265–267
- circRNA** circular RNA. 20
- CLIP-Seq** UV crosslinking and IP followed by sequencing, a.k.a. HITS-CLIP. 176, 202, 253, 256
- cMI** conditional mutual information. 5, 6, 123–126, 211, 222
- cmICO** conditional mutual information of condition and motif occurrence after accounting for previously identified motifs. 6, 123–125
- DBP** DNA-binding protein. 3, 5, 6, 19, 22, 55, 96, 114, 116, 163
- DFREQ** difference of relative frequency. 99, 100, 113, 114, 129, 136, 155, 156, 199
- DLOGL** difference of log likelihood. 7, 105, 106, 129, 133, 155, 156
- DMD** discriminative motif discovery. 4, 6–9, 11, 25, 52, 93, 129, 137, 141, 142, 145, 156, 158, 163, 165, 170, 173, 174, 179, 180, 194, 197, 199, 201, 202, 205–207, 210, 211
- DNA** deoxyribonucleic acid.
- dsRNA** double stranded RNA. 42
- EM** expectation-maximization. 85
- ESC** embryonic stem cell. 11, 145, 179, 182, 185, 201, 211, 261
- ESE** exonic splicing enhancer. 10, 11, 174, 176, 202–204
- FDR** false discovery rate. 132

- FWER** family-wise error rate. 109
- HMM** hidden Markov model. 4, 5, 7, 55, 59, 65–73, 77, 79, 81, 82, 85, 86, 90, 93, 106–108, 110, 113, 114, 119, 120, 124, 125, 127, 129, 133–135, 155–158, 165, 167–170, 180, 184, 195, 199, 211, 217, 218
- IC** information content. 63, 64, 133, 147, 159, 161, 168, 170, 184, 193–195, 197, 198, 201, 202
- IP** immunoprecipitation. x, xii, 9, 22, 267
- iPSC** induced pluripotent stem cell. 179, 182
- IUPAC** International Union of Pure and Applied Chemistry; table 3.1: code for nucleic acids.
- KL** Kullback-Leibler. 63, 103, 221
- lncRNA** long non-coding RNA. 19
- MCC** Matthews correlation coefficient. 100, 101, 113, 114, 152, 155, 156
- MCMC** Monte-Carlo Markov chain. 8, 207
- MD** motif discovery. 3–12, 57, 64, 95, 96, 99, 109, 110, 142, 145, 147, 151, 155, 156, 158, 159, 161, 162, 193–201, 204–209, 211, 231–233, 235, 265, 266
- MFE** minimum free energy. 33
- MI** mutual information. 123–126
- MICO** mutual information of condition and motif occurrence. 4, 5, 9, 10, 103, 113, 114, 117, 129, 141, 156–160, 162, 164–171, 174, 176, 180, 184, 196, 201, 205, 208, 209, 211, 229, 248–251
- miRNA** microRNA. 20, 21, 26–28, 30–35, 37, 39–44, 51, 163
- ML** maximum likelihood. 85, 105, 164, 195
- MMIE** maximum mutual information estimation. 106–108, 129, 156, 164, 165, 167–170
- mRNA** messenger RNA. xii, 4, 9, 10, 18–22, 33, 39–44, 51, 163, 203
- MT** multiple testing. 109, 111, 114, 167, 168
- NCBI** National Center for Biotechnology Information, U.S. National Library of Medicine, located at the National Institutes of Health. 17, 28, 267
- nCC** nucleotide correlation coefficient. 152–154, 156–161, 235
- ncRNA** non-coding RNA. 19, 20, 35, 37, 39, 42
- OOPS** one occurrence per sequence. 64, 65, 106

- PAR-CLIP** photoactivatable-ribonucleoside-enhanced crosslinking and IP, followed by sequencing. 9, 10, 22–24, 51, 52, 117, 145, 164–170, 174–177, 201–204, 208, 209, 211, 245, 253–259
- PCR** polymerase chain reaction. 22
- piRNA** PIWI-interacting RNA. 19, 20, 37, 42–44
- PRE** PUF recognition element, UGUAHAUA. 166, 201
- PSCM** position-specific count matrix. 59, 60, 130, 134
- PSFM** position-specific frequency matrix. 60, 62, 63, 132, 136, 139, 140, 194
- PSSM** position-specific score matrix. 60, 63, 65, 193–195, 207
- PWM** position weight matrix. 6, 60, 129, 134–136, 139, 140, 188
- RBP** RNA-binding protein. 3–7, 9, 10, 19–23, 51, 55, 96, 145, 163–168, 171, 201, 203, 208, 211, 245–247
- regex** regular expression. 5–8, 58, 59, 110, 114, 116, 117, 129, 130, 139, 140, 155–157, 159, 174, 183, 188, 199, 206, 237, 243, 245
- RIP-Chip** DNA microarray quantification of co-IPed mRNA. 9, 10, 22, 145, 163–170, 201, 208, 209, 211, 245
- RISC** RNA-induced silencing complex. 20, 34
- RNA** ribonucleic acid.
- RNAi** RNA interference. 37, 40, 42
- RNA-Seq** RNA sequencing. 10
- rRNA** ribosomal RNA. 20, 39–41, 43
- sAP** average site performance. 153, 154, 158, 161, 235, 240
- sF<sub>1</sub>** site F<sub>1</sub> score. 153, 235
- siRNA** small interfering RNA. 37, 40–42
- snoRNA** small nucleolar RNA. 20, 39
- SNP** single nucleotide polymorphism. 31, 33, 34
- snRNA** small nuclear RNA. 20, 39
- snRNP** small nuclear ribonucleic particles, involved in forming the spliceosome. 10
- sPPV** site positive predictive value. 152, 153, 158, 161, 168, 194, 235, 242
- sSn** site sensitivity. 152, 153, 158, 161, 168, 193, 194, 235, 241



- TF** transcription factor. 11, 19, 22, 145, 179, 183, 184, 201, 211
- TFBS** transcription factor binding site. 19, 21
- tRNA** transfer RNA. 20, 39–41, 43
- TSS** transcription start site. 19, 21, 46, 48–50, 140
- ZOOPS** zero or one occurrence per sequence. 64–66, 106, 140

# Nomenclature

$X$	Random variable
$x$	Realization of a random variable $X$
$\mathcal{X}$	Alphabet of $X$ , with $X \in \mathcal{X}$ ; values that a random variable $X$ can assume
$\theta$	Parameter
$\mathbb{P}$	Probability, e.g. $\mathbb{P}(X = x \theta)$ , the probability that the random variable $X$ assumes the value $x$ given the parameter $\theta$
$\mathbb{L}$	Likelihood, $\mathbb{L}(\theta X) = \mathbb{P}(X \theta)$ , where $X$ are data, and $\theta$ parameters
$\mathbb{E}$	Expected value, e.g. $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}(x)$
$\mathbb{H}$	Information entropy, see equation (C.1), page 220
$\mathbb{I}$	Mutual information, see equation (C.4), page 221

Variables in bold font, e.g.  $\mathbf{X}$ , are vector or matrix valued.

Given a function  $f(\boldsymbol{\theta})$  of a vector valued argument  $\boldsymbol{\theta} = (\theta_i)_{i=1, \dots, n}$ , we denote the gradient of  $f$  with respect to the components of  $\boldsymbol{\theta}$  by

$$\nabla f(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left( \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_i} \right)_{i=1, \dots, n}.$$

*The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning.*

Claude E. Shannon, 1948  
A Mathematical Theory of Communication

*What lies at the heart of every living thing is not a fire, not warm breath, not a 'spark of life'. It is information, words, instructions.*

*If you want a metaphor, don't think of fires and sparks and breath. Think, instead, of a billion discrete, digital characters carved in tablets of crystal.*

*If you want to understand life, don't think about vibrant, throbbing gels and oozes, think about information technology.*

Richard Dawkins, 1986  
The Blind Watchmaker

*Indeed, the very idea is somewhat baffling: If there is a code, then who invented it? What kinds of messages are written in it? Who writes them? Who reads them?*

James Gleick, 2011  
The Information: A History, a Theory, a Flood. Chapter 10: Life's own code

*Perhaps I can best describe my experience of doing mathematics in terms of a journey through a dark unexplored mansion. You enter the first room of the mansion and it's completely dark. You stumble around bumping into the furniture, but gradually you learn where each piece of furniture is. Finally, after six months or so, you find the light switch, you turn it on, and suddenly it's all illuminated. You can see exactly where you were. Then you move into the next room and spend another six months in the dark. So each of these breakthroughs, while sometimes they're momentary, sometimes over a period of a day or two, they are the culmination of—and couldn't exist without—the many months of stumbling around in the dark that proceed them.*

Andrew Wiles, 2000  
NOVA interview

*It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.*

Albert Einstein, June 10, 1933  
“On the Method of Theoretical Physics,” the Herbert Spencer Lecture, Oxford

*Il semble que la perfection soit atteinte non quand il n'y a plus rien à ajouter, mais quand il n'y a plus rien à retrancher.*

Antoine de Saint-Exupéry, 1939  
Terre des hommes. Chapter III: L'avion

*Everything should be made as simple as possible, but not simpler.*

Common aphorism, simplifying the above Einstein quotation

# Acknowledgments

I would like to thank my dear office colleagues and fellow computational biologists Marvin Jens, Azra Krek, Marc Friedländer, Dominik Grün, Sebastian Mackowiak, Andranik Ivanov, and Anna Elefsinioti for many fruitful discussions. Among the wet-lab scientists I had the pleasure of working with, I would particularly like to thank Marlon Stoeckius. Further members of the Rajewsky group that I had insightful exchanges with include Catherine Adamidi, Pinar Önal, Svetlana Lebedeva, Kathrin Theil, Toshiaki Kogame, Sebastian Memczak, and Jordi Solana Garcia.

Other people with whom I enjoyed discussing various topics addressed in this dissertation include Matthias Heinig, Andreas Döring, and Aysam Gürler. I thank Benedikt Obermayer for feedback on the manuscript of the publication on which this dissertation is based.

The information theory and information geometry journal club I co-organized with Kawe Yoocef and Manuela Benary was, although at times challenging, at the same time highly inspirational.

The interaction with Markus Landthaler was of great influence on the choice of my dissertation topic. I am grateful to Ulrike Ziebold for providing feedback for the sections on ChIP-Seq analysis presented in this thesis.

I am thankful for the supervision and support provided through my advisors Nikolaus Rajewsky and Martin Vingron.

I want to thank the Deutsche Forschungsgemeinschaft for a stipend that financed the first part of my PhD research.

Finally, my deepest gratitude needs to be expressed for the support my parents provided during my work on this dissertation. Without, I could not have hoped to approach the level of cogency I wanted to achieve.



# Abstract

This dissertation presents a study of discriminative learning techniques for probabilistic sequence analysis that find application in pattern discovery of binding sites in nucleic acid sequences. Sets of positive and negative example sequences define contrasts that are mined for sequence motifs whose occurrence frequency varies between the sets. A discriminative motif discovery method based on hidden Markov models (HMMs) is described that allows choice of different objective functions, two of which are used for the first time for motif finding with HMMs: mutual information of condition and motif occurrence (MICO), and Matthews correlation coefficient.

We perform an extensive and systematic comparison of motif discovery performance of our method and numerous published tools. Using MICO or several other of the implemented objective functions, our method's performance exceeds that of all other tools. MICO is also the most generally useful discriminative objective function, as it is applicable both to the analysis of probabilistic as well as discrete binding motif models, can leverage contrasts of more than two conditions, and provides natural extensions to quantify conditional association that are used to build models of multiple motifs.

The investigation concludes with several case studies comprising 30 datasets from transcriptome-scale technologies—ChIP-Seq, RIP-ChIP, and PAR-CLIP—of embryonic stem cell transcription factors and of RNA-binding proteins. The case studies demonstrate practicality and utility of the method, and validate it by reproducing motifs of well-studied proteins. In addition, they provide novel insights by connecting previously known splicing-relevant motifs to an alternative splicing regulator.

The presented motif discovery method scales to large data sizes, makes use of available repeat experiments for increased statistical power, and aside from binary contrasts also more complex data configurations can be utilized. It is implemented in the open source software Discover (portmanteau of *discriminative* and *discover*), and is available from <https://github.com/maaskola/discover>.





# Summary

This chapter gives a summary of the contents of this thesis. It summarizes the methodology, concentrating on the reasons of design choices rather than their details. Also summarized are the main findings whose details are presented throughout the manuscript. Where appropriate, pointers are provided to figures and tables in later chapters.

**Objective** The objective of this thesis is to describe a novel method for motif discovery (MD) and to study its properties by applying it to various datasets. The method is designed for the discovery of patterns in nucleic acid sequences (motifs) that serve as recognition signals for nucleic acid binding proteins.

**Background and relevance** Nucleic acid binding proteins play central roles in regulating important biological processes (chapter 1). In particular, DNA-binding proteins (DBPs) act as epigenetic and transcriptional regulators. They regulate accessibility to the transcriptional apparatus and synthesis rates of genes, respectively, by influencing the structural and biochemical properties of chromatin, and by determining the composition of protein complexes occupying gene enhancer and promoter regions. RNA-binding proteins (RBPs), together with regulatory RNAs, act as post-transcriptional regulators, carrying out the processes leading up to protein synthesis, and so direct protein production rates. Thus, DBPs and RBPs jointly execute gene regulatory programs, and are informed by signals encoded in sequences of the genome. As these sequence motifs encode regulatory instructions, their study continues to be of central relevance to understanding cellular programs. Consequently, and because this study involves the analysis of fairly large datasets, methods to decipher sequence motifs are an important area for the application of computational methods in biology.

## Methodology

Various models are available to describe properties of sequence motifs (section 3.1). These include discrete and non-discrete models, with complementing properties. Learning discrete models is computationally less demanding than for non-discrete models, while non-discrete sequence motif models, e.g. biophysically motivated probabilistic ones, are more accordant with principles of statistical mechanics.

It is for this reason that many MD methods, including the one presented here, consist of two phases. An initial phase discovers discrete models, that are subsequently refined

in a second phase to yield parameters for a smooth model.

MD methods generally utilize a set of example sequences (signal sequences) which are known to contain the sequence motif of interest. Aside from the motifs themselves, these sequences typically contain context that is unrelated to recognition by the nucleic acid binding proteins. Thus, MD methods have to model both the motifs and the surrounding context (section 3.2). Although not directly bound by the proteins, the context can also be of importance. For example, it may contain signals for co-factors or competing factors, or it may influence binding of RBPs through accessibility of motifs by favoring or disfavoring secondary structures of mRNAs.

### **Discriminative motif discovery**

Aside from signal sequences, MD methods can utilize sequences which are assumed to contain no, or few, occurrences of the sequence motifs. Some MD methods use such negative example sequences (control sequences) only to learn models of the sequence context, or, after learning, to estimate the statistical significance of motifs discovered from the signal sequences. Discriminative motif discovery (DMD) methods use control sequences for additional purposes. In particular, they search for such motifs that occur more frequently in the signal sequences than in the control sequences.

**Dissertation subject matter** The DMD method I developed and describe in this thesis is named Discover (portmanteau of *discriminative* and *discover*). It uses hidden Markov models (HMMs) (chapter 4) to model binding sites (chapter 5). HMMs offer efficient learning algorithms (chapter 6), using re-estimation- (section 6.1) and gradient-based methods (section 6.2).

**Contrasts and objective functions** DMD methods can leverage different kinds of contrasts (chapter 7 and figure 7.1), including simple binary contrasts, but also more general ones. Various measures can be used as discriminative objective functions to measure the association that a motif has with the conditions of contrasts. Contingency tables of counts of motif occurrences throughout the conditions of a contrast (section 7.2) form the basis for some measures (chapter 8), others involve probabilistic modelling of the contrast's conditions (chapter 9). Discover implements multiple discriminative objective functions, thereby enabling experiments to study the consequences of different objective function choices. One of the objective functions has not previously been used for MD, and others have previously only been used for optimization of discrete motif models.

**Mutual information and significance of association** Due to its versatility, an information theoretic discriminative objective function, mutual information of condition and motif occurrence (MICO), is of particular interest in this study. MICO is related to the log-likelihood ratio of the independence hypothesis for events tallied in contingency tables, and, in consequence, to the  $\chi^2$  test for independence. Discover uses this connection (chapter 10) to efficiently compute significance of association, where an earlier MICO-based DMD method resorted to computationally costly sampling-based significance cal-

culations. Unlike some other objective functions, MICO is not limited only to the analysis of binary contrasts. Another distinctive advantage that MICO has over alternative discriminative objective functions is the availability of natural extensions to quantify conditional association via the conditional mutual information (cMI), which is used by Discover when discovering multiple motifs (see below).

**Seeding: word-based motif finding** As the HMM parameter learning methods used by Discover are local search methods, the choice of initial parameters is of influence. To this end the method described here initially finds suitable word-based motifs (chapter 11). This word-based MD method is named Plasma and is automatically used by Discover to determine starting points, but it can also be used independently to allow manual experimentation. It is a fast, progressive algorithm that allows to optimize IUPAC regular expression (regex) based motifs for many of the discriminative objective functions available in Discover.

**Motif discovery with Discover** The general procedure for MD with Discover is as follows. First, Plasma is used to discover IUPAC regex based motifs that are discriminative for a given contrast. The user selects the length range for which to discover regex motifs and the number of motifs to find per length. Then, for each discovered regex motif, an HMM is initialized by Discover. Each HMM is then optimized using a combination of re-estimation- and gradient-based learning techniques (chapter 12). After learning HMM parameters, either the best is selected according to significance of association, or multiple motifs may be combined (see below).

Both Plasma and Discover can be applied to the analysis of RBP and DBP datasets, by optionally also considering motif occurrences on the reverse complementary strand.

When repeat experiments are available, multiple analyses can be performed, and their results compared to reveal robustly identified motifs. Alternatively, both Plasma and Discover can utilize multiple contrasts to find motifs that are jointly discriminative for the specified contrasts. The benefit offered by joint analyses is increased statistical power.

**Discovering multiple motifs** It is frequently useful to find more than the single highest-scoring motif. In particular, when the cognate motif of a factor is less enriched than other, more recognizable motifs, it may be necessary to consider sub-maximally scoring motifs. Also, e.g. ChIP-Seq<sup>1</sup> data often contain motifs of associated co-factors.

For this purpose the Discover framework offers a MICO-based procedure (chapter 13) designed to yield a non-redundant set of motifs with maximal discrimination between the conditions. It first finds seeds and independently optimizes HMMs for each, selecting the best according to MICO-based *p*-value (figure 13.3). Then, progressively more motifs are added that (a) have sufficient residual discriminatory contribution after accounting for previously accepted motifs and (b) are not redundant with previously accepted motifs (figure 13.4).

These two conditions are ensured by filtering based on conditional mutual information (cMI) in two ways. We determine (I) cMI of conditions of the contrast and occurrences

<sup>1</sup>See section 1.3 and figure 1.3a.

of the newly added motif given occurrences of previously accepted motifs (cMICO), and (II) cMI between occurrences of new and previous motifs given the conditions of the contrast (motif pair cMI). cMICO quantifies the discriminatory contribution of the new motif after accounting for previous ones, while motif pair cMI quantifies association between occurrences of the new and previous motifs.

**Related work** A number of DMD methods have previously been published (chapter 14). These methods use either IUPAC regex based or position weight matrix (PWM) based motif models. They employ a variety of different objective functions that are maximized with different optimization procedures. Consequently, in comparing MD performance of these different tools, the influences of motif models, objective functions, optimization procedures, and implementation details are confounded.

## Results and contributions

The second part of the thesis is an empirical study of MD methodology. Performance of MD methods is studied on synthetic data, and real data of DBPs and RBPs are analyzed to glean insights into the biology of nucleic acid binding proteins.

### Synthetic data experiments

Synthetic data allows controlled experiments to systematically study performance of MD methods. Aside from Discover, several published discriminative and non-discriminative MD methods are considered. First, synthetic data is generated. Then, MD methods are applied to discover motifs, and predict their occurrences. Finally, the methods' performance is quantified.

**Parameters varied in synthetic data** To allow for supervised MD performance evaluation, multiple synthetic datasets are constructed (chapter 15). The synthetic datasets vary parameters with influence on MD performance. These include length and number of sequences, information content (IC) and implantation frequency of signal (and decoy) motifs. Parameter values are varied in a combinatorial manner, yielding experiments consisting of thousands of pairs of sequence sets.

**Quantifying motif discovery performance** The data construction process involves generating motifs and implanting occurrences of them into artificial or real sequences. With the knowledge where the motif occurrences are implanted during data generation, it is possible to classify motif occurrences predicted by MD methods as true or false positives, and predicted non-motif positions as true or false negatives (figures 16.1 and 16.2). Consequently, MD performance can be quantified with supervised metrics (chapter 16).

**Motif discovery tools selected for evaluation** The original intention was to evaluate on the synthetic data all previously published DMD methods, as well as a selection of

classical non-discriminative MD methods. However, three of the published DMD methods (Dispom, DIPS, and DEME) proved impractically slow (see below), and were therefore excluded. Thus, the comparison includes six published DMD methods, two non-discriminative MD methods, as well as my HMM-based method Discover with several different objective functions. Also included in the comparison is Plasma, the IUPAC regex based seed finding method of Discover.

We exemplarily determined the runtimes of the three excluded methods on one single pair of signal and control sequence sets (table H.1), and found them to be more than 1000 times higher than Discover's: the dataset was analyzed in about two minutes by Discover; Dispom needed 40 hours, DIPS more than 25 days, and DEME did not finish in 74 days.

**Recognizability: paragon of motif discovery** A limit for the maximally achievable MD performance is given by how well implanted motif occurrences can be recognized by the true model. As the generation process defines the true models for the synthetic data, they are also available to predict motif occurrences in the generated sequences. Here, we refer to the motif occurrence prediction performance of the true models as motif recognizability. In the controlled synthetic data experiments, recognizability thus serves as a reference for the performance of MD tools.

### Motif discovery performance

Discover is found to yield the highest MD performance of the considered methods (chapter 17, figure 17.1, and table H.2). Using MICO as objective function, Discover consistently achieves MD performance exceeding 96% of motif recognizability. Two published DMD methods, DREME and MoAn, also show good MD performance, albeit lower than that of Discover. They respectively achieve MD performance at 83–90% and 85–91% of motif recognizability. The other published methods yield much lower performance, not exceeding 75% of motif recognizability.

At 90–96% of motif recognizability the MD performance of Plasma without subsequent HMM optimization by Discover is lower than Discover's but also surpasses that of most previously published DMD methods.

**Additional methods** The synthetic data experiments were extended to include several additional MD methods, as well as multiple parameter settings for some methods (figure H.2).

The MD performance of the different discriminative objective functions implemented in Discover was evaluated on the synthetic datasets. With one exception<sup>2</sup>, the discriminative objective functions yield comparable MD performance, superior to that of other methods.

An initial evaluation of MD performance used the then-current version 4.7.0 of DREME, which did not support single-strand MD that is appropriate for the analysis of RBP data as

<sup>2</sup>The difference of log likelihood (DLOGL) objective function performs worse.

simulated in the synthetic data. The current version 4.9.0 fixed this, but MD performance did not improve much (figure H.2 and table H.2).

MoAn is a Monte-Carlo Markov chain (MCMC) sampling based MD method. Such methods allow global optimization of parameters, but are computationally expensive. By default, MoAn performs 30 million iterations. This number of iterations proved too slow, and MoAn could only be included in the comparison by performing fewer iterations<sup>3</sup>. We could evaluate MoAn's MD performance with the default number of iterations only on part of the synthetic datasets. This further increased MoAn's MD performance (figure H.2 and table H.2), yet not to the level of Discoverer's.

### **Motif discovery runtime**

There is considerable variation in the time that the different methods need to analyze the synthetic datasets. The runtimes varied more than 360 fold between the fastest and slowest methods (figure H.3).

The fastest method was Plasma, the IUPAC regex based seeding method of Discoverer. It required 2.6 hours to process all synthetic data experiments. DREME and Discoverer respectively took 3.4 and 6.5 times as long as Plasma, and were the third and fourth fastest methods. MoAn, in spite of the reduced number of iterations, still required about 960 hours.

### **Signal-only and discriminative motif discovery**

Motif recognizability and the performance of signal-only and discriminative MD are further analyzed as functions of individual variates realized in the synthetic data (section 17.3 and figures 17.2 and H.4 to H.6). This reveals how the individual variates influence recognizability and MD performance, and how they contribute to the inferior MD performance of signal-only learning.

### **Discriminative filtering**

We compare the performance of signal-only and discriminative MD with and without subsequent filtering based on discriminative significance (section 17.4 and figure H.7). This shows that MD performance of signal-only learning increases significantly, while discriminative MD performance is only slightly affected by subsequent discriminative filtering.

### **Conclusions**

Of the considered methods, Discoverer, the DMD method I developed and describe in this thesis, yields the highest MD performance. Two published DMD methods, DREME and MoAn, also perform well, with MoAn slightly better than DREME. Plasma, the seeding method of Discoverer, achieves MD performance higher than that of DREME, and comparable to that of MoAn. Plasma is also the fastest of the considered methods, while DREME

---

<sup>3</sup>A tenth of the default number of iterations was performed.

and Discover respectively took 3.4 and 6.5 times as long as Plasma, and MoAn being much slower.

Due to its speed and second-best MD performance, Plasma is thus suited for fast, initial analyses of MD datasets. Yielding the best MD results while still being reasonably fast, Discover is recommended for more definitive analyses.

## Post-transcriptional regulation: PUF RBP family

The PUF proteins are a wide-spread eukaryotic family of conserved post-transcriptional regulators. Their RNA binding properties have been well studied, and for this reason the PUF family of RBPs poses an excellent opportunity to demonstrate applicability of the presented MD methodology to real biological data (chapter 18).

### Materials

**Data** PUF family RBPs in different species are considered (section 18.1), including *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*. DNA microarray quantification of co-IPed mRNA (RIP-Chip) was used to define targets of Puf1, Puf2, Puf3, Puf4, and Puf5 in yeast, of the worm homolog FBF-1, of Pumilio in adult fly ovaries, and of human PUM1 and PUM2 from HeLa S3 cells. Additionally, PAR-CLIP<sup>4</sup> data of human PUM2 from HEK293 cells is analyzed.

**Contrasts** Due to the heterogeneity of the PUF family datasets, contrasts for DMD were set up in different ways, including comparison of bound genes versus unbound ones, of bound genes versus the genomic complement, of multiple groups of genes ranked by binding evidence, as well as the comparison of signal to shuffled sequences.

### Methods and results

DMD was performed with Discover (section 18.2), and essentially recovered the known sequence specificities of the individual members of this family (section 18.3 and table 18.1).

**Motifs from signal-only, likelihood-based learning and from discriminative learning** We also studied the results of applying two discriminative objective functions, among them MICO, as well as signal-only, generative learning (section 18.4). The results of using discriminative objective functions agreed, but signal-only learning using likelihood as objective function did not yield useful results.

**Higher spatial resolution of PAR-CLIP confirms relevance of motif variants** The analysis results for the PUM2 motif differed on the second half of the motif between the PAR-CLIP data and the two RIP-Chip datasets. We investigated this further (section 18.5). The PAR-CLIP sequences are much shorter than the full 3'UTR sequences used for RIP-Chip analysis (figure I.1). We reasoned that shorter sequence lengths of PAR-CLIP sequences allow to recognize more diverse, and less strongly bound variants of the motif. Indeed, diluting the

<sup>4</sup>See section 1.3 and figures 1.2 and 1.3b.



PAR-CLIP sequences by adding increasingly longer flanks and subsequently performing MD produced motifs agreeing with the results of RIP-Chip analysis (table 18.3). Confirming these results, scatter-plots of word frequencies (figures I.3 to I.6), as well as a progressive, informative word mining algorithm (figure 18.1) revealed that the longer sequences of RIP-Chip compress observable word frequency variability, and that, conversely, the higher spatial resolution of binding sites of PAR-CLIP data yields a more finely resolved picture of the binding site spectrum of PUM2.

### Conclusions

Using MICO as objective function and the corresponding length corrected *p*-values, the Discover framework successfully recovered the known motifs, automatically yielding the individual family members' motif length preferences. Furthermore, we found that the finer spatial resolution of PUM2 PAR-CLIP data allows to uncover the relevance of weak-affinity variants that do not conform to the classic PUF recognition element, UGUAHUA.

### Alternative splicing regulation: RBM10

Another RBP that is studied in this thesis is the alternative splicing regulator RBM10 (chapter 19), mutations of which are known to be associated with congenital diseases and have been found in tumors. The molecular mechanism of how mutations in this splicing factor contribute to pathology remained elusive until a recent publication<sup>5</sup> revealed by analysis of RNA-Seq and PAR-CLIP data that RBM10 mediates exon skipping.

#### Discriminative motif discovery with RBM10 PAR-CLIP data

Here we perform a motif analysis for RBM10 by mining PAR-CLIP data (section 19.1) for sequence motifs using the MICO-based multiple motif framework of Discover (section 19.2). The analysis uses PAR-CLIP cross-link centered regions, and is performed separately for sequences whose central position lies within exons or introns. The discriminative motif analysis use shuffled sequences as contrasts. Repeat experiments are available, and are jointly analyzed to increase statistical power.

#### Known splicing motifs discovered for RBM10

The analyses reveal motifs (table 19.1) previously implicated in splicing regulation, whose relevance for the alternative splicing factor RBM10 was not well appreciated. The exonic sequences exhibit a purine-rich motif, known as exonic splicing enhancer (ESE) signal, which is known to be bound by the splicing factor SFRS1 and by eIF4AIII, a member of the exon-junction complex and involved in nonsense-mediated RNA decay.

The intronic sequences contain a pyrimidine-rich motif that resembles the signal of the polypyrimidine tract, another well known splicing signal. RBM5, a related splicing factor, is known to compete for binding to the polypyrimidine tract with U2AF65 which is required for the binding of U2 snRNP to the pre-mRNA branch site.

---

<sup>5</sup>I co-authored this publication.



The analysis reveal furthermore a pyrimidine-rich motif in the exonic sequences that is reverse-complementary to the ESE motif, as well as a palindromic, previously uncharacterized motif in the intronic sequences.

### **Most published RBM10 motifs un-corroborated by PAR-CLIP data**

Two recent publication reported many<sup>6</sup> short words to be bound by RBM10. We inspected whether these are also enriched in the PAR-CLIP data analyzed here, but did not find corroborating evidence for most of them (>85%), while the minority that are enriched are consistent with our own analyses. As the validation rate of the previously reported motifs is so low, it seems justified to stress that our findings are first in underlining the central importance for RBM10 binding of the motifs discovered by us.

### **Conclusions**

Intriguingly, the polypyrimidine tract binding protein, PTB, has been reported to bind to the double stranded region of a secondary structure motif in the form of a hairpin whose one arm consists of pyrimidine-rich sequence, while the other consists of purine-rich sequence. It is conceivable, that similar secondary structure might also be of importance to the regulation exerted by RBM10, which could either favor or disfavor the formation of such hairpins and influence splicing through this mechanism.

In summary, based on the presented motif analyses, two mechanisms might be responsible for the reported exon-skipping mediated by RBM10: (I) competition with splicing enhancers for the ESE motif, and (II) competition with U2AF65 for binding to the polypyrimidine tract.

### **Transcriptional regulation: mouse ESC TF ChIP-Seq**

The empirical study of MD methodology concludes with an analysis of sequence binding motifs of mouse embryonic stem cell (ESC) transcription factors (TFs) (chapter 20). We consider 17 ChIP-Seq datasets of 14 different sequence specific TFs coming from two different studies (section 20.1). DMD is performed with Discover in multiple motif mode, using sequences of 101 nt centered on midpoints of ChIP-Seq regions as signal, and shuffles as controls.

**Motif discovery reveals known motifs** One or more motifs are discovered in each of the analyzed datasets (section 20.2 and table 20.1). In total 44 motifs are discovered. Database searches reveal 40 of these 44 motifs to be bound by previously characterized factors. Many motifs are discovered multiple times in different datasets, and in these cases the motifs are quantitatively highly consistent.

The cognate motifs are found in 14 of 17 datasets, and in 12 of these they are the top-ranking motifs. For two of the cases in which the purported cognate motifs are not discovered, our findings agree with literature results that dispute these motifs' veracity. A simple cross-contrasting approach (section 20.4 and table 20.3) directly reveals the cognate

<sup>6</sup>They list nearly 10% of all 5mers.

motifs for the other cases in which they are not discovered or in which they are discovered but not as top-ranking.

**Co-discovered motifs are those of known co-factors** In all datasets where—aside from the cognate motifs—further known motifs are co-discovered, there is evidence supporting the functional relevance of jointly discovered motifs. This includes overlapping sets of ChIP-Seq target genes, mutual transcriptional activation, shared biological functions, and physical interactions.

**Spatial distribution of motif occurrences** Inspection of the positional distribution of motif occurrences relative to the ChIP-Seq regions' midpoints is not used in our MD approach, and thus serves as orthogonal evidence to corroborate the discovered motifs (section 20.3, figure 20.1, and table K.1). Cognate motifs are strongly peaked around the centers. Physically interacting co-factor motifs are also—yet less strongly—peaked around centers. One of the four discovered but unidentified motifs exhibits a centrally peaked distribution providing corroboration for its relevance as true motif for an as-yet unknown co-factor. For the other three discovered unidentified motifs, broader or more uniform spatial occurrence distributions provide, respectively, less or no further corroboration.

**Oct4 motifs of Discover, DREME, and FIRE** Exemplarily, we compare the results of two published MD tools, DREME and FIRE, with ours on one particular dataset (section 20.5 and table 20.4). For this purpose we generate two additional sets of shuffled sequences as controls, and determine the motifs that Discover, DREME, and FIRE discover for the three contrasts formed from the same signal sequences and the three sets of shuffled sequences. Our method robustly identifies the same two motifs for the three contrasts. The other methods report more motifs than Discover, but not all motifs reported by the other methods are reproduced when using different sets of shuffled sequences. The motifs discovered by Discover are full-length, while about half of the motifs reported by the other methods are redundant, corresponding to partially overlapping segments of the same motif. Thus, while potentially missing some true motifs, Discover consistently and robustly identifies a non-redundant set of full-length motifs with a higher true positive rate than competing methods.

## Conclusions

We conclude that the Discover analyses of ChIP-Seq data are stringent and robust, as indicated by (I) the strong similarity of multiply discovered motifs, (II) the high proportion of previously described motifs recovered, (III) the high proportion of known co-factor motifs among the previously described motifs, and (IV) the consistent results when applied to multiple sets of shuffled sequences.

## List of publications

The method described herein was published after this dissertation was handed in.

Maaskola, J. and Rajewsky, N. (2014). Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. *Nucleic Acids Res* 42(21):12995–13011.

Prior to the work that lead up to this dissertation I was involved in several research projects. Following is a list of the publications engendered by these projects. The contents of these publications are discussed in chapter 2 with a focus on my personal contributions.

## Shared first authorships

I was shared first author of the following publications. My contributions to these collaborative investigations constituted their main computational analyses.

Stoeckius, M., Maaskola, J., Colombo, T., Rahn, H.-P., Friedländer, M. R., Li, N., Chen, W., Piano, F., and Rajewsky, N. (2009). Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression. *Nat Methods*, 6(10):745–751.

von Eyss, B., Maaskola, J., Memczak, S., Möllmann, K., Schuetz, A., Loddenkemper, C., Tanh, M.-D., Otto, A., Muegge, K., Heinemann, U., Rajewsky, N., and Ziebold, U. (2012). The SNF2-like helicase HELLS mediates E2F3-dependent transcription and cellular transformation. *EMBO J*, 31(4):972–985.

## Further publications

I contributed computational biology analyses to the following publications.

Anders, G., Mackowiak, S. D., Jens, M., Maaskola, J., Kuntzagk, A., Rajewsky, N., Landthaler, M., and Dieterich, C. (2012). doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res*, 40(Database issue):D180–D186.

Chen, K., Maaskola, J., Siegal, M. L., and Rajewsky, N. (2009). Reexamining microRNA site accessibility in *Drosophila*: a population genomics study. *PLoS One*, 4(5):e5681.

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRD-eep. *Nat Biotechnol*, 26(4):407–415.

Song, J. L., Stoeckius, M., Maaskola, J., Friedländer, M., Stepicheva, N., Juliano, C., Lebedeva, S., Thompson, W., Rajewsky, N., and Wessel, G. M. (2012). Select microRNAs are essential for early development in the sea urchin. *Dev Biol*, 362(1):104–113.

Wang, Y., Gogol-Döring, A., Hu, H., Fröhler, S., Ma, Y., Jens, M., Maaskola, J., Murakawa, Y., Quedenau, C., Landthaler, M., Kalscheuer, V., Wiczorek, D., Wang, Y., Hu, Y., and Chen, W. (2013). Integrative analysis revealed the molecular mechanism underlying RBM10-mediated splicing regulation. *EMBO Mol Med*, 5(9):1431–1442.



## **Part I**

# **Biological Background**



# Chapter 1

## Systems biology of gene regulation

This chapter gives a high-level review of the biological background of this dissertation. Its structure is as follows. First, section 1.1 comments on the current state of research in the field of life science. Then, section 1.2 reviews the main processes of the systems involved in gene regulation. The chapter concludes in section 1.3 with an overview of recently developed experimental technologies that provide data analyzed in this thesis.

### 1.1 State of life science research

Life-science research has recently blossomed, having reached a state of maturity and providing levels of understanding exceeding even wild speculations of only few decades ago. New insights continue to be found at a high pace thanks to the introduction of new experimental technologies and matching analytical techniques. Combined, they provide detailed and vast amounts of data, that have moved biological research from small-scale to high-throughput experiments that require the efforts and collaboration of experts coming from fields of research spanning from physical sciences like chemists, physicists, and molecular biologists to data analysis specialists like mathematicians, statisticians, and computer scientists.

This chapter cannot but scratch the surface of the involved biological processes, and the interested reader is instead referred to textbooks such as Alberts et al. (2007) and Nelson and Cox (2012) for more comprehensive discussion of these topics. While the high pace of research also means that efforts to give comprehensive overviews of biology quickly become outdated and can not include the latest insights, fortunately the move towards digital publishing, especially combined with open-access publishing policies, means that research literature is increasingly available at the fingertips of the interested. Thus I limit discussion of the biological background in this chapter to main topics, and encourage the reader to find more information in the original literature, for example through portals such as NCBI's literature database PubMed<sup>1</sup>.

---

<sup>1</sup><http://www.ncbi.nlm.nih.gov/pubmed>

## 1.2 Gene regulatory processes

Processes in living cells are subject to many different levels of regulation. The focus of this dissertation is that of gene regulation, which determines the activity of genes. Gene regulation is exerted by a complex system of finely tuned processes, many of which so intimately intermeshed as to contemporaneously act on the same molecule. It is organized in a spatio-temporal cascade of successive processing steps in different cellular compartments. Developmental time points and tissue types are characterized by differential gene expression and activity resulting from differences in gene regulation.

On a high level, gene regulation can be understood to be organized in four layers: epigenetic, transcriptional, post-transcriptional, and post-translational regulation. Epigenetic regulation determines the degree of condensation of DNA and thus accessibility to the transcriptional apparatus. Transcriptional regulation determines the rate at which primary transcripts of a given gene locus are being produced. Post-transcriptional regulation encompasses the regulatory steps leading through maturation and up to translation of messenger RNAs (mRNAs) into polypeptide chains, and ultimately to destruction or degradation of transcripts. Finally, post-translational regulation is the regulation of enzymatic activity of amino acid chains.

The methods presented in this thesis—based on probabilistic sequence modeling—are geared towards the analysis of data involved in research on the first three of these layers, which all involve identification of regions within nucleic acid molecules by proteins or other nucleic acids<sup>2</sup>.

### 1.2.1 Epigenetic regulation

As has long been known, genetic information is stored on chromosomes (Griffith, 1928), large nuclear structures composed of DNA and proteins. While the primary carrier of genetic information is the DNA (Avery, Macleod, and McCarty, 1944), particularly its linear sequence (Watson and Crick, 1953), further mechanisms outside of DNA sequence changes exist that allow long-term information transmission. Some of these mechanisms carry information through cell divisions and play roles in development, others are stable even through generations. Epigenetics is the study of information transmission and inheritance mechanisms that do not rely on DNA sequence changes.

Epigenetic mechanisms comprise modifications to DNA bases, such as methylation, as well as post-translational modifications of histones. Together, these determine the degree of condensation of regions of DNA which allow or hinder transcription. Vast regions of DNA may condense, forming heterochromatin, thus becoming long-term transcriptionally inactivated. Conversely, transcribed regions of the DNA are significantly less condensed and known as euchromatin. While DNA modifications are rather stable, histone modifications can be rearranged much more dynamically. There are many different kinds of histone modifications, forming a histone modification code or language (Jenuwein and Allis, 2001; J.-S. Lee, E. Smith, and Shilatifard, 2010; Rando, 2012).

---

<sup>2</sup>Post-translational regulation is also frequently informed by sequence-based recognition, but the larger alphabet of amino acids, as well as larger contributions due to structural properties of proteins put less emphasis on sequence based modeling.



Transcriptional (in-)activity and modifications to histones and DNA all feed back on each other, but a comprehensive picture of these connections is only beginning to emerge. Also, certain ribonucleic acids (RNAs) are involved in epigenetic regulation, e.g. long non-coding RNAs (lncRNAs) (J. T. Lee and Bartolomei, 2013) and piRNAs (Peng and H. Lin, 2013).

### 1.2.2 Transcriptional regulation

As illustrated in figure 1.1 the biogenesis of an mRNA involves many finely regulated steps. Transcriptional regulation of a gene is informed by the presence of binding sites for transcription factors (TFs) in the vicinity of the gene. TFs are the class of DNA-binding proteins (DBPs) that bind to chromatin to regulate in a positive or negative manner the rate of transcription of genes. Regulatory regions proximal to the transcription start site (TSS) are termed promoter, and regulatory regions distal to the TSS are known as enhancer regions.

Transcription factor binding sites (TFBSs) in enhancer and promoter regions constitute regulatory signals interpreted depending on the activity of corresponding TFs, which in turn is determined by expression strength and post-translational modifications such as phosphorylation status. The regulatory signals of TFBSs promote or hinder the assembly and determine the composition of the transcriptional machinery, a protein complex that constitutes the transcriptional apparatus. When the necessary components of the transcriptional machinery are present transcription is initiated and mRNA is synthesized, complementary in sequence to the transcribed strand of DNA.

### 1.2.3 Post-transcriptional regulation

Once a gene has been transcribed it becomes subject to post-transcriptional regulation. There are five important processes that constitute post-transcriptional regulation: splicing, RNA editing, RNA localization control, RNA stability control, and translational control.

Already before synthesis of the nascent mRNA transcript is completed, proteins assemble on it that add a 5' methyl guanosine cap to the transcript. A second characteristic that marks mRNAs, and which is added upon completion of transcription, is the presence of a polyadenylated 3' tail. The absence of either signal poises transcripts for degradation. Splicing removes introns, yielding mature transcripts, often in a variety of alternative isoforms. Some mRNAs undergo RNA editing, in which individual bases are chemically modified, potentially changing the resulting amino acid when coding nucleotides are affected, or otherwise influencing, i.e. introducing, abrogating, or modifying, regulatory signals in the message.

Mature mRNA is exported from the nucleus into the cytoplasm and may subsequently be transported to specialized localizations in the cell, for spatial or temporal sorting. Finally, protein synthesis may take place involving mRNA as template, or mRNAs may specifically be degraded.

Post-transcriptional regulatory processes are mediated by specific RNA-binding proteins (RBPs), and often involve various non-coding RNAs (ncRNAs). Diverse classes of

small, ncRNAs (Ghildiyal and Zamore, 2009) execute important regulatory roles, and dysregulation of small ncRNAs is relevant to many diseases (Esteller, 2011). Structural, but also regulatory roles are executed by larger ncRNAs, such as small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), transfer RNAs (tRNAs), and ribosomal RNAs (rRNAs). Circular RNAs (circRNAs) are another class of recently discovered RNAs with regulatory potency (Hansen et al., 2013; Memczak et al., 2013).

Splicing, editing, sorting, stability, storage, and degradation of an mRNA are all influenced by regulatory signal in the transcript. Also, the efficiency with which an mRNA is translated depends on various factors, including regulatory signals in the transcript, as well as usage of differential efficiency of various synonymous tRNAs.

**miRNA** The best-studied examples of regulatory small ncRNAs are microRNAs (miRNAs) (Ambros, 2001; Lau, L. P. Lim, et al., 2001; R. C. Lee and Ambros, 2001) which are RNA molecules of about 22 nt length, many of which deeply conserved in sequence across the phylogenetic tree (Berezikov, 2011; Wheeler et al., 2009). In conjunction with specific RBPs miRNAs assemble the RNA-induced silencing complexes (RISCs) that regulates stability and translation efficiency of target mRNAs. A single kind of mRNA may be subject to regulation by dozens of miRNAs, and one miRNA may regulate hundreds of mRNAs (Baek et al., 2008; Selbach et al., 2008). The modus operandi for the RISC is recognition of target mRNAs by the miRNA through sequence complementarity. The constitution of the RISC determines the effect on the targets, and miRNAs are sorted in a regulated manner into different members of the Argonaute family (Czech and Hannon, 2011). Target mRNA may then be stored inaccessible for the translational machinery in RNA granules which are condensed droplets in the cytosol that are composed of proteins and RNAs. While sequestered in RNA granules mRNAs may be translationally repressed or subject to degradation (Huntzinger and Izaurralde, 2011). miRNA biogenesis is itself regulated on transcriptional and post-transcriptional level (Krol, Loedige, and Filipowicz, 2010). Functionally, miRNAs seem to be involved in canalization, i.e. robustification or stabilization (Waddington, 1942), of gene expression programs and changes in the miRNA regulation accompany many differentiation processes in developmental decisions (Hornstein and Shomron, 2006; Peterson, Dietrich, and McPeck, 2009; C.-I. Wu, Shen, and T. Tang, 2009). miRNA diversity and regulation contributed to the evolution of organismal complexity (Berezikov, 2011).

**piRNA** PIWI-interacting RNAs (piRNAs) play a role in the germline and stem cells of diverse organisms (C. Juliano, J. Wang, and H. Lin, 2011), but also have epigenetic and somatic functions (Peng and H. Lin, 2013). Many species' piRNAs are about 28 nucleotides long, yet some species' piRNAs are shorter, such as e.g. the 21U-RNA of *C. elegans*. They protect the integrity of the genome from invasion by genomic parasites, and have for this reason been nicknamed 'guardians of the genome' (Senti and Brennecke, 2010), as well as 'vanguard of the genome defense' (Siomi et al., 2011).

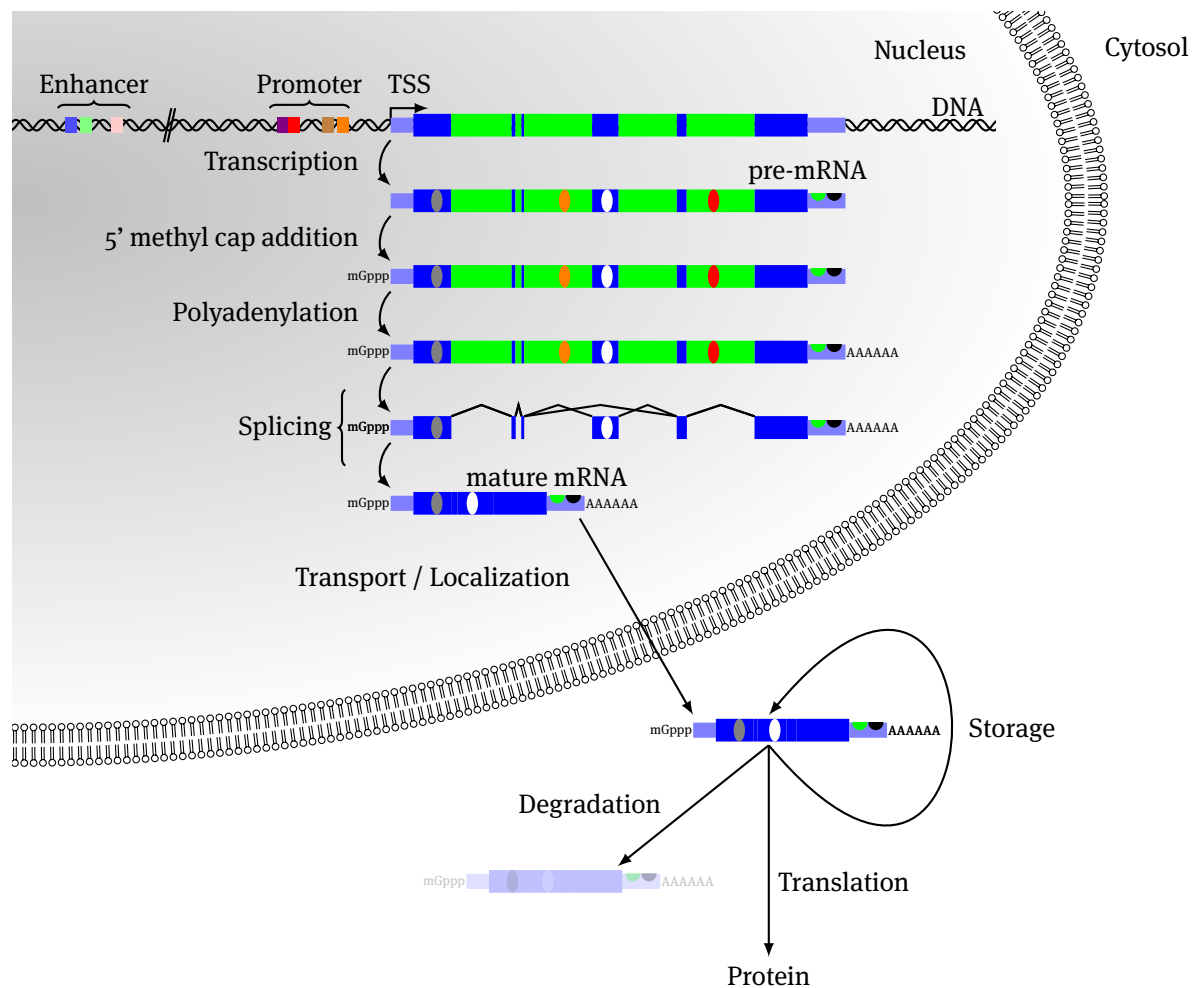


Figure 1.1: Biogenesis of an mRNA. The illustration shows a DNA region containing a gene locus consisting of a TSS, a transcribed region downstream of the TSS, as well as a proximal promoter region and a distal enhancer, with DNA regulatory elements in the form of transcription factor binding sites (squares). Transcription creates primary transcripts (pre-mRNA) of the gene locus, consisting of 5'UTR (light blue region on left side), coding exons (thick, blue segments), introns (thick, green segments), and 3'UTR (light blue region on right side). The transcripts are 5' methyl capped, and polyadenylated. Splicing may create multiple alternative isoforms. Individual nucleotides may be modified by RNA-editing (not shown). The mature transcripts are transported out of the nucleus and into the cytosol, or to specific places in the cell. Mature mRNA may be subject to translation, storage, or targeted degradation. Splicing, RNA-editing, transport, translation, storage, and degradation of mRNA is directed by the tissue specific expression of RBPs and RBP binding sites (ovals) in the transcripts. miRNAs may direct the cytosolic fate of mRNAs via binding to miRNA target sites (semicircles), which are primarily located in the 3'UTR.

### 1.3 Sequencing technologies

Microarray-based approaches fueled much of the last two decades' molecular biology research, but more recently advancements in sequencing technology opened up new avenues for biological research.

Classically, Sanger's chain-termination method (Sanger, Nicklen, and Coulson, 1977) was used to sequence DNA molecules of up to a few hundred bases length. Roche's 454 platform (Margulies et al., 2005) pioneered the next-generation sequencing technologies. It allows the simultaneous sequencing of hundreds of thousands of nucleotide sequences of length  $\geq 400$  bases, producing about 100Mb sequence in a single sequencing run. Illumina's Solexa sequencing platform (Bentley, 2006) initially produced relatively short reads of length 36 nucleotides, yielding about 1 GiB sequence per run. Later generations of the platform allow longer reads of up to 100 bases, and yield up to 300 GiB of sequence per run.

More recently developed sequencing technologies, such as ABI's SOLiD platform (Valouev et al., 2008) or Pacific Biosciences' single molecule real-time (SMRT) sequencing (Eid et al., 2009) offer further advantages. These include the ability to sequence even deeper and more accurately, to detect chemical modifications of bases such as methylation, and to directly sequence full isoforms for improved understanding of splicing.

**Assaying transcriptional regulation by chromatin immunoprecipitation** Prior to widespread adoption of microarray-based technology, using chromatin IP (ChIP), PCR and Sanger sequencing, only few sequences (on the order of dozens) were available that were known to be bound *in vivo* by a given factor. The introduction of ChIP followed by microarray quantification (ChIP-Chip) (Iyer et al., 2001; Ren, Robert, et al., 2000; L. V. Sun et al., 2003; J. Wu et al., 2006) technologies, increased this to hundreds of sequences. The step from few, well-characterized targets to hundreds of targets meant also a reduction of the frequency of true positives in the sequence set.

**ChIP-Seq** The introduction of ChIP followed by sequencing (ChIP-Seq) (D. S. Johnson et al., 2007; Robertson et al., 2007) boosted the number of identified candidate TF-bound regions by another order of magnitude or two. ChIP-Seq also yields higher spatial resolution than previous ChIP-Chip approaches, and by sophisticated computational analysis (e.g. Y. Zhang et al., 2008) targeted regions are also identified with higher confidence. See figure 1.3 for an overview of the protocol used for ChIP-Seq. ChIP-Seq not only is applicable to transcriptional regulation research, but also to assay chromatin state by targeting epitopes characteristic of certain DNA or histone modifications, and thus is also useful to epigenetic research (Ku et al., 2011).

**Sequencing technologies for post-transcriptional research** Analogously to the introduction of ChIP-Seq for DBP research, RIP-Seq technologies improving on the earlier DNA microarray quantification of co-IPed mRNA (RIP-Chip) (Tenenbaum et al., 2000), like HITS-CLIP (Licatalosi et al., 2008), iCLIP (König et al., 2010), and PAR-CLIP (Hafner et al., 2010) boosted RBP research.

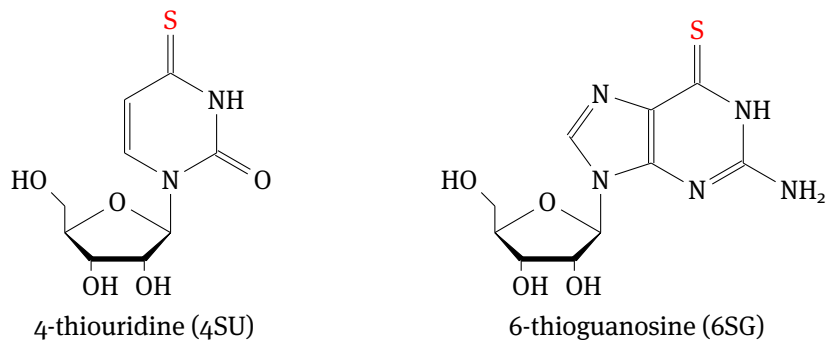


Figure 1.2: Structure of 4-thiouridine (4SU) and 6-thioguanosine (6SG), two of the photo-activatable nucleoside analoga used in PAR-CLIP (Hafner et al., 2010). Red color marks the sulfur groups that differ from regular uridine and guanosine.

**PAR-CLIP** PAR-CLIP uses photo-activatable nucleoside analoga to increase cross-linking efficiency (Hafner et al., 2010). In particular, 4-thiouridine (4SU) and 6-thioguanosine (6SG) are typically used for this, see figure 1.2. The advantage of using these nucleoside analoga is that by using UV-light they induce covalent bonds between proteins and the bound RNA. See figure 1.3 for an overview of the protocol. The cross-linked nucleotide analoga are incorporated into RNA like the corresponding regular nucleosides, but have different basepairing properties. During synthesis of the cDNA, cross-linked 4SU tends to basepair with guanosine instead of adenosine, and cross-linked 6SG with thymidine instead of cytidine. This has the consequence that upon sequencing and mapping the reads to the genome, mismatches between the sequenced and the reference sequence are observed. For 4SU T to C conversions are observed, and for 6SG G to A. A prerequisite for cross-linking is intimate vicinity of the activatable group of 4SU or 6SG with amino acid sidechains. Thus, nucleotide conversion events are typically observed within or next to the cognate nucleotide sequence of the bound RNA. This allows bioinformatics analyses to pinpoint the exact binding sites of an RBP. The spatial resolution that PAR-CLIP allows is on the order of tens of nucleotides.

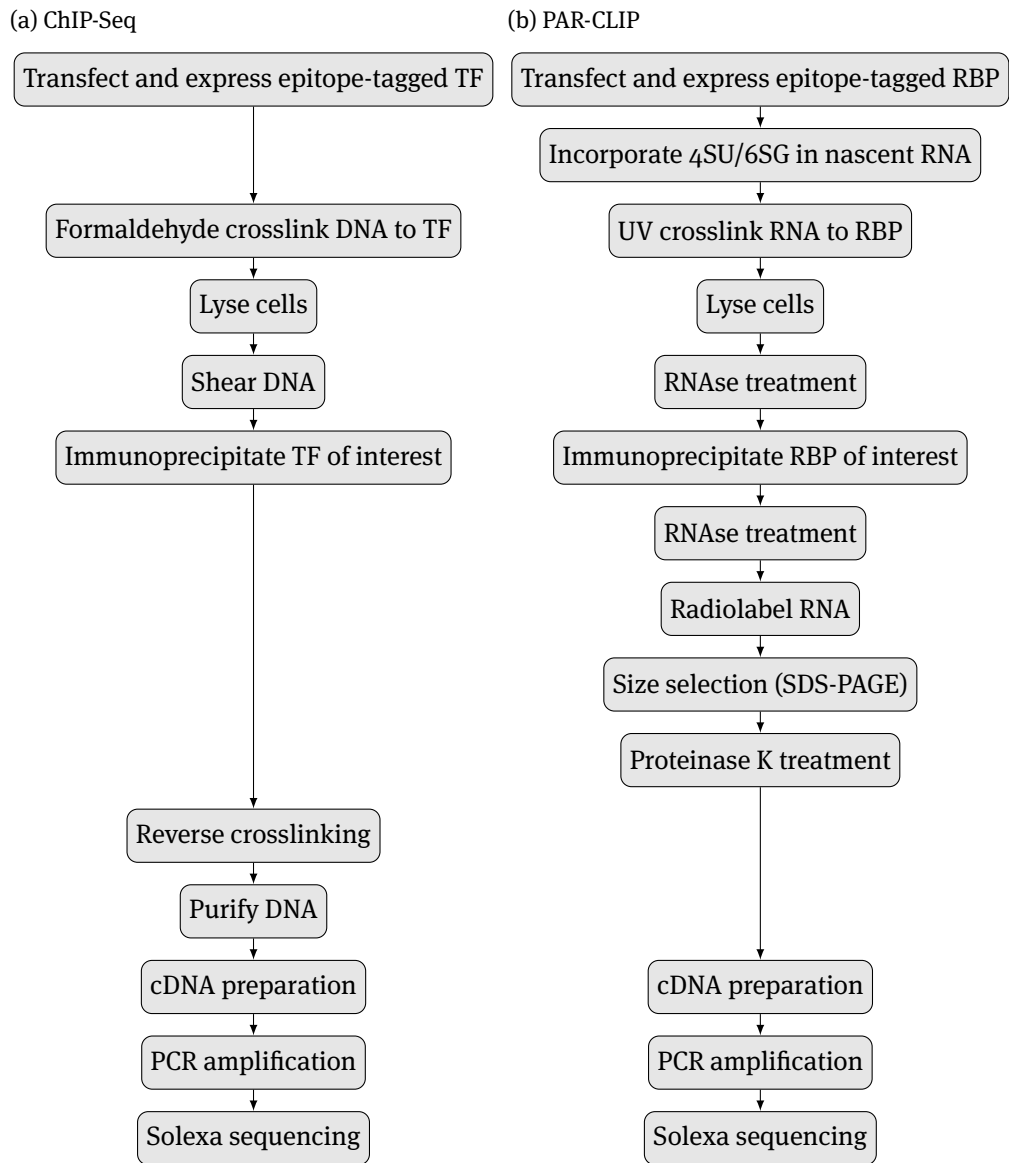


Figure 1.3: (a) ChIP-Seq and (b) PAR-CLIP methodologies. Corresponding steps are aligned vertically. The first step of both methodologies is not necessary when antibodies against endogenous proteins are used.

## Chapter 2

# Published computational biology research

This chapter gives an overview of the biological research that I have been involved in while doing my PhD under the supervision of Nikolaus Rajewsky at the Laboratory for Systems Biology of Gene Regulatory Elements, Berlin Institute for Medical Systems Biology, Max-Delbrück-Centrum für Molekulare Medizin, Berlin.

Throughout this chapter, discussion of the publications is generally structured as follows. First, abstract and introduction of the publications are presented. Then, the parts of the investigation are outlined that I worked on. To this end, I discuss the methods I applied, and then present research findings that derive from application of my methodological contributions. For the sake of brevity, findings that did not directly grow out of my contributions are elided.

The discriminative motif discovery (DMD) methods described in this thesis have been published in a peer-reviewed journal. Being the subject matter of this dissertation, they will not be discussed further in this chapter.

Maaskola, J. and Rajewsky, N. (2014). Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. *Nucleic Acids Res* 42(21):12995–13011.

### 2.1 List of publications

#### Shared first authorships

I was shared first author of the following publications. My contributions to these collaborative investigations constituted their main computational analyses.

Stoeckius, M., Maaskola, J., Colombo, T., Rahn, H.-P., Friedländer, M. R., Li, N., Chen, W., Piano, F., and Rajewsky, N. (2009). Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression. *Nat Methods* 6 (10): 745–751.

von Eyss, B., [Maaskola](#), J., Memczak, S., Möllmann, K., Schuetz, A., Loddenkemper, C., Tanh, M.-D., Otto, A., Muegge, K., Heinemann, U., Rajewsky, N., and Ziebold, U. (2012). The SNF2-like helicase HELLS mediates E2F3-dependent transcription and cellular transformation. *EMBO J* 31 (4): 972–985.

### Further publications

The following is a list of publications about investigations to which I contributed computational biological analyses.

Friedländer, M. R., Chen, W., Adamidi, C., [Maaskola](#), J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 26 (4): 407–415.

Chen, K., [Maaskola](#), J., Siegal, M. L., and Rajewsky, N. (2009). Reexamining microRNA site accessibility in *Drosophila*: a population genomics study. *PLoS One* 4 (5): e5681.

Anders, G., Mackowiak, S. D., Jens, M., [Maaskola](#), J., Kuntzagk, A., Rajewsky, N., Landthaler, M., and Dieterich, C. (2012). doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* 40 (Database issue): D180–D186.

Song, J. L., Stoeckius, M., [Maaskola](#), J., Friedländer, M., Stepicheva, N., Juliano, C., Lebedeva, S., Thompson, W., Rajewsky, N., and Wessel, G. M. (2012). Select microRNAs are essential for early development in the sea urchin. *Dev Biol* 362 (1): 104–113.

Wang, Y., Gogol-Döring, A., Hu, H., Fröhler, S., Ma, Y., Jens, M., [Maaskola](#), J., Murakawa, Y., Quedenau, C., Landthaler, M., Kalscheuer, V., Wiczorek, D., Wang, Y., Hu, Y., and Chen, W. (2013). Integrative analysis revealed the molecular mechanism underlying RBM10-mediated splicing regulation. *EMBO Mol Med* 5 (9): 1431–1442.

## 2.2 Computational analysis of deep-sequencing data

NATURE BIOTECHNOLOGY VOLUME 26 NUMBER 4 APRIL 2008

### Discovering microRNAs from deep sequencing data using miRDeep

Marc R Friedländer<sup>1</sup>, Wei Chen<sup>2</sup>, Catherine Adamidi<sup>1</sup>, Jonas Maaskola<sup>1</sup>, Ralf Einspanier<sup>3</sup>, Signe Knespel<sup>1</sup> & Nikolaus Rajewsky<sup>1</sup>

**Abstract** The capacity of highly parallel sequencing technologies to detect small RNAs at unprecedented depth suggests their value in systematically identifying miRNAs. However, the identification of miRNAs from the large pool of sequenced transcripts from a single deep sequencing run remains a major challenge. Here, we present an algorithm, miRDeep, which uses a probabilistic model of miRNA biogenesis to score compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor. We demonstrate its accuracy and robustness using published *C. elegans* data



and data we generated by deep sequencing human and dog RNAs. miRDeep reports altogether approximately 230 previously unannotated miRNAs, of which four novel *C. elegans* miRNAs are validated by northern blot analysis.

**Introduction** Animal genomes harbor numerous small, non-coding miRNA genes believed to post-transcriptionally regulate many protein-coding genes to influence processes ranging from metabolism, development and regulation of the nervous and immune systems to the onset of cancer (Bushati and Cohen, 2007). Despite concerted efforts to discover and profile miRNAs, even the number of miRNAs in the human genome remains controversial, with estimates ranging from a few hundred (Bartel, 2004) to tens of thousands (Miranda et al., 2006). Traditional experimental approaches to miRNA discovery have relied on cloning and Sanger sequencing protocols (Aravin and Tuschl, 2005) and human and murine miRNAs have been profiled in hundreds of cDNA libraries from dozens of tissues (Landgraf et al., 2007). However, the vast dynamic range of miRNA expression (from tens of thousands to a few molecules per cell) complicates profiling of miRNAs expressed in low numbers. A complementary approach, involving miRNA discovery by computational predictions that analyze genomic DNA for structures that resemble known miRNA precursors (Bentwich, 2005), is compromised by sensitivity problems and substantial numbers of false positives (Bentwich, 2005). Therefore, purely computational approaches require experimental follow-ups, which are again difficult for miRNAs with low expression levels in the sample.

‘Deep-sequencing’ technologies opened the door to detecting and profiling known and novel miRNAs at unprecedented sensitivity. Next-generation sequencing platforms, such as those from Solexa / Illumina and 454 Life Sciences / Roche, can sequence DNA orders of magnitude faster and at lower cost than Sanger sequencing and are evolving so rapidly that increases in sequencing speed by at least another order of magnitude seem likely over the next few years. Although the Solexa / Illumina system can produce ~32 million sequencing reads in one run, read length is currently limited to 35 bp. In contrast, the current 454 platform yields reads up to 200 bases each, although the number of reads is an order of magnitude less than that of Solexa / Illumina. The nature of sequencing errors also contributes further to the different output characteristics of the two approaches.

Despite the ability of both technologies to sequence — and thus to detect — miRNAs at previously unmatched throughput, deep sequencing presents formidable computational challenges and suffers from biases such as those arising from the preparation of small RNA libraries. Even mapping deep-sequencing reads to the genome is itself not trivial, as no animal genome besides that of *C. elegans*, has been sequenced completely. Moreover, sequencing errors and polymorphisms, as well as RNA editing and splicing are but some of the factors that contribute to ambiguity. Although currently almost all of these problems remain mostly unsolved, deep sequencing can successfully survey the small RNA contents of animal genomes with unmatched sensitivity (Aravin, Sachidanandam, et al., 2007; Berezikov et al., 2006; Brennecke et al., 2007; Girard et al., 2006; Houwing et al., 2007; Lau, Seto, et al., 2006; Pak and Fire, 2007; Ruby, Jan, et al., 2006; Tarasov et al., 2007).

When profiling small RNAs by deep-sequencing technology, separating miRNAs from

the pool of other sequenced small RNAs or degradation products is a central problem that is often not described or only partially addressed (Berezikov et al., 2006; Ruby, Jan, et al., 2006). Furthermore, despite a growing need to analyze deep-sequencing data, there is no publicly available algorithm to detect miRNAs in these data.

miRDeep, our publicly available software package, can be used to solve this problem at least in part. Importantly, it also includes stringent statistical controls to estimate the false positive rate and the sensitivity of miRDeep predictions. Therefore, users can not only run miRDeep on their own deep-sequencing data to detect known and novel miRNAs, but can also estimate the quality of their results. At the heart of miRDeep is the idea of detecting miRNAs by analyzing how sequenced RNAs are compatible with how miRNA precursors are processed in the cell. As deep sequencing permits statistical analysis of this model, one can assign a score of the likelihood that a detected RNA is indeed a mature miRNA. Therefore, the foreseeable advances in sequencing capacity of deep-sequencing technologies should further boost the power of miRDeep. In order to address an ongoing discussion about the importance of non-conserved miRNAs (K. Chen and Rajewsky, 2007) and to be as unbiased as possible, we designed miRDeep to detect miRNAs without cross-species comparisons. Finally, given the rapid evolution of deep-sequencing technology, we designed miRDeep to be as flexible as possible and tested it using both Solexa- and 454-derived data from human, the domestic dog and *C. elegans* – animals from the two main branches of Bilateria, representing very different genomic complexity.

## My contributions

I developed a deep-sequencing read alignment tool, described below, and used it to map the dog small RNA deep-sequencing data to the canine genome. The other datasets were mapped by my colleague Marc Friedländer using NCBI megablast (BLAST version 2.2.14).

Adapters were removed from the dog lymphocyte Solexa dataset with a custom mapping tool based on suffix arrays (Manber and Myers, 1993). First, the adapter sequences were identified in the deep-sequencing reads. We required the presence of minimum 10 nt of the 5' adapter sequence with a maximum of three edits (mismatches and/or insertions/deletions). Reads that contained an identified adapter sequence had the adapter removed and were retained, the rest were discarded. The retained reads were mapped to the dog genome (*Canis familiaris* version canFam2, from UCSC) using the custom mapping tool, allowing for up to two edits. For each read, mappings of suboptimal edit distance were discarded (if the best-mapping was edit distance 1, all edit distance 2 mappings were discarded).

**Suffix array based alignment** The suffix array based alignment tool allows to look up perfect occurrences of queries in the database in time linear in the length of the query, independent of the size of the database. Finding all  $m$  perfect matches for a query of length  $n$  has a worst-case runtime of  $\mathcal{O}(An + m)$ , where  $A$  is the size of the alphabet. For genome sequences  $A = 4$  or  $A = 5$ , depending on whether Ns in the genome are treated as characters of the alphabet, or are randomly assigned one of the other four symbols.

The suffix array based alignment also allows to find imperfect occurrences of queries,


using either Hamming distance or Levenshtein distance, which is also known as edit distance. Hamming distance allows substitutions of individual symbols, while edit distance additionally allows insertions or deletions of symbols. Generally, alignment is done by first searching for perfect occurrences, then for occurrences in (Hamming or edit) distance 1, and finally for occurrences in distance 2. Frequently, only best matches are of interest, so search can be stopped once matches have been found at a given distance. Optionally, higher distance matches can be searched by in turn either incrementing by 1, or by doubling the allowed distance.

To find genomic matches, alignment needs to be done twice, once for the original query to find occurrences on the sense strand of the database, and once for the reverse complement of the query to find occurrences on the antisense strand. Alternatively, using twice as much memory, a suffix array of both the sense and antisense strand of the genome can be used, that only needs to be searched once. Genomic suffix arrays require considerable amounts of memory, in the range of dozens of GiBs for mammalian genomes; e.g. the suffix array for only the sense strand of the human genome requires 52 GiB of RAM.

Allowing for imperfect occurrences in finding matches to a query increases runtime. Occurrences of Hamming distance  $\leq k$  to a query of length  $n$  are found in worst-case runtime of  $\mathcal{O}(A^{k+1}n + m)$ , where  $m$  is the number of matches in Hamming distance  $k$ .

Finding matches with edit distance  $\leq k$  is done by performing a banded sequence alignment of the query against the suffix array (Durbin et al., 1998; Gusfield, 1997; Mount, 2004). A banded alignment of a query of length  $n$  against a sequence of length  $N$  with  $n \leq N$  allowing up to  $k$  edits can be done in  $\mathcal{O}((2k + 1)n)$ . In principle, such a banded alignment needs to be done against all unique prefixes of length  $n + k$  of the suffixes of the database, yielding a maximal runtime of  $\mathcal{O}((2k + 1)nN)$ , when there are  $N$  unique length  $n + k$  suffix-prefixes in the database. However, runtime can be significantly reduced by respecting two factors. First, banded alignments of a query  $q$  against any suffix-prefix  $s$  can be stopped as soon it is clear that there is no alignment with edit distance  $d(q, s) \leq k$ . Second, the lexicographical order of the suffixes in the suffix array means that by performing the alignment column-wise<sup>1</sup> work can be saved when proceeding from one suffix to the lexicographically next suffix due to the common prefixes of lexicographically ordered strings. Thus, only those columns of the dynamic programming table need to be recomputed that correspond to the part of the next suffix that differ from the previous.

## 2.3 Population genomics of drosophilid miRNAs

May 2009 | Volume 4 | Issue 5 | e5681 

### Reexamining microRNA Site Accessibility in *Drosophila*: A Population Genomics Study

Kevin Chen<sup>1,2\*</sup>, Jonas Maaskola<sup>2</sup>, Mark L. Siegal<sup>1</sup>, Nikolaus Rajewsky<sup>2</sup>

<sup>1</sup>I.e. orienting the dynamic programming matrix such that positions of the query correspond to rows, and positions of the template to columns.

**Abstract** Kertesz et al. (2007) described PITA, a miRNA target prediction algorithm based on hybridization energy and site accessibility. In this note, we used a population genomics approach to reexamine their data and found that the PITA algorithm had lower specificity than methods based on evolutionary conservation at comparable levels of sensitivity. We also showed that deeply conserved miRNAs tend to have stronger hybridization energies to their targets than do other miRNAs. Although PITA had higher specificity in predicting targets than a naïve seed-match method, this signal was primarily due to the use of a single cutoff score for all miRNAs and to the observed correlation between conservation and hybridization energy. Overall, our results clarify the accuracy of different miRNA target prediction algorithms in *Drosophila* and the role of site accessibility in miRNA target prediction.

**Introduction** Population genomics has been suggested as a method of evaluating the accuracy of genome-wide predictions of cis-regulatory sites (Boffelli et al., 2004; K. Chen and Rajewsky, 2006; Fairbrother, Holste, et al., 2004; Saunders, Liang, and Li, 2007). The idea is to use polymorphism data and population genetics techniques to estimate the level of purifying selection on predicted cis-regulatory sites genome-wide and to use this quantity as a proxy for the accuracy of the prediction algorithm. The underlying assumption is that an accurate prediction algorithm should identify functionally important sites that are likely to be under selective constraint. This is the same assumption underlying comparative genomics approaches but the population genomics approach is sensitive to natural selection of a different strength and on a different time scale. It is likely to become more useful in the future with the advent of high-throughput genome resequencing.

In this note we used a population genomics approach to reexamine the methods and data presented in Kertesz et al. (2007). There the authors presented a method for predicting miRNA binding sites in *Drosophila* using the score  $ddG = dG(\text{duplex}) - dG(\text{open})$  where  $dG(\text{duplex})$  is the hybridization energy of the miRNA to the binding site and  $dG(\text{open})$  is the energy required to open the local RNA secondary structure around the binding site. The  $ddG$  score was used to rank all possible miRNA seed matches in 3'UTRs (see Kertesz et al. (2007) for details on the method). On a set of 190 experimentally validated target sites, the method was shown to perform more accurately than several leading methods, including Pictar (Krek et al., 2005; Lall et al., 2006) and the method of Stark, M. F. Lin, et al. (2007), that do not use site accessibility but instead require conservation of seed matches between species. We found this result surprising because we expected that conservation would implicitly select for all sequence determinants of functional miRNA binding, including site accessibility. We therefore sought to corroborate the results of Kertesz et al. (2007) using a population genomics approach.

## My contributions

In this investigation, my colleague Kevin Chen analyzed genetic variation in predicted miRNA target sites, while I analyzed genetic variation in miRNA genes. As evolutionary patterns within miRNA had been studied previously, my contribution allowed us to establish consistency of our genetic data with previous observations. To this end I examined

genetic variation in miRNA genes, and compared its distribution across structurally defined parts of miRNA genes.

**Summary** We used whole genome shot-gun sequence data from six inbred lines of the fruit fly *Drosophila simulans* from the Drosophila Population Genomics Project (Begun et al., 2007) to estimate levels of polymorphism within *D. simulans* and divergence between *D. simulans* and *D. melanogaster* (see below). To verify the accuracy of the data and our data processing methods, we first examined the patterns of polymorphism and divergence in miRNA genes. These patterns have been established in previous studies of divergence across species (e.g. Lai et al., 2003; Lunter, Ponting, and Hein, 2006) and within species (Lu, Fu, et al., 2008; Saunders, Liang, and Li, 2007) and thus are a good test of data quality. We note that such an analysis was not possible in a previous study of single nucleotide polymorphisms (SNPs) in human miRNAs (K. Chen and Rajewsky, 2006) or in miRNA re-sequencing studies in humans and Arabidopsis (Diederichs and Haber, 2006; Ehrenreich and Purugganan, 2008; Iwai and Naraba, 2005) because of the low rate of polymorphism in these species compared to *Drosophila*.

Our analysis of evolutionary patterns in miRNA genes confirmed the following hierarchy of selective constraint on the different parts of the miRNA precursor: seed > mature miRNA > star miRNA > loop > flanking control region (see below). Our analysis of indel patterns also confirmed that *D. simulans* miRNAs are more strongly depleted of indels than nucleotide substitutions compared to flanking control regions (see below), as previously observed between mammalian species (Lunter, Ponting, and Hein, 2006). A notable observation from our analysis is that the miRNA precursor loop length is under stabilizing selection since we observed a strong depletion of indels in the loop relative to flanking control regions (see below) (one-sided Z test, insertions  $Z > 3.4$ ,  $p$ -value  $< 0.0003$ , deletions  $Z > 3.9$ ,  $p$ -value  $< 4.8e-5$ ). This suggests that miRNA precursor loop length is functionally important, consistent with previous experimental (Zeng, R. Yi, and Cullen, 2005) and computational (Rabani, Kertesz, and Segal, 2008) data.

**Evolution of miRNA genes in *D. simulans*** We mapped all *D. melanogaster* miRNA precursors from Rfam 10.1 to the six *D. simulans* lines using the *D. simulans* syntenic assembly produced by the Drosophila Population Genomics Project<sup>2</sup>. 141 out of 152 miRNAs had complete sequence coverage in the *D. simulans* assembly, while the other 11 miRNAs did not have complete coverage, typically because they are in heterochromatic regions. Of the 141 fully sequenced miRNAs, we identified eight miRNAs that had an unusually high number of differences between *D. melanogaster* and *D. simulans* (miR-303, miR-982, miR-983-1, miR-983-2 and miR-984, which form a cluster, as well as miR-979, miR-985, and miR-997). The high rate of divergence and the positions of the differences in the miRNAs strongly suggest that these miRNAs are not functional in *D. simulans* (see Chen et al., 2009, Supplementary Table 1 for details). Although we cannot completely exclude the possibility that these are functional miRNAs that have diverged in function in *D. simulans*, we chose to exclude them from the remaining analysis, leaving a set of 133 confidently identified *D. simulans* miRNAs. Regardless of whether the eight miRNAs are non-functional or

<sup>2</sup><http://www.dpgp.org>

highly diverged in *D. simulans*, these data demonstrate that miRNAs can evolve rapidly over short evolutionary distances.

The McDonald-Kreitman (MK) test is commonly used to infer directional selection by comparing the ratio of divergence to polymorphism in a putatively selected region (typically non-synonymous sites) to that in a putatively neutral region (typically synonymous sites). Adaptive mutations contribute more to divergence than to polymorphism, so an excess of divergence relative to polymorphism is indicative of positive selection. Conversely, weakly deleterious mutations contribute more to polymorphism than to divergence and so an excess of polymorphism is consistent with weak negative selection. We observed a high ratio of divergence to polymorphism in miRNA genes compared to control flanking regions (see next section,  $\chi^2$  test,  $p$  value  $< 5e-5$ ). This pattern is consistent with, but does not necessarily prove, the notion that at least some miRNAs, like protein coding genes, have evolved adaptively in *Drosophila*.

Despite the significant result from the MK test, caution in interpretation is warranted when applying the MK test on a set of genes with potentially different genealogies scattered across the genome because it is possible for the combination of a set of neutral MK tables to result in an MK table that rejects neutrality (see Shapiro et al., 2007, for a detailed discussion). Indeed, we observed that one family in particular, miR-310-313, had an unusually high amount of divergence and low level of polymorphism, both in the miRNA and control flanking regions. We noticed this family because there are two fixed substitutions each in the mature miRNA sequences of miR-312 and miR-313 (though not in the seed). Using the *D. yakuba* sequence as an outgroup, we found that substitutions in this region frequently occurred on both the *D. simulans* and *D. melanogaster* lineages and we verified that the pattern of polymorphism and divergence observed was not due to unusually low sequence coverage in this region.

Although the low level of polymorphism in this family is consistent with a selective sweep, another possible explanation for the pattern could be an unusually low recombination rate in the miR-310-313 region. Since accurate genome-wide recombination rates are not available for *D. simulans*, we used an estimate of the recombination rate in *D. melanogaster*. There is significant correlation between *D. simulans* polymorphism and *D. melanogaster* recombination rates, suggesting that recombination rates between the two species are comparable. The recombination rate in this region was estimated to be 3.7 cM/Mbp, which is close to the highest rate on chromosome 2R (range: 0 to 3.76 cM/Mbp, average: 2.43 cM/Mbp)<sup>3</sup>, thus we can exclude the effect of suppressed recombination at this locus. In summary, the pattern of divergence and polymorphism in this gene family is suggestive of one or more selective sweeps on the region containing the miR-310-313 family, though experimental tests of the functional changes are needed to exclude other possible non-selective explanations, such as demography. Taken together, our data indicate that miRNA genes have evolved rapidly between *D. melanogaster* and *D. simulans* and that in at least one case, evolution may have been adaptive. Our results are consistent with two recent studies on the evolution of miRNAs, including the miR-310-313 family, between *D. melanogaster* and *D. simulans* (Lu, Fu, et al., 2008; Lu, Shen, et al., 2008).

---

<sup>3</sup><http://cgi.stanford.edu/~lipatov/recombination/recombination-rates.txt>



**Variation in miRNA genes strongly correlates with miRNA precursor structure** We used RNAfold (Hofacker et al., 1994) to predict the minimum free energy (MFE) secondary structures of the miRNA precursors. Based on these predicted secondary structures and the annotated mature miRNA sequences from Rfam 10.1, we defined the following five segments of the miRNA gene: annotated mature region; the star-region that base-pairs to it, taking into account the stereotypical 2 nt 3'-overhang; the loop region consisting of the bases between the mature and star regions; the lower, base-paired segment; and the flanking region defined as the single-stranded flanks of the miRNA precursor (Chen et al., 2009, Supplementary Figure 1). In the case of miR-1017, this procedure produced a lower stem only downstream of the hairpin, so this segmentation was manually corrected, leading to a reduction by 10 counts of substitutions in the lower stem region.

MiRNA precursors can be divided into five regions: mature miRNA, miRNA star, loop, lower stem and flank (Chen et al., 2009, Supplementary Figure 1). Each of these regions is expected to experience different levels of selective constraint. We segmented each miRNA precursor into these five different regions using the predicted MFE secondary structure of the miRNA, as described above. We then tabulated all fixed and polymorphic substitutions and indels in the precursor miRNAs (Chen et al., 2009, Supplementary Tables 1 and 2, Supplementary Figures 1 and 2). We used 50 nt of flanking sequence on either side of the miRNA precursors as (putatively) neutral controls. Using control sequences in the local neighborhood of the miRNAs is important to control for variation in rates of polymorphism and divergence across the genome. There are no known functional constraints on the flanking sequences of miRNA precursors and since we observed increased selective constraint on miRNA genes compared to flanking sequences, the presence of functional elements in the flanking sequences would only make our approach more conservative. We did not include control sequences for the mirtrons (i.e. miRNAs that constitute an entire intron) since this would be exonic sequences which are unlikely to evolve neutrally.

We found the seed region (positions 2–8) of the mature miRNA to be under very strong selective constraint since we did not find any substitutions in this region. However, we found fixed substitutions in the first position of miR-960 and miR-973, consistent with the target model that the first position of the mature miRNA is not bound to the target mRNA. We also observed polymorphisms in the seeds of miR-133, miR-280, miR-966 and miR-990 that could be of functional interest since they are expected to affect miRNA targeting. Of these polymorphisms, we attempted to validate the one in miR-133 because it is the best-studied of these four miRNAs. However, upon resequencing of the relevant *D. simulans* line, this was found to be a sequencing error (data not shown).

In addition to the seed, we detected strong selective constraint on the remainder of the mature sequence, with SNP density only 5% relative to the controls (Chen et al., 2009, Supplementary Table 1). This level of constraint may indicate that either 3' compensatory miRNA binding sites or regulatory elements in this region of the miRNA that mediate post-transcriptional control are more common than currently believed. Also, consistent with previous studies across species, selective constraint on the loop and miRNA flanking sequences was low compared to the rest of the miRNA, but still below the level of the controls (Chen et al., 2009, Supplementary Table 1).

Finally, we observed significant selective constraint on the star sequence, with SNP

density 13% that of controls. This pattern likely reflects base pairing constraints both to pair to the mature miRNA and to prevent the inappropriate strand from being incorporated into the RISC complex. An additional source of constraint is that the star sequence can be incorporated into the miRNA-mediated silencing complex and the scope of targeting by the star sequence is of interest. To address this question, we used the lower stem of the miRNA to model the selective constraint expected due to base pairing constraints and tested for excess constraint on the star miRNA relative to the lower stem. We did not detect any excess constraint on the star miRNA either with respect to SNP and substitution density or with respect to the ratio of divergence to polymorphism (see below, Chi square test,  $p$ -value  $> 0.63$ ).

**Strong selection against insertions and deletions in miRNA genes** An advantage of whole-genome shotgun sequence data over SNP data is the ability to study genome rearrangements, such as indel variation. Overall, we observe that indels are depleted in miRNA genes relative to control sequences, making indel depletion a good feature for identifying miRNAs (Chen et al., 2009, Supplementary Table 2). A similar pattern was previously observed using cross-species data in mammals (Lunter, Ponting, and Hein, 2006).

It is an interesting question whether the size of the loop region of the miRNA is under selective constraint. It has been argued that the loop cannot be too big (Ruby, Stark, et al., 2007) otherwise it would appear to be single-stranded RNA and cause inappropriate Drosha processing while experiments have shown that small loops have impaired processing (Han et al., 2006). We observed a strong depletion of indels in the loop (Chen et al., 2009, Supplementary Table 2) (one-sided Z test, insertions  $Z > 3.4$ ,  $p$ -value  $< 0.0003$ , deletions  $Z > 3.9$ ,  $p$ -value  $< 4.8e-5$ ). This pattern suggests the action of stabilizing selection on loop length, which in turn implies that the length of the loop is functionally important. We verified that this result was not due to a significantly elevated level of insertions in the flanking regions of miRNAs by checking that the rate of insertions in intergenic regions (8.9 / kb) was higher than in miRNA control regions (7.3 / kb). The rate of insertions across the entire genome was lower (6 / kb), presumably due to the lower rate of insertions in genes.

## 2.4 Small RNAs in *C. elegans* embryogenesis

NATURE METHODS | VOL.6 NO.10 | OCTOBER 2009

### Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression

Marlon Stoeckius<sup>1,4</sup>, Jonas Maaskola<sup>1,4</sup>, Teresa Colombo<sup>1,3</sup>, Hans-Peter Rahn<sup>1</sup>, Marc R Friedländer<sup>1</sup>, Na Li<sup>1</sup>, Wei Chen<sup>1</sup>, Fabio Piano<sup>2</sup> & Nikolaus Rajewsky<sup>1</sup>

**Abstract** *C. elegans* is one of the most prominent model systems for embryogenesis, but collecting many precisely staged embryos has been impractical. Thus, early *C. elegans*



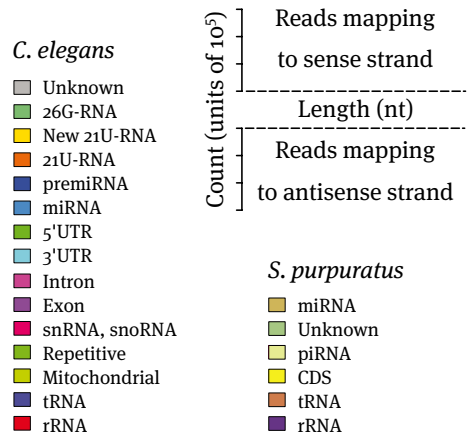
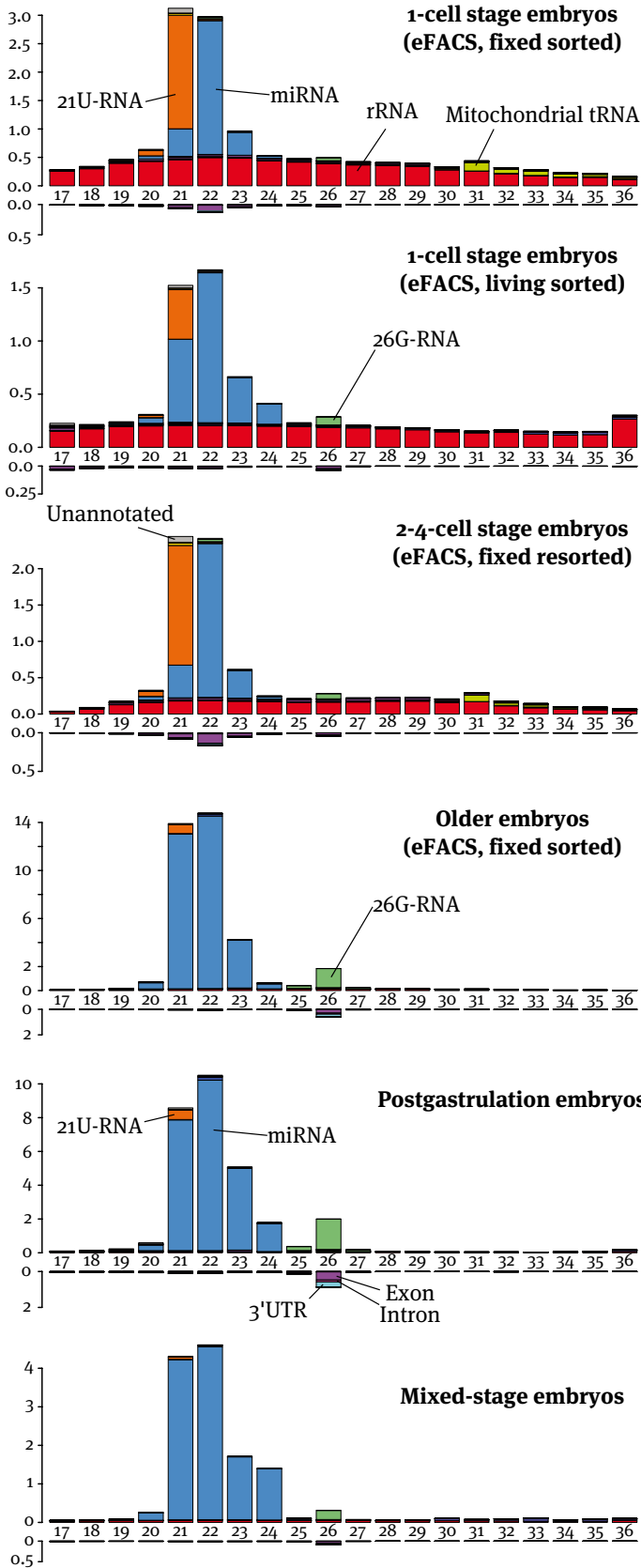
embryogenesis has not been amenable to most high-throughput genomics or biochemistry assays. To overcome this problem, we devised a method to collect staged *C. elegans* embryos by fluorescence-activated cell sorting (eFACS). In a proof-of-principle experiment, we found that a single eFACS run routinely yielded tens of thousands of almost perfectly staged 1-cell stage embryos. As the earliest embryonic events are driven by post-transcriptional regulation, we combined eFACS with second-generation sequencing to profile the embryonic expression of small ncRNAs. We discovered complex and orchestrated changes in the expression between and within almost all classes of small RNAs, including miRNAs and 26G-RNAs, during embryogenesis.

**Introduction** The nematode *C. elegans* is one of the best-explored model organisms for developmental biology. The mechanistic basis of embryogenesis in *C. elegans* has been dissected by describing the entire cell lineage (Sulston et al., 1983) and by performing many molecular and genetic analyses. Various key proteins involved in early cell division as well as hundreds of essential genes required for early embryogenesis and their knockdown phenotypes have been described (Fernandez et al., 2005; Gönczy and Rose, 2005; Kamath et al., 2003; Oegema and Hyman, 2006; Piano, Schetter, Mangone, et al., 2000; Piano, Schetter, Morton, et al., 2002; Sönnichsen et al., 2005). However, a true understanding of embryogenesis will require the knowledge of stage-specific gene expression. Modern high-throughput technologies such as deep sequencing, proteomics and their many applications can be used, for example, to identify and quantify the transcriptome, protein amounts and protein-protein interactions on a genome-wide scale. Prerequisite to the study of embryogenesis progression with many of these methods are large amounts of precisely staged embryos to yield enough RNA or other material. However, this is currently not possible. Isolated embryos are mixtures of embryos at developmental stages ranging from the early 1-cell zygote to the almost hatching worm larvae with approximately 600 cells. To date, staged embryos are usually obtained by manual sorting using a mouth pipette, making it impractical to apply large-scale techniques that require tens of thousands of embryos. Alternatively, one can obtain many semi-synchronized embryos by blocking their development with fluorodeoxyuridine (Stroeher et al., 1994), or one can isolate young embryos from hermaphrodites that have just begun to produce mature oocytes (Schauer and Wood, 1990). Although these methods can yield reasonable quantities of young embryos, the collected embryos are not synchronous, and these approaches cannot be used to investigate specific developmental stages.

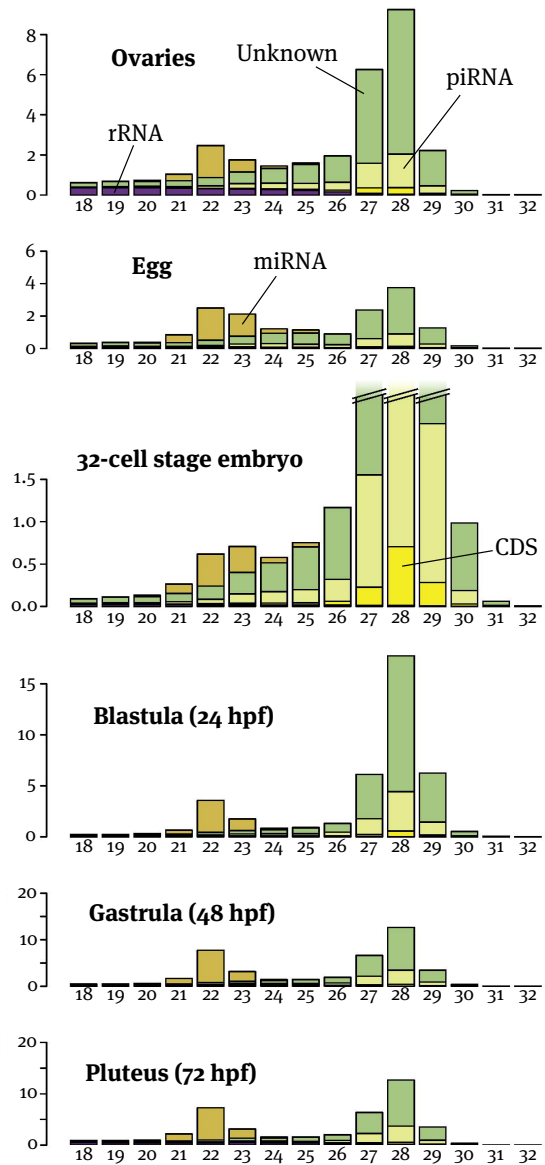
Here we describe a method to collect many precisely staged embryos by fluorescence-activated cell sorting (eFACS). As *C. elegans* embryos have the same size throughout development, eFACS can in principle be applied to any embryonic stage in which a specific fluorescent marker protein can be stably expressed. Thus, eFACS allows the resolution of embryonic stages with sufficient yield of embryos for high-throughput analyses that require large amounts of starting material.

In *C. elegans* embryos, some zygote-specific transcription is initiated at the 4-cell stage, although pharmacological and genetic experiments have suggested that zygotic genes are not required until later in embryogenesis (Edgar, Wolf, and Wood, 1994; Seydoux and Dunn, 1997). Maternal components seem sufficient to direct the embryo through the ini-

*C. elegans*



*S. purpuratus*



tial cleavage rounds up to approximately the onset of gastrulation. Interference with key enzymes involved in the RNA interference (RNAi) pathway lead to numerous defects including embryonic lethality, suggesting functional roles for ncRNAs in embryogenesis (Denli et al., 2004; Grishok et al., 2001; Knight and Bass, 2001). It is unknown which of the previously described small RNA populations in *C. elegans* (Ambros et al., 2003; Baugh et al., 2003; Evans and Hunter, 2005; Ruby, Jan, et al., 2006) such as miRNAs, endogenous small interfering RNAs (siRNAs), 21U-RNAs (thought to be germline-specific and characterized by a length of 21 nt, a strong bias for 5' uracils and their interaction with PIWI proteins) and the virtually uncharacterized class of 26G-RNAs (26 nt length and strong bias for a 5' guanine) are present in the early embryo, and it is unclear how the complexity and composition of the small RNA transcriptome changes during the very first cell cycles (Ambros et al., 2003; Baugh et al., 2003; Evans and Hunter, 2005; Ruby, Jan, et al., 2006). We thus set out to use eFACS in combination with deep sequencing to profile small RNA expression during early embryogenesis.

## My contributions

My contribution to this investigation consisted in the computational analysis of the deep-sequencing data. This included mapping the reads to the genome using a previously developed suffix array based alignment program, and classifying the reads according to the annotation of genomic features that the reads mapped to. For a description of the suffix array based alignment please refer to page 28f.

Subsequently, I mined the deep-sequencing reads for novel 21U-RNAs, quantified expression of small RNAs, and computed expression fold changes for them, which were validated by qPCR assays. Figure 2.1 displays the annotated small RNA read length distribution at different developmental time points.

**Mapping** Mapping of Genome Analyzer 2 (Illumina) deep-sequencing reads was performed using an in-house developed pipeline. This pipeline consists of an initial 3' adaptor removal step, low-complexity read filtering, a mapping routine using a suffix array

---

Figure 2.1 (*facing page*): Small RNA expression in early development of *C. elegans* (Stoekius et al., 2009) (left column) and *S. purpuratus* (Song et al., 2012) (right column), assayed by Solexa sequencing. Bars correspond to read lengths in nucleotides. Vertical extent corresponds to numbers of reads in units of  $10^5$ . Upwards oriented bars represent reads mapping in sense direction to annotated features. For *C. elegans*, antisense mapping reads are shown as downward oriented bars. For *S. purpuratus*, the displays are scaled so as to yield identical vertical extent for the 22 nt miRNA fraction. Conservatively only those reads are annotated as piRNAs in the *S. purpuratus* samples that occur paired with a 10 nt overlapping antisense mapping read. Consequently, many of the  $\approx 28$  nt reads annotated as Unknown or CDS may in fact be piRNAs. In the 32-cell stage piRNA expression is massively increased, and the full bar height is not shown in order to reveal the dynamics of the less highly expressed RNA moieties. The earlier four *C. elegans* samples were prepared by eFACS, sorting either once or twice, with or without prior fixing with methanol. The later *C. elegans* samples and the *S. purpuratus* samples were selected by time relative to fertilization (hpf: hours after fertilization).

based alignment program and a 3' adaptor identification refinement phase.

Briefly, initial adaptor removal was performed by using dynamic programming to find in each read the suffix that best matched to a prefix of the 3' adaptor. For this, all alignments of adaptor prefixes to suffixes of the read sequence were considered. In addition, occurrences of the full adaptor sequence anywhere in the read sequence were considered. Among these alignments, the best alignment was determined according to a simple one-parameter model  $p(\text{alignment}|\Theta) = \Theta^n(1 - \Theta)^{n-k}$ , where  $n$  is the length of the alignment,  $k$  is the edit distance of the alignment, and  $\Theta$  is a parameter describing the error rate. A  $\Theta$  value of 0.9 was heuristically chosen to reflect the relatively high error rate toward the end of Illumina reads.

The alignment program proceeds by determining all genomic matches to a read in edit distance  $k$ . For this application edit distance two was used. The alignment algorithm was implemented using a suffix array of the genome against which each read is sought, incrementally increasing the edit distance until matches are found.

In the 3' adaptor identification refinement phase, the boundary between transcript and adaptor parts of each read was redetermined in light of the genomic context that the read was mapped to. This was done by computing a score  $S(i) = f(i) + r(i)$  for every position  $i$  of the read.  $f(i)$  is derived by aligning prefixes of the genomic context to prefixes of the read, from which  $f(i)$  gives the edit distance of the best match of the read prefix of length  $i$  to the genomic context.  $r(i)$ , the second part of the score is determined from reverse alignments of the reversed read to reverse adaptor prefixes, that is,  $r(i)$  was the edit distance of the best match between the read sequence positions  $i + 1$  to  $n$  and a adaptor prefix, where  $n$  was the length of the read. The 3' adaptor beginning position  $t$  was then determined so as to minimize  $S(t)$ . In case of ties, the minimum of the tied positions was used.

For the subsequent analyses we used weighted matches, i.e., reads mapping to multiple loci have equal weight distributed across these loci (for example, a read represented by two transcripts and mapping equally well to three loci had a weight of two-thirds assigned to each of the three loci).

**Normalization of miRNA reads in-between samples** Expression fold changes of sequencing data were determined using linear models of the log expression, i.e. logistic expression models, as follows. Assume we are given expression vectors  $\mathbf{a}, \mathbf{b} \in \mathbb{R}_{\geq 0}^n$ , where  $n$  is the number of genes and  $\mathbb{R}$  is the set of real numbers, with  $\mathbf{a} = (a_i)_{i=1, \dots, n}$  and  $a_i$  the expression of gene  $i$ . If reference values  $z_a, z_b \in \mathbb{R}_{\geq 0}$  are known, then the normalized expression values are  $\frac{\mathbf{a}}{z_a}$  and  $\frac{\mathbf{b}}{z_b}$ , and the fold change  $fc_i \in \mathbb{R}_{\geq 0}$  of gene  $i$  is given by

$$fc_i = \frac{\frac{b_i}{z_b}}{\frac{a_i}{z_a}} = \frac{b_i z_a}{a_i z_b} = \text{constant} \cdot \frac{b_i}{a_i}.$$

Here 'constant' denotes an arbitrary constant determined by the ratio of the unknowns  $b_i$  and  $a_i$ . Thus, the log fold changes are

$$\log fc_i = \log b_i - \log a_i + \log z_a - \log z_b = \log b_i - \log a_i + \text{constant},$$

which is equivalent to

$$\log b_i = \log a_i + \log fc_i + \text{constant.} \quad (2.1)$$

Typically, the expression of a RNA species that is known to be constant between the two samples is used for the reference values. However, for the present study no such constants were known. We resorted to fitting a linear model of the form response = predictor + residual + intercept to equation (2.1), in which  $\log a_i$  is the predictor,  $\log b_i$  the response, the intercept term determines the ratio of normalizers, and finally the log fold changes correspond to the residuals. In fitting, the slope of the linear model is fixed to unity, essentially only fitting the intercept term. From the fitted models the log fold changes are found as the prediction residuals. Owing to the slope of unity, it is possible to trivially accumulate the pairwise intercept terms of a sequence of expression samples for a joint normalization.

The proposed normalization method is equivalent to assuming that the mean log fold change is zero. The calculated fold changes of miRNA expression by this normalization method were validated by qPCRs and showed to be in good agreement (see Stoeckius et al., 2009, Fig. 4).

**Classification and quantification of small RNA deep-sequencing reads** Known miRNA coordinates were retrieved from miRBase release 12 (Griffiths-Jones et al., 2008). Other ncRNA, 3'UTR, 5'UTR, exon and intron coordinates were retrieved from WormBase, matching genome release WS190<sup>4</sup>. The 21U-RNA coordinates were retrieved from the supplementary materials of previous studies (Batista et al., 2008; Ruby, Jan, et al., 2006). Coordinates of RepeatMasker annotations (Jurka, 2000) and simple repetitive sequences (G. Benson, 1999) were obtained from the UCSC genome browser (Karolchik, R. M. Kuhn, et al., 2008). Reads were annotated by intersecting the mapped coordinates subsequently with the following sets of annotated feature coordinates and subtracting intersecting coordinates before proceeding with the next annotation set. The order in which annotation categories were used was: (i) miRNA, (ii) rRNA, (iii) tRNA, (iv) snRNA, (v) snoRNA, (vi) 21U-RNA, (vii) mRNA and (viii) repetitive sequences. This order roughly reflects the number of genomic bases represented by the different classes. For each annotated feature, all overlapping mapped reads were determined, and quantification was done by summing the overlapping weighted matches.

**Identification of new 21U-RNAs** New 21U-RNAs were predicted using the motif scoring modules provided by Ruby, Jan, et al. (2006), which use position-specific nucleotide frequency matrices of the large and small 21U-RNA upstream motifs, as well as a model for the distance between the two motifs to determine occurrences of 21U-RNA loci. The matrices are parameterized from ungapped alignments of reads deriving from manually selected portions of the genome that are rich in 21U-RNA. The motif scoring was applied to the set of mapped loci that remained after removing other known ncRNAs (including previously known 21U-RNA loci) in which the sequences scored consisted of the upstream

<sup>4</sup><http://www.wormbase.org>; WS190

100 nt and the read itself. We used the same score cutoff of 15.5 to call loci as was used previously (Ruby, Jan, et al., 2006).

We observed expression from 7506 of 15341 known 21U-RNA loci (Stoeckius et al., 2009, Supplementary Tables 5,6). Reads mapping to known 21U-RNA loci derived almost exclusively from the sense strand, had almost always a 5' uracil, and their length distribution sharply peaked at 21 nt. We discovered 389 new 21U-RNAs (Stoeckius et al., 2009, Supplementary Table 7). Their genomic distribution followed the published pattern (Batista et al., 2008; Ruby, Jan, et al., 2006) with additional dispersed genomic loci.

**Differential expression across and within small RNA classes** We next compared the expression of all known classes of small RNAs during embryogenesis. However, we note that we most likely only observed small RNAs with a 5' monophosphate owing to the cloning protocol. Overall, we observed strong, orchestrated changes in the composition of small RNAs between the sequenced samples (figure 2.1). Older embryos were dominated by miRNAs whereas in very early stages we observed additional small RNA classes. Those include mitochondrial tRNA as well as a sizable fraction of rRNA. The rRNA- and tRNA-derived fractions in all samples had a uniform length distribution and thus were likely to be degradation products. The 21U-RNAs were highly expressed in early embryos but difficult to detect in older embryos. We also observed differential expression of endo-siRNAs and 26G-RNAs. The relative abundance of small RNAs in mixed-stage embryo samples convoluted specific changes in small RNA expression during embryogenesis (figure 2.1).

**Endogenous siRNAs are observed in the 1-cell stage embryo** The length distribution of reads mapping sense or antisense to exons or introns of mRNA transcripts varied distinctly (figure 2.1). Sense reads were distributed uniformly, suggesting that they originated from degraded mRNAs. Antisense reads mapping to exons were dominated by 22 nt and 26 nt reads with a strong bias for a 5' uracil or guanine, respectively (consistent with previous reports Ambros et al., 2003; Ruby, Jan, et al., 2006). We will refer to the corresponding small RNAs as endogenous siRNAs (endo-siRNAs). Most 1-cell stage embryo endo-siRNAs mapped to mitochondrial enzymes. The majority of these mRNAs are known to be up-regulated in RNAi pathway defects (*rrf-1*, *eri-1*, *rde-3* and *dcr-1* mutants), which suggests that they are under control of small RNAs (Stoeckius et al., 2009, Supplementary Table 8). We also consistently observed possible degradation products of mitochondrial tRNAs in the early embryo but not in other samples (figure 2.1). Notably, we found more ~22 nt endo-siRNA in the 1-cell stage and 2–4-cell stage embryos, whereas ~26 nt endo-siRNA dominated in the older samples. Additionally, we observed in older embryos a twofold enrichment of antisense reads mapping to 3' untranslated regions (UTRs) (27–32%) when compared to 1-cell or 2–4-cell stages (15%).

**Genomic organization and expression of 26G-RNAs** After removing known RNA classes, we studied the set of remaining reads. The length distribution of these RNAs peaked at 26 nt and were most highly expressed in the older embryonic stages. These 26-mers did not map to any annotated loci and had a strong 5' guanine bias (75.7%). Hereafter, we refer to 26 nt reads with a 5' guanine as 26G-RNAs (Ghildiyal and Zamore, 2009). Although these

26G-RNAs were present only in low numbers in early embryos, we observed high 26G-RNA expression in older embryos. Computational analyses revealed that 26G-RNAs mapped to several clusters in intergenic regions on different chromosomes (Stoeckius et al., 2009, Fig. 6a). We validated five (out of five tested) 26G-RNAs from two clusters (Stoeckius et al., 2009, Fig. 6b).

**Discussion** Previous large-scale studies of small RNA expression had used samples composed of mixed-stage embryos. These studies could not detect the orchestrated and dynamic changes between and within different classes of small RNAs that we observed when comparing the 1-cell stage embryos to later stages. First, the majority of miRNAs is already expressed in the 1-cell stage embryo, suggesting that they are maternally deposited. The reason remains to be determined. Second, we showed that miRNAs from the miR-35 cluster are likely early embryo-specific. Genetic knockouts and mutations for 95 miRNAs have been published (Miska et al., 2007). Notably, the miR-35 cluster is the only known miRNA cluster with an embryonic lethal knockout phenotype. Third, we observed many small RNAs of uniform length mapping sense to rRNAs in 1-cell stage embryos (live-sorted or methanol-fixed), with decreased expression in 2–4-cell embryos, but virtually absent in samples from older stages. Thus, although we do not have independent validation, it seems unlikely that the observed rRNA expression is an experimental artifact. rRNAs, unlike mRNAs, are already transcribed in the 1-cell stage embryo (Seydoux and Dunn, 1997). One may speculate about a turnover of maternally and paternally provided rRNAs to zygotically transcribed rRNAs upon fertilization during very early embryogenesis. Finally, we found consistent evidence for a turnover of mitochondrial components in the 1-cell stage embryo. We observed degradation products of mitochondrial tRNAs in the early embryo as well as many siRNAs directed against mitochondrial enzymes. Thus, it is tempting to speculate about mechanisms that selectively degrade paternal mitochondria in early zygotes, as described in vertebrates (Sutovsky, 2003).

Our data allowed us to study as yet virtually undescribed classes of small RNAs such as 26G-RNAs. Observations of small RNAs, in particular ~26-nt-long with a 5' guanine bias have been reported earlier (Ambros et al., 2003; Ruby, Jan, et al., 2006) and were recently dubbed 26G-RNAs (Ghildiyal and Zamore, 2009). We found that 26G-RNAs are dynamically expressed and that they cluster in several intergenic regions. Northern blot analysis suggested that they may be initially generated with heterogeneous lengths or post-transcriptionally modified such that they appear as having different sizes on the northern blot. In addition to an increase in expression of 26G-RNAs in older embryos, we also observed increased expression of 26-nt endo-siRNAs mapping to the antisense strand of coding mRNAs. We did not computationally detect a 'ping pong' biogenesis mechanism (Aravin, Sachidanandam, et al., 2007; Brennecke et al., 2007) between 26G-RNAs and 26-nt endo-siRNAs.

Our eFACS data and analyses raise many more questions. However, altogether we are tempted to conclude that the complexity of small RNA expression dynamics in very early embryogenesis is comparable to the expression dynamics of protein-coding genes, and that the use of eFACS will contribute to a more complete understanding of gene regulatory networks during early animal development.



## 2.5 Small RNAs in *S. purpuratus* early development

Genomes &amp; Developmental Control

Developmental Biology 362 (2012) 104–113

Select microRNAs are essential for early development in the sea urchin

Jia L. Song <sup>a,d,\*</sup>, Marlon Stoeckius <sup>c</sup>, Jonas Maaskola <sup>c</sup>, Marc Friedländer <sup>c,1</sup>, Nadezda Stepicheva <sup>d</sup>,  
Celina Juliano <sup>a,2</sup>, Svetlana Lebedeva <sup>c</sup>, William Thompson <sup>b</sup>, Nikolaus Rajewsky <sup>c,\*\*</sup>, Gary M. Wessel <sup>a,\*\*\*</sup>

**Abstract** miRNAs are small ncRNAs that mediate post-transcriptional gene regulation and have emerged as essential regulators of many developmental events. The transcriptional network during early embryogenesis of the purple sea urchin, *S. purpuratus*, is well described and can serve as an excellent model to test functional contributions of miRNAs in embryogenesis. We examined the loss of function phenotypes of major components of the miRNA biogenesis pathway. Inhibition of *de-novo* synthesis of Drosha and Dicer in the embryo led to consistent developmental defects, a failure to gastrulate, and embryonic lethality, including changes in the steady state levels of transcription factors and signaling molecules involved in germ layer specification. We annotated and profiled small RNA expression from the ovary and several early embryonic stages by deep sequencing followed by computational analysis. miRNAs as well as a large population of putative piRNAs had dynamic accumulation profiles through early development. Defects in morphogenesis caused by loss of Drosha could be rescued with four miRNAs. Taken together our results indicate that post-transcriptional gene regulation directed by miRNAs is functionally important for early embryogenesis and is an integral part of the early embryonic gene regulatory network in *S. purpuratus*.

**Introduction** Small RNAs are components of a conserved gene regulatory mechanism that includes miRNAs, siRNAs and piRNAs. miRNAs negatively regulate protein expression by binding to sequence-complementary target sites in mRNAs which induces repression of mRNA translation or transcript destabilization and decay (Bartel, 2009; Brodersen and Voinnet, 2009; Ghildiyal and Zamore, 2009; Guo et al., 2010; Hendrickson et al., 2009; Rajewsky, 2006, 2011). In animals, miRNAs have thousands of targets and altogether regulate a major portion of protein coding genes (Baek et al., 2008; Bartel, 2009; R. C. Friedman et al., 2009; Krek et al., 2005; Lewis, Burge, and Bartel, 2005; Selbach et al., 2008; Stark, Brennecke, et al., 2005; Xie et al., 2005). The vast majority of miRNAs are initially processed by Drosha and its cofactor DGCR8 (Han et al., 2006; Y. Lee et al., 2003) and the maturation of miRNAs and siRNAs requires Dicer. Dicer is a member of the RNase III endoribonuclease family and is responsible for processing double stranded RNA (dsRNA) to siRNAs during RNAi (H. Zhang et al., 2002). It is also the key enzyme that mediates the final processing of most miRNAs from their precursors.

A number of fundamental steps in embryogenesis appear to be regulated by miRNAs and while the documentation of gene regulatory networks involved in cell fate specification and differentiation has revealed the importance of numerous signaling molecules and transcription factors, the diverse regulatory roles of miRNAs in early development are only now emerging (reviewed in Fabian, Sonenberg, and Filipowicz, 2010; Ghildiyal and Zamore, 2009; Pauli, Rinn, and Schier, 2011). Recently a number of miRNAs were



identified in the purple sea urchin, *S. purpuratus* (Campo-Paysaa et al., 2011; Friedländer, Mackowiak, et al., 2012; Peterson, Dietrich, and McPeck, 2009; Wheeler et al., 2009), revealing many deeply conserved miRNAs also present in humans. Echinoderms are a sister group to the chordates and the function of miRNAs in these embryos may reflect transitions in deuterostome development. Armed with the in-depth knowledge of transcriptional gene regulatory networks in the sea urchin<sup>5</sup>, we set out to investigate the importance of miRNAs in early embryogenesis of this animal. We profiled and annotated small RNA expression from the ovary and several early embryonic stages by deep sequencing followed by computational analysis, including application of the miRNA identification tool "miRDeep" (Friedländer, W. Chen, et al., 2008; Friedländer, Mackowiak, et al., 2012). Individual knockdowns of Dicer, Drosha and DGCR8 as well as miRNA rescue experiments suggest that the miRNA pathway plays an important functional role in early cell fate decisions of sea urchin embryogenesis and serves as a paradigm for an ancestral feature of the deuterostome lineage.

### My contributions

My contribution to this investigation was similar to that for the Stoeckius et al. (2009) investigation. Together with Marc Friedländer I mapped and classified the deep-sequencing reads. After initially mapping the data to the genome, we found the sea urchin genome annotation lacking, which made it difficult to classify reads by genomic annotations. Instead, we mapped the reads directly to sets of sequences of genomic features using the Mapper module of miRDeep2 (Friedländer, Mackowiak, et al., 2012). We used a similar hierarchy, as in Berninger et al. (2008) and Stoeckius et al. (2009), specifically, miRNA > mRNA > tRNA > rRNA > unknown. To identify miRNA reads, we mapped to *S. purpuratus* miRNA precursors from miRBase version 16 and to three novel precursors predicted using miRDeep2 (Friedländer, Mackowiak, et al., 2012). To identify mRNA reads, we mapped to the gene\_cds (coding) sequences from SpBase<sup>6</sup>. To identify tRNA reads we analyzed the Spur\_v2.1 genome with tRNAscan-SE-1.23 using default eukaryotic parameters and mapped to the predicted tRNA sequences. To identify rRNA reads we mapped to the *S. purpuratus* 18S and 28S rRNA sequences obtained at GenBank (D. A. Benson et al., 2004). Last, all reads that mapped to the Spur\_v2.1 genome but did not map to any of the above annotations were labeled 'unknown'. Reads were mapped with Bowtie (Langmead et al., 2009) with these options: -f -n 1 -e 80 -l 18 -a -best -strata. Using this stringent procedure we successfully mapped between 43% and 55% of the clipped reads in each of the six datasets, corresponding to between 1.7 and 4.4 million reads. Even though some reads may not have been mapped because of the incomplete state of the genome assembly, these numbers were comparable to previous small RNA studies (Mayr and Bartel, 2009; Persson et al., 2009), showing the consistent high quality of the data. We annotated piRNAs in the following way: all reads that mapped to the genome but did not map to existing annotations ('unknown' reads) were pooled across samples. Using a custom ruby script we identified all instances where two of these reads overlap with each other such that they are

<sup>5</sup>see [www.spbase.org/endomes](http://www.spbase.org/endomes)

<sup>6</sup><http://sugp.caltech.edu/SpBase/download/>

on opposite genomic strands and their 5' ends overlap by exactly ten nucleotides. Since this overlap is in perfect accordance with the 'ping-pong' model of piRNA biogenesis, we annotated all such read pairs as piRNAs.

**Dynamic small RNA expression** We investigated the length distribution and annotation of all sequencing reads. Small RNAs showed a bimodal length distribution with 2 distinct peaks around 22 and 28 nucleotides (figure 2.1). All miRNAs identified by miRDeep account for a characteristic peak around 22 nucleotides. Interestingly, we found that most of the sequenced sea urchin RNAs that do not map to existing annotations have a distinct length profile peaking at 28 nucleotides, as has been observed for piRNAs in other species. piRNAs are associated with silencing of transposable elements in the germline and have recently been shown to be involved in maternal mRNA deadenylation in the early embryo thus mediating the maternal-to-zygotic transition (Rouget et al., 2010). Further, we found that a large portion of these RNAs tend to overlap with each other by exactly ten nucleotides, with one read exhibiting a uridine bias at the 5' end and the other an adenine bias at the tenth nucleotide. These features are consistent with the conserved 'ping-pong' piRNA biogenesis pathway via mutual cleavage of the sense and antisense piRNA precursors by the Piwi proteins (Aravin, Gaidatzis, et al., 2006; Brennecke et al., 2007; Girard et al., 2006; Grivna et al., 2006; Gunawardane et al., 2007; Houwing et al., 2007; Saito et al., 2006; Vagin et al., 2006; Watanabe et al., 2006). We therefore annotate the RNAs that overlap by exactly ten nucleotides as piRNAs and refer to the remaining small RNA species of around 28 nucleotides that do not overlap by ten nucleotides as 'unknown' sequences, although their length distribution suggests that they are likely highly enriched in piRNAs (figure 2.1).

We observed a significant decrease in total reads mapping to miRNAs at the 32-cell stage. This was correlated with an increase of reads mapping to putative piRNAs (figure 2.1). As distributions of sequenced reads do not reflect absolute abundance but rather relative frequencies, two possible interpretations to the 32-cell stage transition are that miRNAs are either cleared from the egg following fertilization, or that piRNAs strongly increase at the 32-cell stage. To distinguish these possibilities, we performed Northern blots for selected piRNA candidates. We observed a pronounced increase in 4 of the 5 detectable piRNAs in the 32-cell stage (Song et al., 2012, Fig. S2). Moreover, RT-qPCR analysis did not illustrate a drastic decrease of the tested miRNAs in the 32-cell stage (Song et al., 2012, Fig. S1). Taken together the results suggest that piRNAs have a dynamic expression pattern in the early sea urchin embryo. This increase of piRNA expression may correspond with the specification of the piwi-positive small micromere lineage (C. E. Juliano et al., 2006; Rodriguez et al., 2005).

## 2.6 CHIP-Sequencing a cell-cycle regulator and a helicase

The EMBO Journal (2012) 31, 972–985

### The SNF2-like helicase HELLS mediates E2F3-dependent transcription and cellular transformation

Björn von Eyss<sup>1,6</sup>, Jonas Maaskola<sup>2,6</sup>, Sebastian Memczak<sup>1,6</sup>, Katharina Möllmann<sup>1</sup>, Anja Schuetz<sup>3</sup>, Christoph Loddenkemper<sup>4</sup>, Mai-Dinh Tanh<sup>1</sup>, Albrecht Otto<sup>1</sup>, Kathrin Muegge<sup>5</sup>, Udo Heinemann<sup>3</sup>, Nikolaus Rajewsky<sup>2</sup> and Ulrike Ziebold<sup>1,\*</sup>

**Abstract** The activating E2F-transcription factors are best known for their dependence on the Retinoblastoma protein and their role in cellular proliferation. E2F3 is uniquely amplified in specific human tumours where its expression is inversely correlated with the survival of patients. Here, E2F3B interaction partners were identified by mass spectrometric analysis. We show that the SNF2-like helicase HELLS interacts with E2F3A *in vivo* and cooperates with its oncogenic functions. Depletion of HELLS severely perturbs the induction of E2F-target genes, hinders cell-cycle re-entry and growth. Using ChIP-Seq, we identified genome-wide targets of HELLS and E2F3A/B. HELLS binds promoters of active genes, including the trithorax-related MLL1, and co-regulates E2F3-dependent genes. Strikingly, just as E2F3, HELLS is overexpressed in human tumours including prostate cancer, indicating that either factor may contribute to the malignant progression of tumours. Our work reveals that HELLS is important for E2F3 in tumour cell proliferation.

**Introduction** The E2F-transcription factor family is well conserved and widely known for its role in proliferation and cell-cycle progression (van den Heuvel and Dyson, 2008). Among this family, E2F1–E2F3 are most intriguing, since these E2Fs directly associate with and are antagonized by the pRB (retinoblastoma) tumour suppressor protein (Burkhart and Sage, 2008). Genetic mouse models provided first evidence that E2F3 largely contributes to pRb-attributed tumorigenic events, such as the reduction of *Rb*-deficient pituitary adenomas in E2F3-deficient mice (Ziebold et al., 2003), but also pre-neoplastic lesions of the lung (T. Parisi et al., 2007). In addition, the loss of E2F3, in combination with E2F1 and E2F2, reduces hyperplasia in intestinal epithelia (Chong et al., 2009). Also in human patients, aberrant E2F3-expression was linked to cancer. In some of these tumours such as invasive urinary bladder carcinoma, it was suggested that the loss of pRB as well as amplification of E2F3 are obligate events for tumorigenesis (Feber et al., 2004; Hurst et al., 2008; Oeggerli et al., 2004). In advanced prostate cancers, E2F3 is highly overexpressed and the highest E2F3 levels determine the worst clinical outcome for the patient (Foster et al., 2004). Apart from inappropriate E2F3-expression, the whole pRB/E2F-pathway is regarded as being deregulated in nearly every human tumour (Hanahan and Weinberg, 2011).

In the past, the transcriptional programmes of specific E2Fs were established after overexpression of E2Fs in mammalian cells and subsequent microarray (DeGregori and D. G. Johnson, 2006). These analyses led to the identification of many DNA synthesis, checkpoint control, DNA repair and apoptosis-associated targets and suggested that there are similarities but also specificity among the individual E2Fs (Kong et al., 2007). Using

the more sophisticated ChIP assays coupled to microarrays, it was concluded that there are functional overlaps between the E2Fs, but also that isoform-specific functions exist (Asp et al., 2009; Ren, Cam, et al., 2002; Y. Takahashi, Rayman, and Dynlacht, 2000). Owing to these studies many new E2F-targets were identified, but the impact of most targets for the oncogenic E2F function is still elusive.

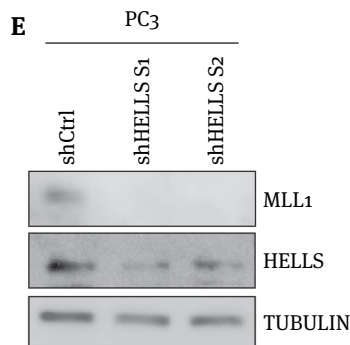
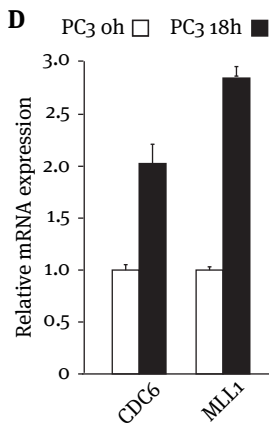
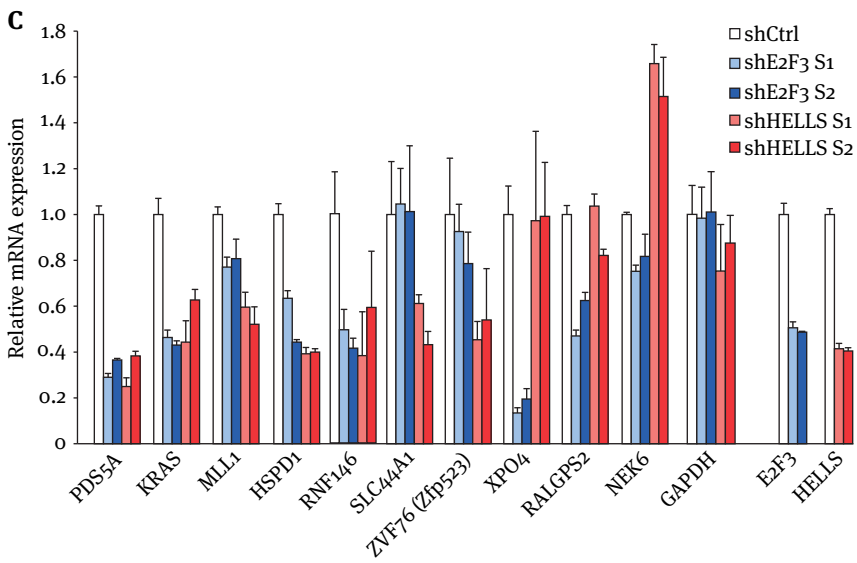
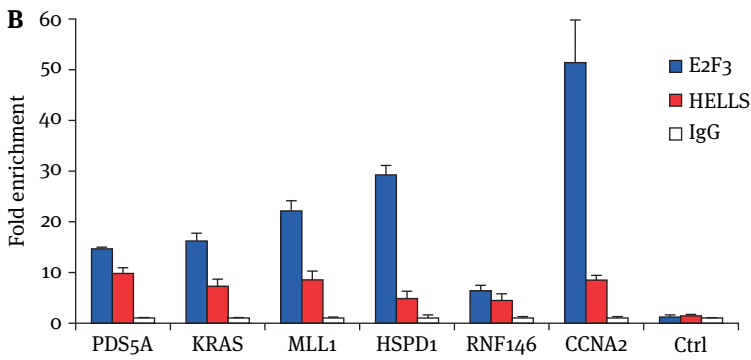
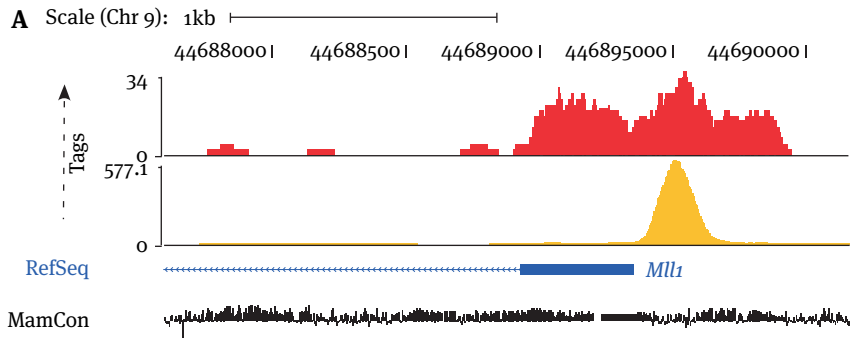
Recently, more attempts were to deconstruct the mechanistic basis of the pRB/E2F complex (Blais and Dynlacht, 2007). Various molecules were identified that are tethered by pRB/E2F, acting both locally and globally, to modify the chromatin of E2F-dependent promoters. Most of the numerous pRB-interacting molecules, such as the histone methyltransferase SUV39H1 contribute to repression. pRB-tethered SUV39H1 leads to heterochromatinization and silencing of S-phase genes (Narita et al., 2003; Nielsen et al., 2001). In order for a cell to proceed through the cell-cycle E2F-dependent promoters have to be inverted from this repressive chromatin state into a state favouring transcription. E2F1–E2F3 commence this transition by recruiting chromatin regulators such as HCF-1 in complex with the mixed-lineage leukaemia (MLL1) protein that possess methyltransferase activity towards histone H3 lysine 4, H3K4 (Tyagi et al., 2007). To alleviate the repressive state of the chromatin, HCF-1 also enlists the histone lysine demethylase PHF8 (W. Liu et al., 2010). That a majority of promoters bound by PHF8 overlap with E2F-bound promoters was demonstrated by ChIP-Seq. Anticipating that all molecules that assist the E2F-transactivating function or revert the state of the chromatin at E2F-dependent promoters harbour immense potential as drug targets in human tumours, we explored such molecules in the context of oncogenic E2F3.

Here, we identified the SNF2-like, ‘lymphoid-specific’ helicase HELLS<sup>7</sup> as a novel E2F3-interaction partner. Previously, it was shown that HELLS is required for the silencing of constitutive heterochromatic regions such as retrotransposons or the pericentromer (Den-

<sup>7</sup>HELLS is also called SMARCA6, PASG or LSH.

Figure 2.2 (*facing page*): E2F3 and HELLS regulate many common targets, most notably MLL1. **(A)** ChIP-Seq tag profiles of E2f3 (red) and Hells (yellow) in WT MEFs in a screenshot from the UCSC genome browser displaying number number of ChIP-Seq tags mapping to the *Mll1* genomic context. Tag profiles were created by extending each read to a length of 100 bases, and normalizing to 10 million mapped reads. RefSeq transcripts and mammalian sequence conservation (MamCon) are also shown. **(B)** Confirmatory ChIP analyses of HELLS target genes in DU145 prostate carcinoma cells. The enrichment of E2F3 and HELLS was assessed using *PDS5A*, *KRAS*, *MLL1*, *HSPD1*, and *RNF146* primers spanning the genomic regions around the TSS. *CCNA2* is an E2F-dependent promoter and *U2* served as negative promoter (Ctrl) and IgG served as antibody control. **(C)** qRT-PCR (qPCR) analysis depicting the relative expression of representative HELLS target genes in human prostate carcinoma cells (PC3) after lentiviral depletion of E2F3 or HELLS using two independent sequences each. The knockdown was verified by qPCR analysis of *E2F3* and *HELLS*. The transcriptional regulation of *PDS5A*, *KRAS*, *MLL1*, *HSPD1*, *RNF146*, *SLC44A1*, *ZNF76* (*M. musculus Zfp523*), *XPO4*, *RALGPS2* and *NEK6* was analyzed. *GAPDH* served as a control. **(D)** *MLL1* was analyzed by qRT-PCR in synchronized PC3 cells, which were serum-arrested and harvested at 0 or 18 h after serum addition. *CDC6* is a S-phase-specific control. **(E)** Effects of depleting HELLS in PC3 cells determined by western blots for MLL1, HELLS and TUBULIN, using two independent hairpins (shHELLS S1 or S2) or control infections (shCtrl).

2.6. CHIP-SEQUENCING A CELL-CYCLE REGULATOR AND A HELICASE



nis et al., 2001; L. Sun et al., 2004) via an interaction with DNA methyltransferases (Myant and Stancheva, 2008). Now, we demonstrate that HELLS binds to E2F3 *in vivo*, aiding induction of E2F-target genes and cell-cycle re-entry. Furthermore, we provide evidence that HELLS, akin to E2F3, is overexpressed in several human tumours. In prostate carcinomas, HELLS/E2F3 co-expression is marking the most aggressive stages. The vast overlap of the identified co-bound promoters suggests that E2F3 and HELLS co-regulate target genes. Notably, a highly HELLS-enriched promoter is the trithorax-related *MLL1*. We suggest that in cancer cells *MLL1*, that is important for histone methylation and activation of key cyclin genes, depends itself on E2F3:HELLS.

## My contributions

Before describing my contributions to the published part of this investigation, I would like to refer the reader to an anecdote about a memorable episode during this investigation that is described in appendix L.2.

**Solexa read processing and mapping** Solexa reads were base called using the manufacturer's software SCS 2.5 / RTA 1.5. Subsequently, 3' ligation adapter sequences were identified and removed from the reads. Reads were clipped to 25 nt, and reads shorter than 17 nt were discarded. The remaining reads were mapped against the mouse genome (NCBI37/mm9) allowing up to edit distance 1 per read using a custom read mapping pipeline. Only reads mapping uniquely to the genome were considered. The mapped reads were reduced to non-redundant tag position sets.

**Adapter removal** To remove 3' ligation adapter sequences, all adapter prefix occurrences up to edit distance 4 were considered. The best-matching adapter prefix occurrence of each read was identified using a binomial model with parameters  $p$ ,  $n$ , and  $k$ , where  $p$ , the average sequencing accuracy, was set to 0.95, and  $n$  is the length of the prefix occurrence and  $k$  is the edit distance, i.e. the number of mismatches and indels (=insertions and deletions) in the alignment of the adapter sequence to the adapter prefix occurrence in the read. For each read we removed the prefix occurrence that had the highest log-likelihood according to this model.

**Tag density normalization** To compare distributions of aligned deep-sequencing reads between the different libraries we applied the following normalization procedure. First, aligned tag position sets were smoothed by applying a sliding window of size 100 nt, calculating for each position the relative frequency of positions with aligned tags in the window centered on that position. This relative frequency was divided by the number of aligned tag positions and multiplied by 10 million to yield mapping density units of tags per 10 million aligned tags.

**Tag density analysis around transcription start sites** A non-redundant set of RefSeq (Pruitt, Tatusova, and Maglott, 2005) TSS was extracted from the UCSC table browser

(Karolchik, Hinrichs, et al., 2004). For each library, we determined the expected tag densities per 10 million aligned tags for each position in windows of size 1.5 around the TSS. From these densities we computed the relative enrichment of each library to the IgG library in the same genotype.

**Identification of binding regions** The mapped reads were reduced to non-redundant tag position sets. Regions enriched for E2F3 and HELLS were determined using MACS (Y. Zhang et al., 2008) version 1.4.0beta. The IgG mapped read sets were pooled and used as control in the analyses. Default parameter settings were used.

**Determining target genes** Target genes were defined by finding transcription start sites situated closer than 1 kb to summits of enriched regions according to the MACS analysis of the ChIP-Seq libraries.

**Analysis of H3K4me3 data** MEF histone modification data of Mikkelsen et al. (2007), in particular H3K4me3, were retrieved from the supplement of Young et al. (2011). These data are given in terms of Ensembl gene IDs, that we mapped to RefSeq genes. We performed the analysis of overlapping target genes and H3K4me3 on the level of RefSeq genes.

**HELLS and E2F3 are enriched at a large fraction of promoters** The joint regulation of cell-cycle entry by E2F3 and HELLS and their co-expression in human tumours are strong evidence that these proteins share functional overlaps. To reveal the full extent of the genome-wide chromatin association of E2f3 and Hells, we applied ChIP-Seq in mouse fibroblasts. Since Hells was previously reported to localize to heterochromatin (Yan et al., 2003) and enforce DNA methylation and silencing of selected Hox genes (Xi et al., 2007), we expected to detect Hells also at inactive, repressed loci. Such regions are also targets of the polycomb group-mediated repression and characterized by trimethylation of histone H3 at lysine 27, H3K27me3 (Sawarkar and Paro, 2010). Thus, the genome-wide distribution of Hells and E2f3 was assessed in WT MEFs and compared with that of the polycomb-associated, repressive H3K27me3 mark. The ChIP-Seq data were mapped to the genome and subjected to a bi-modal transcription factor type peak analysis to identify enriched regions (von Eyss et al., 2012, supplementary data). The highest number of significant peaks within 1 kb of RefSeq TSSs were found for E2f3 (von Eyss et al., 2012, figure 6A; supplementary tables SII and SIII). Importantly, only 22% of the Hells targets coincide with H3K27me3, but 93% of all Hells-associated TSS coincide with E2f3-bound TSS (von Eyss et al., 2012, figure 6B). These findings indicate that Hells is more likely to co-bind E2f3 targets than the H3K27me3-enriched, presumably 'repressed' promoters. For a graphical overview of E2f3 and Hells peaks, we chose the chromosomal context of *Mll1* (figure 2.2A) because for Hells it is the highest-ranking region according to the ChIP-Seq data (von Eyss et al., 2012, supplementary table SIII). The ChIP-Seq data confirm our conventional ChIP analysis that Hells and E2f3 co-occupy overlapping sets of promoters. Knowing that E2F3 have strong preference for genomic regions closest to the TSS (DeGregori and D. G. Johnson, 2006), we assessed the global positional preference of E2f3 and Hells (von Eyss et al.,



2012, figure 6D). As anticipated, E2F3 sharply concentrates at TSS. Also for Hells, the ChIP-Seq summits clustered closely to the TSS, but with a preference towards a region of about 120 nt downstream of the TSS. Having observed that there is a global overlap of Hells and E2F3-bound targets, we determined if the co-bound target genes are likely to be active or repressed by assessing the overlap with previously mapped regions that are enriched for histone H3 trimethylated at lysine 4, H3K4me3 (Mikkelsen et al., 2007). As expected, the majority of E2f3 targets are also enriched for H3K4me3 (von Eyss et al., 2012, figure 6E). Importantly, 86% of the Hells targets are likely to be actively transcribed, since they also show H3K4me3 enrichment at their promoter.

**The activities of HELLS and E2F3 are essential to induce MLL1 and other targets to create a proliferative circuitry**

Next, confirmatory ChIP analysis was performed in DU145 prostate cancer cells, since here the E2F3:HELLS activity is physiologically relevant. Consistently, both E2F3 and HELLS were enriched at the promoter of *MLL1* and other ChIP-Seq targets such as *PDS5A*, *KRAS*, *HSPD1* and *RNF146* (figure 2.2B; von Eyss et al., 2012, supplementary figure S5). We used the established E2F-target *CCNA2* that was previously shown to depend on HELLS (von Eyss et al., 2012, figure 3C), as a positive control. Also in the mouse fibroblasts, Hells ChIP-Seq reads were found at the *Ccna2* promoter (von Eyss et al., 2012, supplementary figure 6C). Next, we used human prostate carcinoma cell lines again to determine if in this context the HELLS activity would affect the expression of *MLL1* or any of the other targets (von Eyss et al., 2012, supplementary table SII). Either HELLS or E2F3 was acutely depleted by lentiviral shRNA and the knockdown of E2F3 or HELLS was verified using qRT-PCR. Importantly, HELLS-depletion led to a reduction of *MLL1* and also of *PDS5A*, *KRAS*, *HSPD1*, *RNF146*, *SLC44A1*, *ZNF76* (*Mus musculus Zfp523*), *XPO4* and *RALGPS2*. Only *NEK6* expression was elevated upon HELLS loss (figure 2.2C). This is in stark contrast to the effect of Hells loss on *MCM4/6* or *CDC6* shown before (von Eyss et al., 2012, figure 3A). Importantly, also E2F3 is needed for normal *MLL1* levels and also regulates the expression of most of the other targets, except *SLC44A1*. Since we also detected a diminished expression of *MLL1* and other HELLS targets in T98G cells after HELLS-depletion (von Eyss et al., 2012, supplementary figure S7A–C), we assessed whether the expression pattern of *MLL1* resembles that of typical E2F-targets. In synchronized human prostate cancer cells, *MLL1* was induced more strongly than *CDC6* upon serum-induced cell-cycle re-entry (figure 2.2D). Given that the expression of the *MLL1* RNA not always correlates with the protein level (H. Liu, Cheng, and Hsieh, 2007), we tested the effect of HELLS-depletion upon the expression of the *MLL1* protein in the same prostate cells and clearly found *MLL1* to be reduced (figure 2.2E). Furthermore, the *Mll1* RNA and protein also declined in E2f3-deficient mouse cells, where it is associated with a reduction of H3K4me3. In accordance to the ChIP-Seq data, the *Mll1* promoter is indeed bound by Hells in both WT and E2f3-deficient cells (von Eyss et al., 2012, supplementary figure S7D–F). Combined, our data potentially support that the top-ranking ChIP-Seq targets of HELLS, among them *MLL1*, are HELLS targets *in vivo*. Importantly, *Mll1*, *Kras* and *Rnf146* are Hells-regulated targets in both mouse and human cells. Lastly, both E2F3 and HELLS are necessary for proper activation of *MLL1* and other targets.



## 2.7 Data curation for post-transcriptional research

D180–D186 *Nucleic Acids Research*, 2012, Vol. 40, Database issue  
doi:10.1093/nar/gkr1007

Published online 15 November 2011

### doRiNA: a database of RNA interactions in post-transcriptional regulation

Gerd Anders<sup>1</sup>, Sebastian D. Mackowiak<sup>2</sup>, Marvin Jens<sup>2</sup>, Jonas Maaskola<sup>2</sup>,  
Andreas Kuntzagk<sup>1</sup>, Nikolaus Rajewsky<sup>2,\*</sup>, Markus Landthaler<sup>3,\*</sup> and  
Christoph Dieterich<sup>1,\*</sup>

**Abstract** In animals, RBPs and miRNAs post-transcriptionally regulate the expression of virtually all genes by binding to RNA. Recent advances in experimental and computational methods facilitate transcriptome-wide mapping of these interactions. It is thought that the combinatorial action of RBPs and miRNAs on target mRNAs form a post-transcriptional regulatory code. We provide a database that supports the quest for deciphering this regulatory code. Within doRiNA, we are systematically curating, storing and integrating binding site data for RBPs and miRNAs. Users are free to take a target (mRNA) or regulator (RBP and/or miRNA) centric view on the data. We have implemented a database framework with short query response times for complex searches (e.g. asking for all targets of a particular combination of regulators). All search results can be browsed, inspected and analyzed in conjunction with a huge selection of other genome-wide data, because our database is directly linked to a local copy of the UCSC genome browser. At the time of writing, doRiNA encompasses RBP data for the human, mouse and worm genomes. For computational miRNA target site predictions, we provide an update of PicTar predictions.

#### My contributions

I curated the analyses for the PAR-CLIP datasets for the RBPs AGO1–4, IGF2BP1, PUM2, and QKI, produced by Hafner et al. (2010) that were included in the database.

## 2.8 Splicing regulation

Research Article  
RBM10 mediated alternative splicing



### Integrative analysis revealed the molecular mechanism underlying RBM10-mediated splicing regulation

Yongbo Wang<sup>1†</sup>, Andreas Gogol-Döring<sup>1†</sup>, Hao Hu<sup>2</sup>, Sebastian Fröhler<sup>1</sup>, Yunxia Ma<sup>3</sup>, Marvin Jens<sup>4</sup>,  
Jonas Maaskola<sup>4</sup>, Yasuhiro Murakawa<sup>5</sup>, Claudia Quedenau<sup>1</sup>, Markus Landthaler<sup>5</sup>, Vera Kalscheuer<sup>2</sup>,  
Dagmar Wieczorek<sup>6</sup>, Yang Wang<sup>7</sup>, Yuhui Hu<sup>1\*</sup>, Wei Chen<sup>1\*\*</sup>

**Abstract** RBM10 encodes an RNA binding protein. Mutations in RBM10 are known to cause multiple congenital anomaly syndrome in male humans, the TARP syndrome. However, the molecular function of RBM10 is unknown. Here we used PAR-CLIP to identify thousands of binding sites of RBM10 and observed significant RBM10–RNA interactions in the vicinity of splice sites. Computational analyses of binding sites as well as loss-of-function and gain-of-function experiments provided evidence for the function of RBM10 in regulating exon skipping and suggested an underlying mechanistic model, which could be subsequently validated by minigene experiments. Furthermore, we demonstrated the splicing defects in a patient carrying an RBM10 mutation, which could be explained by disrupted function of RBM10 in splicing regulation. Overall, our study established RBM10 as an important regulator of alternative splicing, presented a mechanistic model for RBM10-mediated splicing regulation and provided a molecular link to understanding a human congenital disorder.

### **My contributions**

Using the DMD methods described in this thesis, I analyzed motifs present in RBM10 PAR-CLIP data. Specifically, I compared the PAR-CLIP bound regions with sets of expression stratified control sequences sampled from RNA-Seq data of the same tissue type. I found the exonic splicing enhancer (ESE) motif GAAGAV, originally described by Fairbrother, Yeh, et al. (2002), Lavigne et al. (1993), and Zavolan et al. (2003), to be the most significantly enriched motif present in the data. This finding is consistent with the splicing associated function of RBM10.

## **Part II**

# **Probabilistic Sequence Analysis**



# Overview of probabilistic sequence analysis

Affinity of a protein to a given nucleic acid is determined by unspecific and specific contributions to the binding free energy (von Hippel and Berg, 1986). The unspecific contributions include a large electrostatic component due to ionic interactions between basic amino acid side chains and the acidic sugar phosphate backbone of the nucleic acid. Specific components are caused by hydrogen bonding between nucleobases and protein side chains in the protein's binding pocket. Given a binding pocket, the hydrogen bonding patterns are determined by the sequence identity of the nucleic acids and the ensemble of structural conformations that the nucleic acid assumes. As the structural ensembles of RNA have higher degrees of freedom than those of (typically double stranded) DNA, structural contributions to binding harbor more specificity determining potential for RBPs than for DBPs. However, many RBPs prefer single stranded conformations of binding sites in RNA and, consequently, nucleic acid sequence identity is frequently as important a specificity determinant for RBPs as it is for DBPs. While, apart from sequence and structure, still further properties<sup>8</sup> influence binding to nucleic acids of both DBPs and RBPs, modeling of sequence binding properties remains a fundamental tool to decipher regulatory programs.

The following part discusses modeling techniques for probabilistic sequence analysis. Chapter 3 deals with models for binding motifs. As binding motifs are typically not found in isolation, but rather are embedded into unrelated sequence context, the same chapter also presents models for binding sites. Hidden Markov models (HMMs) are versatile modeling tools that find manifold applications in biological sequence analysis, and their theory is presented in chapter 4. Subsequently, chapter 5 presents a model for binding sites based on HMMs. Learning methods for HMMs are presented in chapter 6.

---

<sup>8</sup>Such as chemical modifications of nucleic acids, or protein and nucleic acid interaction partners of DBPs and RBPs.



## Chapter 3

# Motif and binding site models

Numerous models are available to describe binding preferences of nucleic acid binding proteins (Stormo, 2000). Section 3.1 gives an overview of common models for the binding specificity contribution of core cognate nucleotides. We will refer to these as binding motif models.

In motif discovery (MD) problems involving real biological data motifs rarely present themselves in isolation. Rather, the available sample sequences include functionally unrelated, non-cognate sequence context into which the binding patterns lie embedded.

Section 3.2 addresses composite models of probabilistic character that describe the binding motif embedded in surrounding sequence context. Models of this kind are referred to as binding site models.

### 3.1 Binding motif models

Word-based models, the oldest and simplest motif models considered here, require little explanation and are addressed in section 3.1.1. Models that allow degeneracy at certain positions without being quantitative about preferences at the degenerate positions are discussed in section 3.1.2. Subsequently, section 3.1.3 discusses probabilistic motif models that leverage the toolbox of graphical modeling. Section 3.1.3.1 explains how inference is performed to detect occurrences of probabilistic motif models, and also briefly mentions how probabilistic motif models relate to statistical-mechanical models from biophysics. The section finishes in section 3.1.3.2 with information theoretic concepts applicable to sequence motifs.

#### 3.1.1 Consensus models

The most basic form of binding preference modeling is the specification of the single most frequently bound word, the so-called consensus sequence. An example is given in figure 3.1a. This form of modeling has two advantages. The first is the simplicity of representation, the second the feasibility of finding words that are globally optimal. Obviously the representational simplicity entails that word-based model should only in rare cases allow a comprehensive characterization of all sequences that a factor binds to.

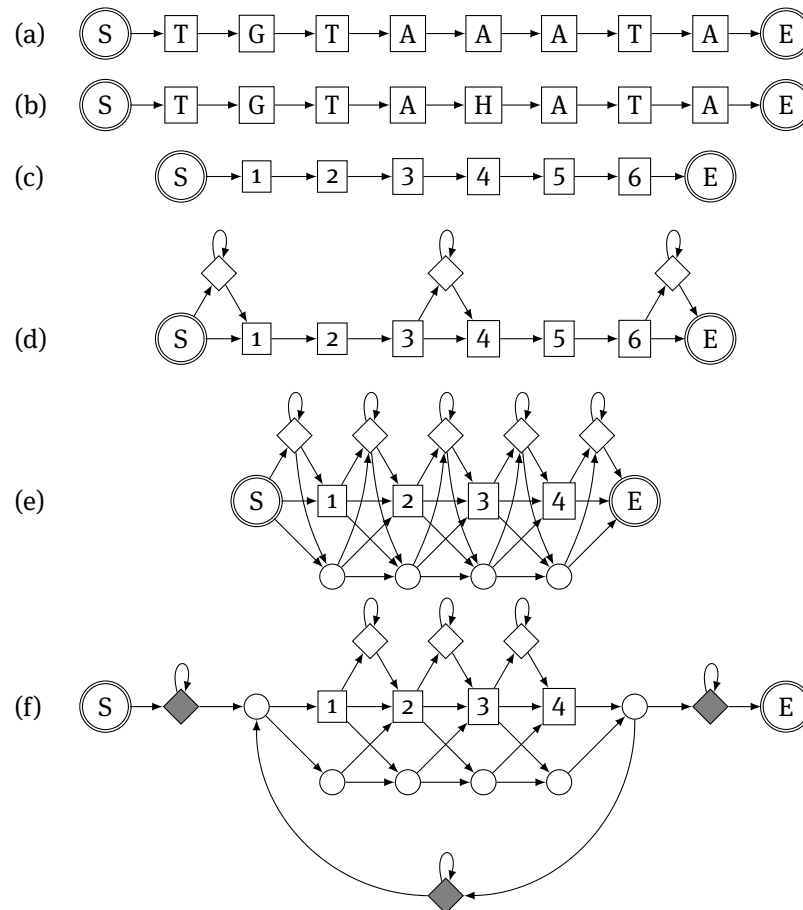


Figure 3.1: Graphical probabilistic models for binding motifs and sites. (a) consensus motif model, (b) IUPAC regex motif model, (c) BLOCKS motif model, (d) META-MEME motif model, (e) Profile HMM motif model, (f) Plan7 HMMER2 binding site model. Start and stop states are labeled S and E, and are marked by double circles. States that are regular parts of the motifs are depicted in rectangle boxes. Insertion states are depicted as diamond shaped nodes. Silent, deletion states are depicted by small circled nodes. In (a) and (b) regular states labeled by a letter emit only characters of that kind. In (c)–(f) numbered regular states and insertion states have individual emission distributions. Figure based on figure 3 of Eddy (1998).



Table 3.1: The IUPAC code for nucleic acids.

Code	Symbols	Mnemonic
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T (or U)	T	Thymine (or Uracil)
R	A or G	Purine
Y	C or T/U	Pyrimidine
S	C or G	Strong
W	A or T/U	Weak
K	G or T/U	Keto
M	A or C	Amino
B	C or G or T/U	Not A
D	A or G or T/U	Not C
H	A or C or T/U	Not G
V	A or C or G	Not T/U
N	A or C or G or T/U	Any

### 3.1.2 Regular expression type models

A simple extension of word-based models is to allow degeneracy at certain positions. We may specify the allowed nucleotides at a given position using e.g. the IUPAC codes for nucleic acids, see table 3.1. An example is given in figure 3.1b. Regular expression (regex) type models are more general than word-based models. Yet they still fail to faithfully represent specific binding preferences in which the presence of a particular nucleotide at a given position may be tolerated but is less conducive to binding than other nucleotides.

### 3.1.3 Probabilistic models

This section summarizes classical, matrix-based probabilistic models for sequence motifs, their usage in probabilistic sequence analysis, and some related information theoretic concepts.

Conceptually and historically the representations presented here build on matrices of nucleotide counts at the different positions of the motif in a set of example sequences. Thus these are discussed first, followed by a probabilistic interpretation of the counts as frequencies. The combination of such a model for a motif of interest with a background model allows to condense the representation into score matrices, that are subsequently discussed.

Matrix-based motif models can be represented by graphical probabilistic models as exemplified in figure 3.1c. Models for motifs of variable lengths due to non-facultative internal positions, such as exemplified in figures 3.1d to 3.1f, are not covered in this section, but are naturally modeled with the tools of HMMs introduced in chapter 4.

**Nucleotide count matrix** The simplest probabilistic model for binding site specificity, a position-specific count matrix (PSCM), is built from a set of example binding site sequences of identical length  $w$ . PSCMs specify the number of times that a given nucleotides

is observed at each position in the example sequences. In a PSCM the parameter at position  $i, j$  specifies how often the symbol  $j$  is observed at the  $i$ -th position of the motif.

**Nucleotide frequency matrix** By transforming the counts of a PSCM at each position into a probability distribution, a position-specific frequency matrix (PSFM) may be derived. In so doing, a pseudo-count is frequently added. In a PSFM the parameter at position  $i, j$  in the matrix gives the probability of observing the symbol  $j$  at the  $i$ -th position of the motif. Formally, in a PSFM  $M = (p_{ij})_{\substack{1 \leq i \leq w \\ 1 \leq j \leq 4}}$  the emission probabilities of each position  $i$  represent a multinomial distribution. Together with an independence assumption of the contributions of the individual positions (Benos, Bulyk, and Stormo, 2002), the columns of a PSFM form a product-multinomial model.

**Score matrix** It is often convenient to directly combine a PSCM with a background model, yielding a position-specific score matrix (PSSM), which is also frequently called position position weight matrix (PWM). It was introduced by Stormo et al. (1982). In a PSSM the observation frequencies of the model are divided by frequencies given by a background model and then log-transformed<sup>1</sup>. Assuming that the background model specifies probabilities ( $b_{ij}$ ) for symbol  $j$  at position  $i$ , the PSSM of the PSFM  $P$  contains scores

$$s_{ij} = \log_2 p_{ij} - \log_2 b_{ij}. \quad (3.1)$$

The advantage of this representation is the numerical convenience it allows in calculating the score  $S$  of a sequence  $\mathbf{X} = (X_i)_{1 \leq i \leq w}$  by simply summing the relevant entries of this table.

$$S(\mathbf{X}|\boldsymbol{\theta}) = \sum_{i=1}^w s_{iX_i} \quad (3.2)$$

Scores of this kind are the log-likelihood ratio of the model to the background, see below.

**Score distribution** There exists an algebraic algorithm due to Staden (1989) to calculate the background score distribution of a given PSSM. As remarked by Rahmann (2003), by applying it to the signal model the same algorithm may be used to compute the signal score distribution. Rahmann also gives a dynamic programming variant of the Staden algorithm.

### 3.1.3.1 Inferring motif occurrences

Where for discrete word-based models the question of motif occurrence is a matter of consistency with the model, for probabilistic motif models it has a probabilistic answer. In situations where probabilistic methods are used it is customarily understood that different kinds of errors may happen, referred to as type I and type II errors. Type I errors are such where a true signal is not recognized as such, in this case where a true motif occur-

<sup>1</sup>Note that we will mostly use logarithms to the base two for the sake of consistence with the conventional usage of units of bits from information theory.

rence is not recognized. Type II errors are such where background is mistaken for signal, i.e. where a background sequence is assumed to be a motif occurrence.

Two schools of thought exist to cope with this, those of statistical hypothesis testing and of Bayesian inference. Statistical hypothesis testing typically aims to determine a suitable threshold  $\lambda$  to accept or reject the hypothesis of a motif occurrence, where Bayesian inference computes posterior probabilities. Both approaches often agree on the outcome, but close to the decision boundary of hypothesis testing Bayesian inference computes intermediate probabilities, interpolating between the decisions and reflecting the uncertainty about which is the right decision.

**Hypothesis testing for motif occurrences** The Neyman-Pearson lemma states that the likelihood ratio test is uniformly most powerful<sup>2</sup> when deciding between two point hypotheses (Neyman and E. Pearson, 1933).

$$\Lambda(X) = \frac{\mathbb{L}(\text{motif}|X)}{\mathbb{L}(\text{background}|X)} = \frac{\mathbb{P}(X|\text{motif})}{\mathbb{P}(X|\text{background})} \geq \lambda, \quad (3.3)$$

where the threshold  $\lambda$  is chosen such that the probability of falsely rejecting the background hypothesis is  $\alpha$ ,  $\mathbb{P}(\Lambda(X) \geq \lambda|H_0) = \alpha$ . It is often numerically advantageous to compute log-likelihoods in place of likelihoods. In this case the test becomes

$$\begin{aligned} \log_2 \Lambda(x) &= \log_2 \mathbb{L}(\text{motif}|X) - \log_2 \mathbb{L}(\text{background}|X) \\ &= \log_2 \mathbb{P}(X|\text{motif}) - \log_2 \mathbb{P}(X|\text{background}) \geq \log_2 \lambda. \end{aligned} \quad (3.4)$$

As remarked at the end of section 3.1.3, the distributions of scores of sequences generated by the signal or background model can be determined. These distributions may then be inspected to determine suitable cutoffs for hypothesis testing. The background score distribution yields the Type-1 error, i.e.  $p$ -value. The signal score distribution yields the statistical power of the test, i.e. the probability to detect true signal.

**Bayesian reasoning for motif occurrences** Within the framework of Bayesian reasoning one may consider a related ratio of probabilities, namely that of joint posterior probabilities,

$$R = \frac{\mathbb{P}(X, \text{motif})}{\mathbb{P}(X, \text{background})} = \frac{\mathbb{P}(X|\text{motif})\mathbb{P}(\text{motif})}{\mathbb{P}(X|\text{background})\mathbb{P}(\text{background})}, \quad (3.5)$$

or, again because of numeric advantages, the difference of logarithmic probabilities,

$$\log_2 R = \log_2 \mathbb{P}(X|\text{motif}) - \log_2 \mathbb{P}(X|\text{background}) + \log_2 \mathbb{P}(\text{motif}) - \log_2 \mathbb{P}(\text{background}). \quad (3.6)$$

It is clearly seen that the Bayesian log ratio of posterior probabilities is identical to the log likelihood ratio statistic plus the log ratio of prior probabilities. In other words, the threshold motivated by the trade-off between Type I and II errors in hypothesis testing

<sup>2</sup>A test is uniformly most powerful when it has the greatest power  $1 - \beta$  among all tests of a given false positive rate  $\alpha$ . Statistical power of a test is also known as sensitivity, and  $\beta$  is the probability of committing a type II error. False positive rate is the probability of committing a type I error, i.e. of rejecting the null hypothesis when it is true.

corresponds in Bayesian reasoning to a constant that reflects the prior knowledge of the frequency of the two classes.

Apart from the log ratio of joint probabilities  $R$  which is practically equivalent to hypothesis testing, the Bayesian formalism offers further possibilities, such as calculating the posterior probability for a motif occurrence,

$$\begin{aligned} \mathbb{P}(\text{motif}|X) &= \frac{\mathbb{P}(X, \text{motif})}{\mathbb{P}(X, \text{motif}) + \mathbb{P}(X, \text{background})} \\ &= \frac{\mathbb{P}(X|\text{motif})\mathbb{P}(\text{motif})}{\mathbb{P}(X|\text{motif})\mathbb{P}(\text{motif}) + \mathbb{P}(X|\text{motif})\mathbb{P}(\text{motif})} \end{aligned} \quad (3.7)$$

It may be noted in passing that the posterior probability is equivalent to the sigmoid function  $\sigma$  applied to the above difference of logarithmic joint probabilities<sup>3</sup>,

$$\begin{aligned} \sigma(\log_2 R) &= \frac{1}{1 + 2^{-\log_2 R}} = \frac{1}{1 + \frac{1}{R}} \\ &= \frac{1}{1 + \frac{\mathbb{P}(X, \text{background})}{\mathbb{P}(X, \text{motif})}} = \frac{\mathbb{P}(X, \text{motif})}{\mathbb{P}(X, \text{motif}) + \mathbb{P}(X, \text{background})} \\ &= \mathbb{P}(\text{motif}|X). \end{aligned} \quad (3.8)$$

This equivalence of posterior motif probabilities and the sigmoid function applied to  $\log_2 R$  establishes a biophysical underpinning for the probabilistic formalism. In statistical mechanics Fermi-Dirac statistics describe the state distribution of systems of many particles obeying the Pauli exclusion principle. In particular, the sigmoid is equivalent to the Fermi-Dirac statistics of a two state system like that of a site that may be occupied by a bound factor or may be free for binding. This equivalence relates the log motif likelihood ratio to the binding energy and the log ratio of prior probabilities (or equivalently the hypothesis testing threshold) to the chemical potential, scaled by the product of temperature and gas constant.

### 3.1.3.2 Information theoretic concepts

**Information entropy of a motif** This and the following subsections use some tools from information theory which are summarized in the appendix C.2. The information entropy (see appendix C.2.1) of a PSFM  $M = (p_{ij})_{\substack{1 \leq i \leq w \\ 1 \leq j \leq 4}}$  at a position  $i$  is given by

$$\mathbb{H}_i(M) = - \sum_{j=1}^4 p_{ij} \log_2 p_{ij}. \quad (3.9)$$

When the base of the logarithm in the above equation is two, the unit of information entropy is bit. The minimum of 0 bit is attained for a degenerate distribution, e.g. for  $p_i = (1, 0, 0, 0)$ , the maximum of 2 bit for a uniform distribution, i.e. for  $p_i = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ .

Assuming independence of the contribution of the individual positions to the binding (Benos, Bulyk, and Stormo, 2002), the information entropy of a PSFM  $M$  of width  $w$  is

<sup>3</sup>Note that the sigmoid function is typically defined as  $\sigma(x) = \frac{1}{1+e^{-x}}$ , but here we use a variant defined as  $\sigma(x) = \frac{1}{1+2^{-x}}$ . Also note the first footnote of this chapter.

given by

$$\mathbb{H}(M) = \sum_{i=1}^w \mathbb{H}_i(M) = - \sum_{i=1}^w \sum_{j=1}^4 p_{ij} \log_2 p_{ij}. \quad (3.10)$$

The minimal value of information entropy of a motif of width  $w$  is 0 bit, its maximum is  $2w$  bit.

**Information content of a motif** Schneider, Stormo, et al. (1986) introduced the concept of information content (IC) of a PSFM. The IC of a PSFM  $M$  of width  $w$  may be determined relative to a background model  $B$ ,

$$\text{IC}(M, B) = \mathbb{I}(M; B) \quad (3.11)$$

$$= \sum_{i=1}^w \sum_{j=1}^4 p_{ij} \log_2 \frac{p_{ij}}{b_{ij}} \quad (3.12)$$

$$= \sum_{i=1}^w \sum_{j=1}^4 p_{ij} \log_2 p_{ij} - \sum_{i=1}^w \sum_{j=1}^4 p_{ij} \log_2 b_{ij} \quad (3.13)$$

$$= - \sum_{i=1}^w \sum_{j=1}^4 p_{ij} \log_2 b_{ij} - \mathbb{H}(M). \quad (3.14)$$

It is the Kullback-Leibler (KL) divergence (see appendix C.2.4) of the model and the background model,  $\text{IC}(M, B) = D_{KL}(M||B)$ . If the background model  $B$  is a uniform, zeroth order Markov chain, i.e.  $b_{ij} = \frac{1}{4}$  for  $1 \leq i \leq w$  and  $1 \leq j \leq 4$ , then the IC of a length- $w$  motif  $M$  is

$$\text{IC}(M) = - \sum_{i=1}^w \sum_{j=1}^4 p_{ij} \log_2 \frac{1}{4} - \mathbb{H}(M) \quad (3.15)$$

$$= 2w - \mathbb{H}(M). \quad (3.16)$$

In the case of a uniform background model, the maximum IC is  $2w$  bit, which is attained when all positions are degenerate, and the minimum information is 0 bit when all positions are uniform.

The IC of a motif may be interpreted as the reduction of information entropy of a motif relative to the background.

Another interpretation of the IC of a PSFM  $M$  is the expected PSSM score of sequences  $\mathbf{X}$  generated by the underlying PSFM  $M$ ,

$$\text{IC}(M) = \mathbb{E}[S(\mathbf{X}|\boldsymbol{\theta})|\mathbf{X} \sim M]. \quad (3.17)$$

Note again our previous remark in section 3.1.3 that the distribution of PSSM scores can be computed efficiently, and by calculating its mean we can experimentally confirm this theoretical expectation.

**Sequence logo** Sequence logos are a way of representing the binding preferences of a nucleic acid binding factor that was introduced by Schneider and Stephens (1990).

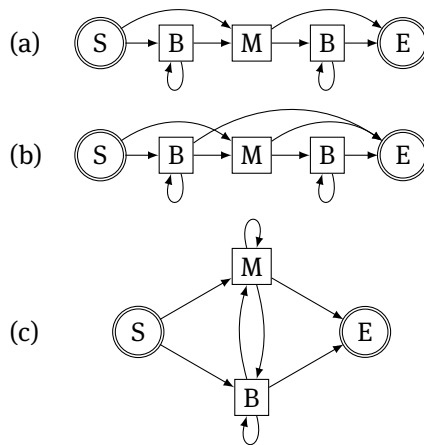


Figure 3.2: Simple binding site model topologies. (a) One occurrence per sequence (OOPS) model. (b) Zero or one occurrence per sequence (ZOOPS) model. (c) Arbitrary number of occurrences (ANOPS) model. Double circles denote start (S) and end (E) states. Rectangles denote motif (M) and background (B) states.

The idea is to scale the total vertical extent of each position by the IC of that position. The letters at this position divide this total height up proportionately to their probability.

Sequence logos will be used to present the outcome of MD applications later in this thesis, such as in chapters 18 and 20.

## 3.2 Binding site models

### 3.2.1 Binding site model topologies

Figure 3.2 shows graphical representations of basic binding site models. The one occurrence per sequence (OOPS) model is for sequences with exactly one occurrence of the binding motif, possibly embedded in background sequence. The zero or one occurrence per sequence (ZOOPS) model represents sequences with either exactly one or no binding motif, optionally surrounded by background sequence. The arbitrary number of occurrences per sequence (ANOPS) model is used to model sequences that may have zero, one or multiple motif occurrences, with optional surrounding background sequence.

### 3.2.2 Likelihoods

Expressions for the probabilistic models corresponding to the OOPS and ZOOPS models are as follows. Assuming the sequence  $X$  to be of length  $n$ , and the motif to be of width  $w$ ,

the likelihood for the OOPS model is given by

$$\begin{aligned}
\mathbb{L}(\text{OOPS}, \boldsymbol{\theta}|\mathbf{X}) &= \mathbb{P}(\mathbf{X}|\text{OOPS}, \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n-w+1} \mathbb{P}(\mathbf{M}_i|\boldsymbol{\theta}) \mathbb{P}(\mathbf{X}_{1\dots i-1}|\mathbf{B}, \boldsymbol{\theta}) \mathbb{P}(\mathbf{X}_{i\dots i+w-1}|\mathbf{M}, \boldsymbol{\theta}) \mathbb{P}(\mathbf{X}_{i+w\dots n}|\mathbf{B}, \boldsymbol{\theta}) \\
&= \mathbb{P}(\mathbf{X}|\mathbf{B}, \boldsymbol{\theta}) \sum_{i=1}^{n-w+1} \mathbb{P}(\mathbf{M}_i|\boldsymbol{\theta}) \frac{\mathbb{P}(\mathbf{X}_{i\dots i+w-1}|\mathbf{M}, \boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}_{i\dots i+w-1}|\mathbf{B}, \boldsymbol{\theta})},
\end{aligned} \tag{3.18}$$

where  $\mathbb{P}(\mathbf{M}_i|\boldsymbol{\theta})$  is a positional prior, that is often chosen to be uniform,  $\mathbb{P}(\mathbf{M}_i) = \frac{1}{n-w+1}$ . By making use of PSSM scores, as defined in equation (3.2), this can be written as

$$\mathbb{L}(\text{OOPS}, \boldsymbol{\theta}|\mathbf{X}) = \mathbb{P}(\mathbf{X}|\mathbf{B}, \boldsymbol{\theta}) \sum_{i=1}^{n-w+1} \mathbb{P}(\mathbf{M}_i|\boldsymbol{\theta}) 2^{S(\mathbf{X}_{i\dots i+w-1}|\boldsymbol{\theta})}. \tag{3.19}$$

The likelihood of the ZOOPS model is given by a mixture model on top of the OOPS model,

$$\mathbb{L}(\text{ZOOPS}, \boldsymbol{\theta}|\mathbf{X}) = \mathbb{P}(\mathbf{X}|\text{ZOOPS}, \boldsymbol{\theta}) = \lambda \cdot \mathbb{P}(\mathbf{X}|\text{OOPS}, \boldsymbol{\theta}) + (1 - \lambda) \cdot \mathbb{P}(\mathbf{X}|\mathbf{B}, \boldsymbol{\theta}), \tag{3.20}$$

where  $\lambda$  is the probability of observing a binding motif in the site.

The ANOPS model is conveniently formulated using the theory of HMMs. After outlining HMM theory in chapter 4, we return to the ANOPS binding site model in chapter 5.

### 3.2.3 Posterior probabilities

Based on the above given expressions we can define expressions for different posterior probabilities under the OOPS and ZOOPS model.

Let us label the hypothesis of observing a motif occurrence at position  $i$  as  $\mathbf{M}_i$ , and, overloading notation, label the hypothesis of observing at least one motif occurrence anywhere in a sequence as  $\mathbf{M}$ . By definition, the posterior probability of observing a motif occurrence anywhere in the sequence is 1 for the OOPS model,  $\mathbb{P}(\mathbf{M}|\mathbf{X}, \text{OOPS}, \boldsymbol{\theta}) = 1$ .

The posterior probability of  $\mathbf{M}_i$  under the OOPS model is

$$\begin{aligned}
\mathbb{P}(\mathbf{M}_i|\mathbf{X}, \text{OOPS}, \boldsymbol{\theta}) &= \frac{\mathbb{P}(\mathbf{M}_i, \mathbf{X}|\text{OOPS}, \boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\text{OOPS}, \boldsymbol{\theta})} \\
&= \frac{\mathbb{P}(\mathbf{X}|\mathbf{B}, \boldsymbol{\theta}) \mathbb{P}(\mathbf{M}_i|\boldsymbol{\theta}) \mathbb{P}(\mathbf{X}_{i\dots i+w-1}|\mathbf{M}, \boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\text{OOPS}, \boldsymbol{\theta}) \mathbb{P}(\mathbf{X}_{i\dots i+w-1}|\mathbf{B}, \boldsymbol{\theta})} \\
&= \frac{\mathbb{P}(\mathbf{X}|\mathbf{B}, \boldsymbol{\theta}) \mathbb{P}(\mathbf{M}_i|\boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\text{OOPS}, \boldsymbol{\theta})} 2^{S(\mathbf{X}_{i\dots i+w-1}|\boldsymbol{\theta})}.
\end{aligned} \tag{3.21}$$

The corresponding posterior probability for  $M_i$  under the ZOOPS model is given by

$$\begin{aligned}
 \mathbb{P}(M_i|\mathbf{X}, \text{ZOOPS}, \boldsymbol{\theta}) &= \frac{\mathbb{P}(M_i, \mathbf{X}|\text{ZOOPS}, \boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\text{ZOOPS}, \boldsymbol{\theta})} \\
 &= \frac{\lambda \mathbb{P}(\mathbf{X}|\mathbf{B}, \boldsymbol{\theta}) \mathbb{P}(M_i|\boldsymbol{\theta}) \mathbb{P}(X_{i\dots i+w-1}|\mathbf{M}, \boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\text{ZOOPS}, \boldsymbol{\theta}) \mathbb{P}(X_{i\dots i+w-1}|\mathbf{B}, \boldsymbol{\theta})} \\
 &= \frac{\lambda \mathbb{P}(\mathbf{X}|\mathbf{B}, \boldsymbol{\theta}) \mathbb{P}(M_i|\boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\text{ZOOPS}, \boldsymbol{\theta})} 2^{S(X_{i\dots i+w-1}|\boldsymbol{\theta})}.
 \end{aligned} \tag{3.22}$$

The posterior probability of  $M$  under the ZOOPS models is given by

$$\mathbb{P}(M|\mathbf{X}, \text{ZOOPS}, \boldsymbol{\theta}) = \sum_{i=1}^{n-w+1} \mathbb{P}(M_i|\mathbf{X}, \text{ZOOPS}, \boldsymbol{\theta}). \tag{3.23}$$

We will refer to the hypothesis of not observing a motif in the ZOOPS model as  $\neg M$  by using the mathematical negation notation for the logical complement. The posterior for  $\neg M$  is given by

$$\begin{aligned}
 \mathbb{P}(\neg M|\mathbf{X}, \text{ZOOPS}, \boldsymbol{\theta}) &= \frac{\mathbb{P}(\neg M, \mathbf{X}|\text{ZOOPS}, \boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\text{ZOOPS}, \boldsymbol{\theta})} \\
 &= \frac{(1 - \lambda) \mathbb{P}(\mathbf{X}|\mathbf{B})}{\mathbb{P}(\mathbf{X}|\text{ZOOPS}, \boldsymbol{\theta})}.
 \end{aligned} \tag{3.24}$$

As  $M$  and its negation  $\neg M$  are complementary events, this yields a second expression for the posterior probability of  $M$ ,

$$\mathbb{P}(M|\mathbf{X}, \text{ZOOPS}, \boldsymbol{\theta}) = 1 - \mathbb{P}(\neg M|\mathbf{X}, \text{ZOOPS}, \boldsymbol{\theta}). \tag{3.25}$$

Based on the theory HMMs, the posterior probability for the ANOPS model is given later in section 5.2.



## Chapter 4

# Hidden Markov models

Hidden Markov models (HMMs) are probabilistic graphical models. They are used in situations that may be described by a set of variables of which only a subset is observed, in so-called partial data modeling.

Hidden Markov modeling involves two spaces, one that represents the observable entities, and another one that represents underlying states. Both state and observation spaces may be continuous or discrete. In this thesis we will only consider discrete state and observation spaces.

Relevant literature on HMMs includes a famous review by Rabiner (1989), as well as some textbooks. Rabiner and Juang (1993) detail speech recognition applications. Durbin et al. (1998) focus on biological sequence applications. Cappé, Moulines, and Rydén (2010) cover more theoretical aspects. Koller and N. Friedman (2009) address the general theory of probabilistic graphical models. Finally, general machine learning textbooks by MacKay (2003) and Bishop (2006) also outline the theory of HMMs.

### 4.1 Formal definition

The notation and definitions here follow Baum, Petrie, et al. (1970) and Rabiner (1989).

**Definition 1** (Hidden Markov model). *Let  $A = (a_{ij})$  be an  $N \times N$  stochastic matrix, i.e.  $\sum_{j=1}^N a_{ij} = 1$  for all  $1 \leq i \leq N$ , and  $a_{ij} \geq 0$  for all  $1 \leq i, j \leq N$ . Let  $a = (a_i), 1 \leq i \leq N$  be a probability distribution, i.e.  $\sum_{i=1}^N a_i = 1$ , and  $a_i \geq 0$  for all  $1 \leq i \leq N$ . For each  $1 \leq i \leq N$  let  $b_i(y)$  be a probability density:  $\int b_i(y) dy = 1$ . Let  $\theta$  be the triple  $A, a, b = \{b_i\}$ . We define a stochastic process  $\mathbf{X} = \{X_t\}$  with density*

$$\begin{aligned} \mathbb{L}(\theta|\mathbf{X}) &= \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_T = x_T | \theta) \\ &= \sum_{i_0, i_1, \dots, i_T=1}^N a_{i_0} a_{i_0 i_1} b_{i_1}(x_1) a_{i_1 i_2} b_{i_2}(x_2) \cdots a_{i_{T-1} i_T} b_{i_T}(x_T). \end{aligned} \quad (4.1)$$

Then  $\theta$  is a hidden Markov model for  $\{X_t\}$ .

### 4.1.1 Modeling beginning and end of sequences

It is possible to simplify the structure of HMMs by adding a special state in which the model is initialized. Below, we assume that the index of this state is 1. Thus,  $a_1 = 1$  and  $a_i = 0$  for  $1 < i \leq N$ . The role of the initial state distribution parameter  $a = (a_i)$ ,  $1 \leq i \leq N$  is thus realized by the transition probabilities  $(a_{ij})$ ,  $1 \leq j \leq N$ .

**Definition 2** (Hidden Markov model with start state). *Let  $A = (a_{ij})$  be an  $N \times N$  stochastic matrix. For each  $1 \leq i \leq N$  let  $b_i(y)$  be a probability density. Let  $\theta$  be the pair  $A, b = \{b_i\}$ . We define a stochastic process  $\mathbf{X} = \{X_t\}$  with density*

$$\begin{aligned} \mathbb{L}(\theta|\mathbf{X}) &= \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_T = x_T | \theta) \\ &= \sum_{i_1, \dots, i_T=1}^N a_{1i_1} b_{i_1}(x_1) a_{i_1 i_2} b_{i_2}(x_2) \cdots a_{i_{T-1} i_T} b_{i_T}(x_T). \end{aligned} \quad (4.2)$$

Then  $\theta$  is a hidden Markov model with start state for  $\{X_t\}$ .

Similarly, it is possible to model the end of sequences. For this one may introduce another special state as end state. It is however also possible to identify start and end states. When doing so, we again assume that the index of the start and end state is 1.

**Definition 3** (Hidden Markov model with start and end state). *Let  $A = (a_{ij})$  be an  $N \times N$  stochastic matrix. For each  $1 \leq i \leq N$  let  $b_i(y)$  be a probability density. Let  $\theta$  be the pair  $A, b = \{b_i\}$ . We define a stochastic process  $\mathbf{X} = \{X_t\}$  with density*

$$\begin{aligned} \mathbb{L}(\theta|\mathbf{X}) &= \mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_T = x_T | \theta) \\ &= \sum_{i_1, \dots, i_T=1}^N a_{i_1 i_1} b_{i_1}(x_1) a_{i_1 i_2} b_{i_2}(x_2) \cdots a_{i_{T-1} i_T} b_{i_T}(x_T) a_{i_T 1}. \end{aligned} \quad (4.3)$$

Then  $\theta$  is a hidden Markov model with start and end state for  $\{X_t\}$ .

The start and end state is the only state to emit a specific symbol,  $\varepsilon$ , the so called empty symbol that is never occurring in the observations and which is not emitted by any other state. It is assumed that the observation sequence is pre- and postfixed by the empty symbol  $\varepsilon$ , thus forcing any valid state path to begin and end in that state.

## 4.2 Fundamental problems and basic inference algorithms

Rabiner (1989) lists three fundamental problems in HMM applications: decoding, evaluation, and learning.

The decoding problem seeks to find the most likely state path that explains an observation given a model. The evaluation problem is how to compute the likelihood with which a given observation sequence is generated by a model. The learning problem is that of determining the parameters of a model that are most suitable to describing a set of observations.

The solution to the decoding problem is often found with the Viterbi algorithm, which is described in section 4.3. The evaluation problem is solved by the forward-backward

algorithm, outlined in section 4.4. Learning methods for HMMs deserve a more extensive discussion. Chapter 6 is dedicated to outlining some approaches for this purpose. Later chapters of this dissertation outline new discriminative learning approaches.

The following two sections of this chapter detail the two fundamental methods of inference for HMMs: the Viterbi algorithm in section 4.3 and the forward-backward algorithm in section 4.4.

As detailed below, both algorithms are of the same order of complexity: for an HMM with  $E$  transitions of non-zero probability, and a sequence of length  $T$  the complexity is in both cases  $\mathcal{O}(TE)$  and thus linear in the size of the data.

Both algorithms run into numerical issues when implemented in a straight-forward manner. However, numerically stable variants are available, and detailed below. For the Viterbi-algorithm we explain a log-space calculation that avoids these problems. For the forward-backward algorithm the numerical issues may be avoided by a scaling method.

Note that the algorithms as given below assume that the HMMs have a start and end state. Observation sequences mentioned below are thus assumed to be pre- and postfixed by the empty symbol  $\varepsilon$ . For example if an observation sequence  $\mathbf{X} = X_1 X_2 \dots X_T$  of length  $T$  is given, it is treated as  $\varepsilon X_1 X_2 \dots X_T \varepsilon$ , i.e. with  $X_0$  and  $X_{T+1}$  equal to  $\varepsilon$ .

### 4.3 Viterbi algorithm

The problem of decoding amounts to determining the state sequence  $\mathbf{q}^*$  that is individually most likely to generate a given observation sequence  $\mathbf{X} = X_1 X_2 \dots X_T$  with a model  $\theta$ ,

$$\mathbf{q}^* = \underset{\mathbf{q}}{\operatorname{argmax}} \mathbb{P}(\mathbf{q}, \mathbf{X} | \theta). \quad (4.4)$$

The Viterbi algorithm is a recursive scheme to determine  $v_t(k)$ , the probability of the most probable path ending in state  $k$  with observation  $X_t$ . During the computation a pointer structure  $ptr$  is created that allows the subsequent determination of the corresponding path via backtracking.

1. Initialization ( $t = 0$ ):

$$v_0(j) = \begin{cases} 1 & \text{for } j = 1 \\ 0 & \text{for } 1 < j \leq N \end{cases} \quad (4.5)$$

2. Recursion ( $t = 1, \dots, T$ ):

$$v_t(j) = b_j(X_t) \max_i (v_{t-1}(i) a_{ij}) \quad \text{for } 1 \leq j \leq N \quad (4.6)$$

$$ptr_t(j) = \underset{i}{\operatorname{argmax}} (v_{t-1}(i) a_{ij}) \quad \text{for } 1 \leq j \leq N \quad (4.7)$$

3. Termination:

$$\mathbb{P}(\mathbf{q}^*, \mathbf{X} | \theta) = \max_i (v_T(i) a_{i1}) \quad (4.8)$$

$$q_T^* = \underset{i}{\operatorname{argmax}} (v_T(i) a_{i1}) \quad (4.9)$$

4. Traceback ( $t = T, \dots, 1$ ):

$$q_{t-1}^* = ptr_t(q_t^*) \quad (4.10)$$

**Complexity** The complexity of the recursion is demonstrated by noting that for each time point  $1 \leq t \leq T$  and each state  $1 \leq j \leq N$  all previous states  $1 \leq i \leq N$  need to be considered, resulting in  $\mathcal{O}(TN^2)$ . The initialization is of complexity  $\mathcal{O}(N)$ , and the trace back phase is of complexity  $\mathcal{O}(T)$ . Thus, the total runtime complexity is  $\mathcal{O}(TN^2)$ .

Considering the fact that most HMM topologies are far from complete graphs, it is possible to implement this algorithm in  $\mathcal{O}(TE)$ , where  $E$  is the number of transitions with non-zero probability.  $E \geq N - 1$  for HMMs that have the topology of a connected graph, and  $E \leq N^2$  for general graphs. This speed-up can be achieved by considering only the topological predecessors of a given state in performing the maximizations during the recursive steps. The topological predecessors of a given state are the states that have a non-zero transition probability to the state. Appendix B illustrates this.

### 4.3.1 Numerical aspects - log-space computation

Numerical problems are encountered when calculating the Viterbi path for long observation sequences. The likelihood is the product of probabilities, i.e. numbers less or equal to one, and any product of sufficiently many probabilities will tend towards zero. Thus simple implementations will lead to numerical underflow.

These numerical issues of the Viterbi algorithm can be easily avoided by resorting to log-space probability calculations. The algorithm then determines  $\log v_t(k)$ , the logarithm of the probability of the most probable path ending in state  $k$  with observation  $X_t$ . The trace-back phase is identical to the previous version.

(4.11)

1. Initialization ( $t = 0$ ):

$$\log v_0(j) = \begin{cases} 0 & \text{for } j = 1 \\ -\infty & \text{for } 1 < j \leq N \end{cases} \quad (4.12)$$

2. Recursion ( $t = 1, \dots, T$ ):

$$\log v_t(j) = \log b_j(X_t) + \max_i (\log v_{t-1}(i) + \log a_{ij}) \quad \text{for } 1 \leq j \leq N \quad (4.13)$$

$$ptr_t(j) = \operatorname{argmax}_i (\log v_{t-1}(i) + \log a_{ij}) \quad \text{for } 1 \leq j \leq N \quad (4.14)$$

3. Termination:

$$\log P(\mathbf{q}^*, \mathbf{X} | \boldsymbol{\theta}) = \max_i (\log v_T(i) + \log a_{i1}) \quad (4.15)$$

$$q_T^* = \operatorname{argmax}_i (\log v_T(i) + \log a_{i1}) \quad (4.16)$$

4. Traceback ( $t = T, \dots, 1$ ):

$$q_{t-1}^* = ptr_t(q_t^*) \quad (4.17)$$

To avoid redundant calculation, the logarithms of the transition and emission probabilities may be precomputed.

## 4.4 Forward-backward algorithm

The forward-backward algorithm consists in computing two matrices  $\alpha$  and  $\beta$  of likelihoods of partial observations, such that

$$\alpha_t(i) := \mathbb{P}(X_1 X_2 \dots X_t, q_t = S_i | \theta) \quad (4.18)$$

is the conditional joint probability of the partial observation sequence  $X_1 X_2 \dots X_t$  and being in state  $S_i$  at time  $t$  given the model  $\theta$ . The matrix  $\beta$  is defined by

$$\beta_t(i) := \mathbb{P}(X_{t+1} X_{t+2} \dots X_T | q_t = S_i, \theta) \quad (4.19)$$

and gives the conditional probability of the partial observation sequence  $X_{t+1} X_{t+2} \dots X_T$  given the model  $\theta$  and given that the state at time  $t$  is  $S_i$ .

$\alpha$  and  $\beta$  can be computed efficiently with algorithms that we give next. They can be used to compute probabilistic values of interest. Among these are the likelihood, the posterior probability of being in a given state at a given time, and expected values of these posterior probabilities. Also the likelihood gradient can be computed efficiently from the forward and backward matrices, as we will show at the end of this chapter.

### 4.4.1 Algorithm to compute the forward matrix

The following algorithm computes the forward matrix.

1. Initialization ( $t = 0$ ):

$$\alpha_0(j) = \begin{cases} 1 & \text{for } j = 1 \\ 0 & \text{for } 1 < j \leq N \end{cases} \quad (4.20)$$

2. Recursion ( $t = 1, \dots, T + 1$ ):

$$\alpha_t(j) = b_j(X_t) \sum_{i=1}^N \alpha_{t-1}(i) a_{ij} \quad \text{for } 1 \leq j \leq N \quad (4.21)$$

Regarding the initialization step in the algorithm, note that we use an HMM with start and end state, assuming that it is initialized in state  $q_0 = S_1$ .

### 4.4.2 Algorithm to compute the backward matrix

The following algorithm computes the backward matrix.

1. Initialization ( $t = T + 1$ ):

$$\beta_{T+1}(j) = \begin{cases} 1 & \text{for } j = 1 \\ 0 & \text{for } 1 < j \leq N \end{cases} \quad (4.22)$$

2. Recursion ( $t = T, \dots, 0$ ):

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(X_{t+1}) \beta_{t+1}(j) \quad \text{for } 1 \leq i \leq N \quad (4.23)$$

Regarding the initialization step in the algorithm, note again that we use an HMM with start and end state, assuming that it has to end in state  $q_{T+1} = S_1$ .

**Complexity** For both the forward and the backward algorithm the complexity is again demonstrated by noting that for each time point  $1 \leq t \leq T$  and each state  $1 \leq j \leq N$  all possible predecessor or successor states  $1 \leq i \leq N$  need to be considered, resulting in  $\mathcal{O}(TN^2)$ .

Like the Viterbi algorithm, also the forward and backward algorithms can be computed in  $\mathcal{O}(TE)$ , where  $E$  are the number of edges in the HMM topology, by respecting the graph topology of the HMM (see section 4.3 and appendix B).

### 4.4.3 Inference

**Computing posterior probability of being in some state** Using the forward and backward matrices we find that the joint probability of being in state  $S_i$  at time  $t$  and an observation sequence  $\mathbf{X} = X_1 X_2 \dots X_T$  given the model  $\boldsymbol{\theta}$  is given by

$$\begin{aligned} \mathbb{P}(\mathbf{X}, q_t = S_i | \boldsymbol{\theta}) &= \mathbb{P}(X_1 X_2 \dots X_t, q_t = S_i | \boldsymbol{\theta}) \mathbb{P}(X_{t+1} X_{t+2} \dots X_T | q_t = S_i, \boldsymbol{\theta}) \\ &= \alpha_t(i) \beta_t(i). \end{aligned} \quad (4.24)$$

Using Bayes' theorem, we see that the posterior probability of being in state  $S_i$  at time  $t$  given an observation sequence  $\mathbf{X} = X_1 X_2 \dots X_T$  is then given by

$$\mathbb{P}(q_t = S_i | \mathbf{X}, \boldsymbol{\theta}) = \frac{\mathbb{P}(\mathbf{X}, q_t = S_i | \boldsymbol{\theta})}{\mathbb{P}(\mathbf{X} | \boldsymbol{\theta})} = \frac{\alpha_t(i) \beta_t(i)}{\mathbb{P}(\mathbf{X} | \boldsymbol{\theta})}. \quad (4.25)$$

Similarly, if we want to compute the posterior probability of a transition from state  $S_i$  to state  $S_j$  at time  $t$  we may use

$$\mathbb{P}(q_t = S_i, q_{t+1} = S_j | \mathbf{X}, \boldsymbol{\theta}) = \frac{\mathbb{P}(\mathbf{X}, q_t = S_i, q_{t+1} = S_j | \boldsymbol{\theta})}{\mathbb{P}(\mathbf{X} | \boldsymbol{\theta})} = \frac{\alpha_t(i) a_{ij} b_j(X_{t+1}) \beta_{t+1}(j)}{\mathbb{P}(\mathbf{X} | \boldsymbol{\theta})}. \quad (4.26)$$

**Expected transitions and emissions** By summing (4.26) over all times  $t$ , we can compute for an observation sequence  $\mathbf{X}$  the expected number of transitions  $\mathcal{A}_{ij}$  from state  $S_i$  to state  $S_j$ ,

$$\mathcal{A}_{ij} = \sum_{t=0}^T \mathbb{P}(q_t = S_i, q_{t+1} = S_j | \mathbf{X}, \boldsymbol{\theta}) = \sum_{t=0}^T \frac{\alpha_t(i) a_{ij} b_j(X_{t+1}) \beta_{t+1}(j)}{\mathbb{P}(\mathbf{X} | \boldsymbol{\theta})}. \quad (4.27)$$

The expected number of times that observation  $x$  occurs in state  $S_i$  in observation sequence  $\mathbf{X}$  is given by summing (4.25) over those times at which  $x$  is observed,

$$\mathcal{B}_i(x) = \sum_{\{t | X_t = x\}} \mathbb{P}(q_t = S_i | \mathbf{X}, \boldsymbol{\theta}) = \sum_{\{t | X_t = x\}} \frac{\alpha_t(i) \beta_t(i)}{\mathbb{P}(\mathbf{X} | \boldsymbol{\theta})}. \quad (4.28)$$

**Computing the likelihood** By marginalizing (4.24) over all states at any time  $t$  we can compute the likelihood of an observation sequence  $\mathbf{X} = X_1 X_2 \dots X_T$  given the model  $\theta$ ,

$$\mathbb{L}(\theta|\mathbf{X}) = \mathbb{P}(\mathbf{X}|\theta) = \sum_{i=1}^N \mathbb{P}(\mathbf{X}, q_t = S_i|\theta) = \sum_{i=1}^N \alpha_t(i)\beta_t(i). \quad (4.29)$$

For HMMs with start and end state  $S_1$  another way of determining the likelihood from the forward matrix alone is

$$\mathbb{L}(\theta|\mathbf{X}) = \mathbb{P}(\mathbf{X}|\theta) = \mathbb{P}(\mathbf{X}, q_{T+1} = S_1|\theta) = \alpha_{T+1}(1), \quad (4.30)$$

and similarly, using only the backward matrix:

$$\mathbb{L}(\theta|\mathbf{X}) = \mathbb{P}(\mathbf{X}|\theta) = \mathbb{P}(\mathbf{X}, q_0 = S_1|\theta) = \beta_0(1). \quad (4.31)$$

## 4.5 Scaled forward-backward algorithm

With increasing length of the observation the numbers in the matrices  $\alpha$  and  $\beta$  quickly become smaller than what can be represented by floating point types. To alleviate this problem a scaling method can be used, which is detailed below.

The scaled forward-backward algorithm determines matrices  $\tilde{\alpha}$  and  $\tilde{\beta}$ , as well as a scaling vector  $\mathbf{s}$  such that

$$\alpha_t(i) = \tilde{\alpha}_t(i) \prod_{k=0}^t s_k = \tilde{\alpha}_t(i) \prod_{k=1}^t s_k, \quad (4.32)$$

and

$$\beta_t(i) = \tilde{\beta}_t(i) \prod_{k=t}^{T+1} s_k. \quad (4.33)$$

In equation (4.32) the latter identity is due to  $s_0 = 1$ .

### 4.5.1 Algorithm to compute the scaled forward matrix

The following algorithm computes the scaled forward matrix  $\tilde{\alpha}$  and the scaling vector  $\mathbf{s}$ .

1. Initialization ( $t = 0$ ):

$$\tilde{\alpha}_0(j) = \begin{cases} 1 & \text{for } j = 1 \\ 0 & \text{for } 1 < j \leq N \end{cases} \quad (4.34)$$

$$s_0 = 1 \quad (4.35)$$

2. Recursion ( $t = 1, \dots, T + 1$ ):

$$\hat{\alpha}_t(j) = b_j(X_t) \sum_{i=1}^N \tilde{\alpha}_{t-1}(i) a_{ij} \quad \text{for } 1 \leq j \leq N \quad (4.36)$$

$$s_t = \sum_{i=1}^N \hat{\alpha}_t(i) \quad (4.37)$$

$$\tilde{\alpha}_t(j) = \frac{\hat{\alpha}_t(j)}{s_t} \quad \text{for } 1 \leq j \leq N \quad (4.38)$$

Note that the algorithm to compute the scaled forward matrix  $\tilde{\alpha}$  differs from the algorithm for the unscaled forward matrix  $\alpha$  in that first an intermediate value  $\hat{\alpha}_t$  is computed, which is subsequently summed over to yield  $s_t$ . This sum is then used to scale the values in the matrix  $\hat{\alpha}$  for time  $t$ , yielding  $\tilde{\alpha}$ .

### 4.5.2 Algorithm to compute the scaled backward matrix

The following algorithm computes the backward matrix  $\tilde{\beta}$  using the scaling vector  $\mathbf{s}$ .

1. Initialization ( $t = T + 1$ ):

$$\tilde{\beta}_{T+1}(j) = \begin{cases} \frac{1}{s_{T+1}} & \text{for } j = 1 \\ 0 & \text{for } 1 < j \leq N \end{cases} \quad (4.39)$$

2. Recursion ( $t = T, \dots, 0$ ):

$$\hat{\beta}_t(i) = \sum_{j=1}^N a_{ij} b_j(X_{t+1}) \tilde{\beta}_{t+1}(j) \quad \text{for } 1 \leq i \leq N \quad (4.40)$$

$$\tilde{\beta}_t(i) = \frac{\hat{\beta}_t(i)}{s_t} \quad (4.41)$$

Note that in the algorithm for the scaled backward matrix the same  $s_t$  values are used that were determined in the calculation of the scaled forward matrix.

**Correctness and complexity** For a proof of the correctness of the scaling variants of the algorithms the reader is kindly referred to appendix A.

Runtime complexity of the scaling variants of the forward and backward algorithms is identical to that of the original algorithms. Note again that while a trivial realization of the above code is in  $\mathcal{O}(TN^2)$ , the algorithms can be implemented in  $\mathcal{O}(TE)$  by only considering transitions with non-zero probability.

### 4.5.3 Inference

**Computing the likelihood** Building on (4.30), and using the scaling coefficient vector  $\mathbf{s}$ , the likelihood of the observation sequence  $\mathbf{X} = X_1 X_2 \dots X_T$  given the model  $\theta$  may be computed by

$$\mathbb{L}(\theta|\mathbf{X}) = \mathbb{P}(\mathbf{X}|\theta) = \alpha_{T+1}(1) = \tilde{\alpha}_{T+1}(1) \prod_{t=0}^{T+1} s_t = \prod_{t=0}^{T+1} s_t = \prod_{t=1}^{T+1} s_t, \quad (4.42)$$



as  $\tilde{\alpha}_{T+1}(1) = 1$ , and  $s_0 = 1$ . Similarly, and numerically preferably, the log-likelihood may be determined by

$$\log \mathbb{L}(\boldsymbol{\theta}|\mathbf{X}) = \log \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) = \sum_{t=1}^{T+1} \log s_t. \quad (4.43)$$

**Computing posterior probability of being in some state** An expression based on the scaled variants of the forward and backward matrices,  $\tilde{\alpha}$  and  $\tilde{\beta}$  corresponding to (4.25) for the posterior probability of being in a state  $S_i$  at time  $t$  given the observation sequence  $\mathbf{X} = X_1 X_2 \dots X_T$  and the model  $\boldsymbol{\theta}$  is

$$\begin{aligned} \mathbb{P}(q_t = S_i | \mathbf{X}, \boldsymbol{\theta}) &= \frac{\alpha_t(i) \beta_t(i)}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})} \\ &= \frac{\tilde{\alpha}_t(i) \prod_{k=0}^t s_k \tilde{\beta}_t(i) \prod_{k=t}^{T+1} s_k}{\prod_{k=0}^{T+1} s_k} \\ &= \tilde{\alpha}_t(i) \tilde{\beta}_t(i) s_t. \end{aligned} \quad (4.44)$$

Similarly, (4.26) may be computed as

$$\begin{aligned} \mathbb{P}(q_t = S_i, q_{t+1} = S_j | \mathbf{X}, \boldsymbol{\theta}) &= \frac{\alpha_t(i) a_{ij} b_j(X_{t+1}) \beta_{t+1}(j)}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})} \\ &= \frac{\tilde{\alpha}_t(i) \prod_{k=0}^t s_k a_{ij} b_j(X_{t+1}) \tilde{\beta}_{t+1}(j) \prod_{k=t+1}^{T+1} s_k}{\prod_{k=0}^{T+1} s_k} \\ &= \tilde{\alpha}_t(i) a_{ij} b_j(X_{t+1}) \tilde{\beta}_{t+1}(j). \end{aligned} \quad (4.45)$$

**Expected transitions and emissions** Instead of using  $\alpha$  and  $\beta$ , the expected transitions may also be computed using the scaled forward and backward matrices  $\tilde{\alpha}$  and  $\tilde{\beta}$ . Then, instead of (4.27), we have

$$\mathcal{A}_{ij} = \sum_{t=0}^T \mathbb{P}(q_t = S_i, q_{t+1} = S_j | \mathbf{X}, \boldsymbol{\theta}) = \sum_{t=0}^T \tilde{\alpha}_t(i) a_{ij} b_j(X_{t+1}) \tilde{\beta}_{t+1}(j). \quad (4.46)$$

Similarly, instead of using (4.28), the expected emissions may be computed as

$$\mathcal{B}_i(x) = \sum_{\{t|X_t=x\}} \mathbb{P}(q_t = S_i | \mathbf{X}, \boldsymbol{\theta}) = \sum_{\{t|X_t=x\}} \tilde{\alpha}_t(i) \tilde{\beta}_t(i) s_t. \quad (4.47)$$

## 4.6 Likelihood gradient

We first consider expressions for the likelihood gradient of a single sequence  $\mathbf{X}$ . Equation (4.48) for the partial derivatives of the likelihood with respect to the transition prob-

abilities was given by Baum (1972),

$$\frac{\partial \mathbb{L}(\boldsymbol{\theta}|\mathbf{X})}{\partial a_{ij}} = \frac{\partial \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial a_{ij}} = \sum_{t=0}^T \alpha_t(i) b_j(X_{t+1}) \beta_{t+1}(j) \quad (4.48)$$

$$= \frac{\mathcal{A}_{ij}}{a_{ij}} \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}). \quad (4.49)$$

Equation (4.49) is derived by using the definition of the expected number of transitions from state  $S_i$  to state  $S_j$ ,  $\mathcal{A}_{ij}$ , in (4.27). From this we have the partial derivative of the log-likelihood with respect to the transition probabilities,

$$\frac{\partial \log \mathbb{L}(\boldsymbol{\theta}|\mathbf{X})}{\partial a_{ij}} = \frac{\partial \log \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial a_{ij}} = \frac{1}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})} \frac{\partial \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial a_{ij}} = \frac{\mathcal{A}_{ij}}{a_{ij}} \quad (4.50)$$

An expression for the partial derivative of the likelihood with respect to the emission probabilities corresponding to the one of Baum (1972) is

$$\frac{\partial \mathbb{L}(\boldsymbol{\theta}|\mathbf{X})}{\partial b_j(k)} = \frac{\partial \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial b_j(k)} = \sum_{\{t|X_t=k\}} \sum_{i=1}^N \alpha_{t-1}(i) a_{ij} \beta_t(j) \quad (4.51)$$

$$= \frac{1}{b_j(k)} \sum_{\{t|X_t=k\}} \alpha_t(j) \beta_t(j) \quad (4.52)$$

$$= \frac{\mathcal{B}_j(k)}{b_j(k)} \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}). \quad (4.53)$$

Equation (4.52) uses the definition of  $\alpha_t(j)$  in (4.21), equation (4.53) uses the definition of the expected number of emissions of kind  $k$  in state  $S_j$ ,  $\mathcal{B}_j(k)$  in (4.28). From this we get the partial derivative of the log-likelihood with respect to the emission probabilities,

$$\frac{\partial \log \mathbb{L}(\boldsymbol{\theta}|\mathbf{X})}{\partial b_j(k)} = \frac{\partial \log \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial b_j(k)} = \frac{1}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})} \frac{\partial \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial b_j(k)} = \frac{\mathcal{B}_j(k)}{b_j(k)}. \quad (4.54)$$

Equations (4.49), (4.50), (4.53) and (4.54) can all be found in Krogh (1994).

### 4.6.1 Transformed probabilities

In order to avoid boundary issues Mao and Hu (2001) suggest to consider quantities  $g_{ij}$  and  $h_{il}$ , which are defined to transform into the corresponding transition and emission probabilities as

$$a_{ij} = \frac{e^{g_{ij}}}{\sum_{k=1}^N e^{g_{ik}}} \quad (4.55)$$

and

$$b_i(l) = \frac{e^{h_{il}}}{\sum_{k=1}^M e^{h_{ik}}}. \quad (4.56)$$

Using that  $\frac{\partial a_{ij}}{\partial g_{kl}} = \delta_{ik} (\delta_{jl} - a_{ij}) a_{il}$  Mao and Hu give (4.57) for the partial derivative of

the likelihood with respect to the transformed transition probabilities  $g_{ij}$ ,

$$\frac{\partial \mathbb{L}(\boldsymbol{\theta}|\mathbf{X})}{\partial g_{ij}} = \frac{\partial \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial g_{ij}} = \sum_{k=1}^N \sum_{t=0}^T \alpha_t(i) b_k(X_{t+1}) \beta_{t+1}(k) (\delta_{kj} - a_{ij}) a_{ik} \quad (4.57)$$

$$= \sum_{k=1}^N (\delta_{kj} - a_{ij}) \sum_{t=0}^T \alpha_t(i) a_{ik} b_k(X_{t+1}) \beta_{t+1}(k) \quad (4.58)$$

$$= \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) \sum_{k=1}^N \mathcal{A}_{ik} (\delta_{kj} - a_{ij}). \quad (4.59)$$

While (4.58) is just a reordering of (4.57), (4.59) uses the definition of  $\mathcal{A}_{ik}$  in (4.27). From this we have the partial derivative of the log-likelihood with respect to the transformed transition probabilities

$$\frac{\partial \log \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial g_{ij}} = \frac{1}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})} \frac{\partial \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial g_{ij}} = \sum_{k=1}^N \mathcal{A}_{ik} (\delta_{kj} - a_{ij}). \quad (4.60)$$

Similarly, with  $\frac{\partial b_i(l)}{\partial h_{jk}} = \delta_{ij} (\delta_{lk} - b_i(l)) b_i(k)$  Mao and Hu give (4.61) for the partial derivative of the likelihood with respect to the transformed emission probabilities  $h_{jk}$  as

$$\frac{\partial \mathbb{L}(\boldsymbol{\theta}|\mathbf{X})}{\partial h_{jk}} = \frac{\partial \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial h_{jk}} = \sum_{l=1}^M \left( \sum_{\{t|X_t=l\}} \sum_{i=1}^N \alpha_{t-1}(i) a_{ij} \beta_t(j) \right) \cdot (\delta_{kl} - b_j(k)) b_j(l) \quad (4.61)$$

$$= \sum_{l=1}^M \frac{1}{b_j(l)} \left( \sum_{\{t|X_t=l\}} \alpha_t(j) \beta_t(j) \right) \cdot (\delta_{kl} - b_j(k)) b_j(l) \quad (4.62)$$

$$= \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) \sum_{l=1}^M \mathcal{B}_j(l) (\delta_{kl} - b_j(k)), \quad (4.63)$$

where  $M$  denotes the size of the alphabet. As above, (4.62) uses the definition of  $\alpha_t(j)$  in (4.21), and (4.62) the definition of the expected emissions  $\mathcal{B}_j(l)$ , in (4.28). From this we have the partial derivative of the log-likelihood with respect to the transformed emission probabilities

$$\frac{\partial \log \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial h_{jk}} = \frac{1}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})} \frac{\partial \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\partial h_{jk}} = \sum_{l=1}^M \mathcal{B}_j(l) (\delta_{kl} - b_j(k)), \quad (4.64)$$

## 4.6.2 Complexity

Here we collect the runtime complexities of HMM likelihood gradient calculations presented in this section. As for the Viterbi algorithm, and the computation of the forward and backward matrices, the runtime of computations of the gradient expressions is reduced by considering only non-zero terms.

Define  $T$  to be the length of the observation sequence,  $N$  the number of states, and  $M$  the size of the alphabet. Further, let  $E$  be the number of edges in the HMM topology (the number of non-zero transition probabilities), and  $F$  be the number of admissible emis-

sions in all states, i.e.  $F$  is the number of state-emission pairs  $(k, x)$  such that  $b_k(x)$  is non-zero. In general,  $N - 1 \leq E \leq N^2$  and  $N \leq F \leq NM$ .

**Likelihood gradient due to Baum** The calculation of partial derivatives of the likelihood with respect to the transition probabilities according to (4.48) can be done in  $\mathcal{O}(TE)$ . Similarly, the partial derivatives of the likelihood with respect to the emission probabilities according to (4.51) may be computed in  $\mathcal{O}(TE + F)$ , where the contribution  $F$  is due to respecting all state-emission pairs' partial derivatives. Together, using the expressions of Baum (1972) the likelihood gradient may thus be determined in  $\mathcal{O}(TE + F)$ .

**Likelihood gradient due to Krogh** The likelihood gradient expressions (4.50) and (4.54) due to Krogh (1994) are simpler than those of Baum (1972) but depend on calculation of the expected statistics, and thus have the same cumulative runtime of  $\mathcal{O}(TE + F)$ .

**Likelihood gradient due to Mao and Hu** The expressions for the partial derivatives of the likelihood with respect to the transformed transition probabilities in (4.57) and with respect to the transformed emission probabilities in (4.61) considerably increase the runtime to  $\mathcal{O}(TEN)$  and  $\mathcal{O}(TFN)$ , respectively. Note that, similar to how considering the number of non-zero transitions  $E$  rather than  $N^2$  yields more precise upper bounds for the runtime of the forward-back algorithm, consideration of numbers like that of pairs of non-zero transitions to a given state (summed over all states) may allow to bound the runtime of the expressions of Mao and Hu further.

Importantly, while the expressions that Mao and Hu also have runtime linear in the size of the data, their runtime has a coefficient that is larger than that of the expressions for the gradient of the untransformed probabilities.

**Likelihood gradient due to Maaskola** Basing the expressions of the likelihood gradient with respect to the transformed probabilities on the expected statistics as proposed here in (4.60) and (4.64) allows to reduce the runtime to the same asymptotic order as that of the untransformed probabilities' gradient, however with a larger constant<sup>1</sup> term.

Note again that the respective runtime of  $\mathcal{O}(EN)$  and  $\mathcal{O}(FM)$  to compute the transformed probabilities' likelihood gradient from the expected statistics in (4.60) and (4.64) could be bounded more stringently by considering the number of non-zero transition pairs to each state, and non-zero emission pairs of each state. Importantly, these expressions for the transformed probabilities' gradient, subsequently to determining the expected statistics, do not further dependent on data size.

---

<sup>1</sup>with respect to the data size

Table 4.1: Runtime complexity and inter-dependence of HMM inference algorithms. Dep.: Depends on algorithm #.  $T$ : length of the observation sequence.  $N$ : number of states.  $E$ : number of non-zero transition probabilities, with  $N - 1 \leq E \leq N^2$ .  $M$ : size of the alphabet,  $F$ : number of state-emission pairs  $(k, x)$  such that  $b_k(x)$  is non-zero, with  $N \leq F \leq NM$ . References: <sup>A</sup> Baum (1972), <sup>B</sup> Krogh (1994), <sup>C</sup> Mao and Hu (2001), <sup>D</sup> Maaskola and Rajewsky (2014).

#	Algorithm	Complexity	Dep.	Equations
1	Viterbi path	$\mathcal{O}(TE)$		(4.11)–(4.17)
2	Forward, unscaled / scaled	$\mathcal{O}(TE)$		(4.20)–(4.21) / (4.34)–(4.38)
3	Backward, unscaled / scaled	$\mathcal{O}(TE)$		(4.22)–(4.23) / (4.39)–(4.41)
4	Expected transitions	$\mathcal{O}(TE)$	2, 3	(4.27)
5	Expected emissions	$\mathcal{O}(TN)$	2, 3	(4.28)
6	$\mathbb{L}$ gradient <sup>A</sup>	$\mathcal{O}(TE + F)$	2, 3	(4.48), (4.51)
7	$\mathbb{L}$ gradient <sup>B</sup>	$\mathcal{O}(E + F)$	4, 5	(4.50), (4.54)
8	$\mathbb{L}$ gradient, transformed <sup>C</sup>	$\mathcal{O}(TEN + TFN)$	2, 3	(4.57), (4.61)
9	$\mathbb{L}$ gradient, transformed <sup>D</sup>	$\mathcal{O}(EN + FM)$	4, 5	(4.60), (4.64)



## Chapter 5

# Binding site hidden Markov model

This chapter presents an HMM-based probabilistic model for binding motifs in sequence context. The model makes use of a specific topology, described in section 5.1. Section 5.2 explains how posterior probabilities of motif occurrences are computed with this probabilistic model.

### 5.1 An HMM for binding sites

The probabilistic model is an HMM with start and end state, as defined in section 4.1.1. For ease of specification, the model is based on a default topology, illustrated in figure 5.1, that includes HMM states for one or more motifs, as well as for the sequence context. The model also includes a start and end state, displayed as separate states S and E in figure 5.1, but the implementation of Discover actually uses a combined start/end state. By default, the transition and emission probabilities of each state depend only on the state, realizing a 0<sup>th</sup> order Markov property. However, Discover allows for the emission probabilities of each state to additionally depend on the  $n$  previous emissions, and the emission order may be specified per state. This feature is intended to allow higher order emissions for the background state B, combined with 0<sup>th</sup> order emissions for the motif states.

#### 5.1.1 Motif chains

Each motif corresponds to a chain of states, indicated by the numbered states in figure 5.1. The motif states are sequentially connected such that the first motif chain state transitions to the second, the second to the third, and so on. There are transitions from the background state B, and from the start state S to the first of each motif chain. The last state of each motif chain may transition to the background state B, to the end state E, and to the first state of each motif chain.

**Insert states** As is visible in figure 5.1, insert states may be included between adjacent motif chain states. To avoid unnecessary parameters, by default no insert states are included, and the user has to specify where insert states should be allowed.

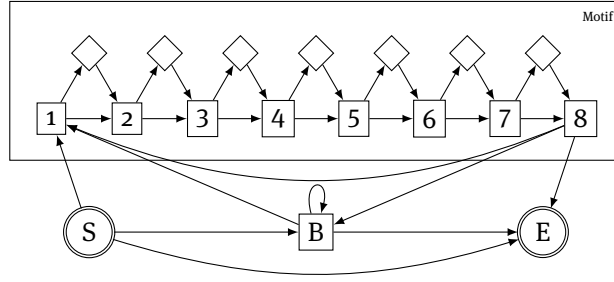


Figure 5.1: Default topology of a binding site HMM with a motif of 8 nucleotides length and a background state B. The model includes a start state S, and an end state E. The numbered states represent motif chain states, the diamond shaped states are (optional) insert states. The box around the motif is an instance of plate notation from the theory of probabilistic graphical models, see Koller and N. Friedman (2009), and indicates that there may be multiple motifs.

## 5.2 Posterior motif occurrence probability

Continuing with definitions of section 3.2.3, we designate the hypothesis of observing at least one motif occurrence as  $M$ . As posterior probability for motif occurrence in observation sequence  $\mathbf{X} = X_1 X_2 \dots X_T$  we compute the probability for a binding site HMM  $\theta$  to transition at least once through the chain of states corresponding to that motif.

Transitioning at least once through the chain of motif states is the complementary event to not transitioning through the chain at all. Thus, the posterior probability of at least one pass through the motif chain is given by subtracting from 1 the posterior probability of not passing through the motif chain,

$$\mathbb{P}(M|\mathbf{X}, \theta) = 1 - \mathbb{P}(\neg M|\mathbf{X}, \theta). \quad (5.1)$$

Due to Bayes' theorem, the posterior probability of not transitioning through the motif chain can be computed as the ratio of the likelihood of not transitioning through the motif chain and the unrestricted likelihood,

$$\mathbb{P}(\neg M|\mathbf{X}, \theta) = \frac{\mathbb{P}(\neg M, \mathbf{X}|\theta)}{\mathbb{P}(\mathbf{X}|\theta)}. \quad (5.2)$$

For the denominator, the likelihood of  $\theta$  given the observation sequence  $\mathbf{X}$ ,  $\mathbb{P}(\mathbf{X}|\theta)$ , we have equation (4.29). For the numerator we have

$$\mathbb{P}(\neg M, \mathbf{X}|\theta) = \sum_{\mathbf{q}} \mathbb{P}(\mathbf{q}, \mathbf{X}|\theta), \quad (5.3)$$

where the summation is over all paths  $\mathbf{q}$  that do not transition through the motif state chain. This summation can be easily performed with a slight modification of the forward-backward procedure that ignores transitions to, within, and from the motif state chain.

Note that when the binding site HMM  $\theta$  comprises chains for multiple motifs, the above described procedure allows to separately compute the posterior probabilities of



each individual motif occurring at least once. Similarly, it is also possible to compute whether at least one occurrence of any of a set of motifs is present in a sequence.

**Gradient** The gradient of the posterior occurrence probability is

$$\begin{aligned}
 \nabla \mathbb{P}(\mathbf{M}|\mathbf{X}, \boldsymbol{\theta}) &= -\nabla \mathbb{P}(\neg \mathbf{M}|\mathbf{X}, \boldsymbol{\theta}) \\
 &= -\nabla \frac{\mathbb{P}(\neg \mathbf{M}, \mathbf{X}|\boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})} \\
 &= \frac{\mathbb{P}(\neg \mathbf{M}, \mathbf{X}|\boldsymbol{\theta}) \nabla \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) - \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) \nabla \mathbb{P}(\neg \mathbf{M}, \mathbf{X}|\boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})^2} \tag{5.4} \\
 &= \frac{\mathbb{P}(\neg \mathbf{M}, \mathbf{X}|\boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})} \left( \frac{\nabla \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})} - \frac{\nabla \mathbb{P}(\neg \mathbf{M}, \mathbf{X}|\boldsymbol{\theta})}{\mathbb{P}(\neg \mathbf{M}, \mathbf{X}|\boldsymbol{\theta})} \right) \\
 &= \mathbb{P}(\neg \mathbf{M}|\mathbf{X}, \boldsymbol{\theta}) (\nabla \log \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) - \nabla \log \mathbb{P}(\neg \mathbf{M}, \mathbf{X}|\boldsymbol{\theta})).
 \end{aligned}$$

This expresses the gradient of the posterior motif occurrence probability in terms of the gradients of the log likelihood,  $\nabla \log \mathbb{P}(\mathbf{X}|\boldsymbol{\theta})$ , and of the log likelihood of no motif occurrence,  $\nabla \log \mathbb{P}(\neg \mathbf{M}, \mathbf{X}|\boldsymbol{\theta})$ .

Expressions for the gradient of the log likelihood are given in section 4.6. In particular, equations (4.50) and (4.54) give expressions for the partial derivatives of the log likelihood with respect to transition and emission probabilities, respectively. Alternatively, when transformed probabilities are used, then equations (4.60) and (4.64) respectively give the corresponding expressions with respect to transformed transition and emission probabilities.

The gradient of the log likelihood of no motif occurrence,  $\nabla \log \mathbb{P}(\neg \mathbf{M}, \mathbf{X}|\boldsymbol{\theta})$ , can be computed similarly, using the same equations as for the gradient of the log likelihood. The only difference is that the expected counts of transitions  $\mathcal{A}_{ij}$  and emissions  $\mathcal{B}_j(k)$  in equations (4.50) and (4.54) or equations (4.60) and (4.64) need to be computed disallowing transitions to, within, and from the motif chain.



## Chapter 6

# Learning algorithms

The solution to the learning problem depends on the choice of criteria for suitability of parameters. The classical paradigm is to choose the parameters such that the likelihood of the set of observations is maximized, and is called maximum likelihood (ML) learning. Baum and Eagon (1967) presented a theorem about an inequality for homogeneous polynomials with non-negative coefficients that guarantees that an expectation-maximization (EM) (Dempster, Laird, and Rubin, 1977) called scheme increases the likelihood in each iteration until a critical point<sup>1</sup> is reached. This inequality was later generalized and re-expressed in terms of an inequality for continuous transformations by Baum (1972) and Baum, Petrie, et al. (1970). The corresponding EM training for HMMs has become known as Baum-Welch (BW) algorithm and is described in section 6.1.

This chapter also outlines alternatives to the BW algorithm. Viterbi learning, which is structurally similar to EM but does not optimize the likelihood, is described in section 6.1.3. Gradient optimization, applicable to the optimization of arbitrary objective functions, can also be used to maximize the likelihood. This chapter concludes by introducing in section 6.2 the gradient optimization framework used later in this thesis to optimize discriminative objective functions.

All learning algorithms presented in this chapter are iterative, local search algorithm. This means that they depend on the choice of starting parameters for the search, which are iteratively improved in the course of optimization. In particular, the algorithms are not guaranteed to find global optima, or might require multiple starts with different initial parameters to increase the likelihood of finding global optima.

### 6.1 Baum-Welch algorithm

Like the general EM approach, BW training is suitable for inference given partial data. Partial data modeling fundamentally characterizes the HMM framework. In particular, an HMM is a representation for systems comprising entities that are observed, i.e. for which data are available, and entities that are unobserved, for which in other words the state must be inferred.

---

<sup>1</sup>Points where the first derivative is zero, i.e. (local or global) maxima or minima, or saddle points.

At its basic level, the BW algorithm operates by iteratively executing two steps. The first, the so called expectation, or E-step, consists in determining the expected states of all modeled entities, in particular of the unobserved ones. Given these expected states, the maximization step consists in modifying the parameters so as to maximize the likelihood of generating this expected state assignment.

### 6.1.1 Expected parameters and likelihood maximization

Algorithm 1 gives pseudo code for the E-step of the BW algorithm. First, the forward and backward matrices  $\alpha^m$  and  $\beta^m$ , or their scaled variants  $\tilde{\alpha}^m$  and  $\tilde{\beta}^m$ , are computed for each observation sequence  $\mathbf{X}^m$ . From these, for each sequence  $\mathbf{X}^m$ , are then computed the expected number of transitions  $\mathcal{A}_{ij}^m$  from state  $S_i$  to state  $S_j$ , as well as the expected number of emissions  $\mathcal{B}_i^m(x)$  of symbol  $x$  in state  $S_i$ . Expressions for the above-mentioned calculations are given in sections 4.4 and 4.5.

Note that upper indices are used here to denote the observation to which a variable pertains. Then, by summing over all  $M$  observation sequences, the total expected numbers of transitions  $\mathcal{A}_{ij}$  from state  $S_i$  to state  $S_j$  are determined,

$$\mathcal{A}_{ij} = \sum_{m=1}^M \mathcal{A}_{ij}^m. \quad (6.1)$$

Similarly, by summing over all observation sequences, the total expected numbers of emissions of kind  $x$  in state  $S_i$  are determined,

$$\mathcal{B}_i(x) = \sum_{m=1}^M \mathcal{B}_i^m(x). \quad (6.2)$$

During the maximization step the parameters are updated so as to maximize the likelihood with which the expected hidden variable assignments are observed. As shown by Baum (1972), this amounts to row-normalizing the expected count matrices  $\mathcal{A}$  and  $\mathcal{B}$ :

$$a_{ij} = \frac{\mathcal{A}_{ij}}{\sum_{k=1}^N \mathcal{A}_{ik}}, \quad (6.3)$$

and

$$b_i(x) = \frac{\mathcal{B}_i(x)}{\sum_k \mathcal{B}_i(k)}. \quad (6.4)$$

Algorithm 2 gives pseudo code for the M-step of the BW algorithm.

### 6.1.2 Pseudocode

The pseudocode in algorithm 3 realizes the BW algorithm. It starts with an arbitrary or random parameterization that is successively improved. The only requirements regarding the initial parameterization needs to fulfill are consistency with the desired topology and emission characteristics of the HMM. These must be such that any transition and emission events that are assumed possible have non-zero probability. The reason for this is that any event with zero probability will keep this zero probability during the BW routine.

**Algorithm 1** Baum-Welch expectation step**Input:** HMM parameters  $\theta$ , observation sequences  $\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^M$ **Output:** Expected number of transitions  $\mathcal{A} = \mathcal{A}_{ij}$  from state  $S_i$  to state  $S_j$  and expected number of emissions  $\mathcal{B} = \mathcal{B}_i(x)$  in state  $S_i$  of symbol  $x$ 

- 1: **for**  $m = 1$  **to**  $M$  **do**
- 2:   compute  $\alpha^m$  or  $\tilde{\alpha}^m$  using  $\mathbf{X}^m$  and  $\theta$
- 3:   compute  $\beta^m$  or  $\tilde{\beta}^m$  using  $\mathbf{X}^m$  and  $\theta$
- 4:   compute  $\mathcal{A}_{ij}^m$  for  $1 \leq i, j \leq N$  with (4.27) or (4.46)
- 5:   compute  $\mathcal{B}_i^m(x)$  for  $1 \leq i \leq N$  and all observations  $x$  with (4.28) or (4.47)
- 6: compute  $\mathcal{A}_{ij}$  for  $1 \leq i, j \leq N$  with (6.1)
- 7: compute  $\mathcal{B}_i(x)$  for  $1 \leq i \leq N$  and all observations  $x$  with (6.2)
- 8: **return**  $\{\mathcal{A}, \mathcal{B}\}$

**Algorithm 2** Baum-Welch maximization step**Input:** Expected number of transitions  $\mathcal{A}$  and expected number of emissions  $\mathcal{B}$ **Output:** HMM parameters  $\theta' = \{\mathbf{A}, \mathbf{b}\}$  with  $\mathbb{P}(\mathbf{X}|\theta') \geq \mathbb{P}(\mathbf{X}|\theta)$ 

- 1: compute  $\mathbf{A}$  using  $\mathcal{A}$  with (6.3)
- 2: compute  $\mathbf{b}$  using  $\mathcal{B}$  with (6.4)
- 3: **return**  $\{\mathbf{A}, \mathbf{b}\}$

As given here, the BW algorithm terminates after a maximal number of iterations has been performed, or when the improvement in log likelihood over the previous iteration falls below a threshold. An alternative for the latter criterion is to threshold on a norm of the parameter vector difference to the previous iteration. Frequently, the  $L^1$ -norm<sup>2</sup> is used for this purpose.

### 6.1.3 Viterbi learning

Viterbi learning is similar to the BW algorithm, as it uses the same algorithmic structure of iteratively re-estimating the parameters. Differently from BW training, in which the expected state assignment determines a distribution over possible states that explain an observation, Viterbi learning only considers the single most likely state path.

**Statistics of transitions and emissions of the Viterbi path** For each sequence  $\mathbf{X}^m$  Viterbi learning first determines the Viterbi path  $\mathbf{q}^{*m}$  using the Viterbi algorithm of section 4.3, and then extracts the number of transitions from state  $S_i$  to state  $S_j$  in the Viterbi path of observation  $m$ ,

$$\mathcal{A}_{ij}^m = \sum_{t=1}^{T^m} \mathbf{1}_{q_t^{*m}=i \wedge q_{t+1}^{*m}=j}, \quad (6.5)$$

and subsequently in all observations,

$$\mathcal{A}_{ij} = \sum_{m=1}^M \mathcal{A}_{ij}^m. \quad (6.6)$$

Similarly, the number of emissions of symbol  $x$  in state  $S_i$  is counted in the Viterbi path

<sup>2</sup>The sum of absolute values.

**Algorithm 3** Baum-Welch

**Input:** initial HMM parameters  $\theta$ , observation sequences  $\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^M$ , tolerance  $\varepsilon$ , maximal number of iterations  $n$

**Output:** HMM parameters  $\theta$  at a critical point of the likelihood

```

1:  $l_0 \leftarrow -\infty$ 
2:  $k \leftarrow 0$ 
3: repeat
4:    $k \leftarrow k + 1$ 
5:    $\mathcal{A}, \mathcal{B} \leftarrow$  Baum-Welch expectation step using  $\mathbf{X}$  and  $\theta$ 
6:    $\theta \leftarrow$  Baum-Welch maximization step using  $\mathcal{A}$  and  $\mathcal{B}$ 
7:    $l_k \leftarrow \log \mathbb{L}(\theta | \mathbf{X})$ 
8: until  $l_k - l_{k-1} < \varepsilon$  or  $k > n$ 
9: return  $\theta$ 

```

of observation sequence  $\mathbf{X}^m$ ,

$$\mathcal{B}_i^m(x) = \sum_{\{t | X_t^m = x\}} 1_{q_t^* = i}. \quad (6.7)$$

Subsequently, the emission counts of symbol  $x$  in state  $S_i$  are added for all observations,

$$\mathcal{B}_i(x) = \sum_{m=1}^M \mathcal{B}_i^m(x). \quad (6.8)$$

**Maximization step** Based on the transition and emission counts observed in the Viterbi paths of all observations, Viterbi learning uses the same expressions (6.3) and (6.4) to re-estimate the transition and emission parameters. BW training increases the likelihood in each iteration until convergence. The similar algorithmic structure of Viterbi learning suggests that it maximizes the likelihood of the most likely decoding.

## 6.2 Gradient learning

Many learning methods used in this thesis are based on gradient optimization, an iterative, local learning method. It allows to optimize arbitrary differentiable objective functions  $f(\mathbf{X}, \theta)$  of some data  $\mathbf{X}$  and a model  $\theta$ , only requiring an expression for the gradient,  $\nabla_{\theta} f(\mathbf{X}, \theta)$ . In each iteration of gradient optimization, the gradient is computed and used to improve the current parameter estimate. In iteration  $k$ , the parameter estimate  $\theta_{k+1}$  is computed by determining the direction of steepest increase of the objective function  $\nabla_{\theta} f(\mathbf{X}, \theta)|_{\theta=\theta_k}$  at the point  $\theta_k$  and taking a step of length  $\alpha_k$  into that direction.

$$\theta_{k+1} = \theta_k + \alpha_k \nabla_{\theta} f(\mathbf{X}, \theta)|_{\theta=\theta_k} \quad (6.9)$$

A suitable step size  $\alpha_k$  is computed in each iteration  $k$  using one of several algorithms to perform line search in the direction of the gradient (Press et al., 1995).

Discover uses the Moré-Thuente line search algorithm (Moré and Thuente, 1994) to ensure sufficient increase and proximity to the local maximum along the search direction. Specifically, the algorithm determines such an  $\alpha_k = \alpha$  that fulfills the following two con-

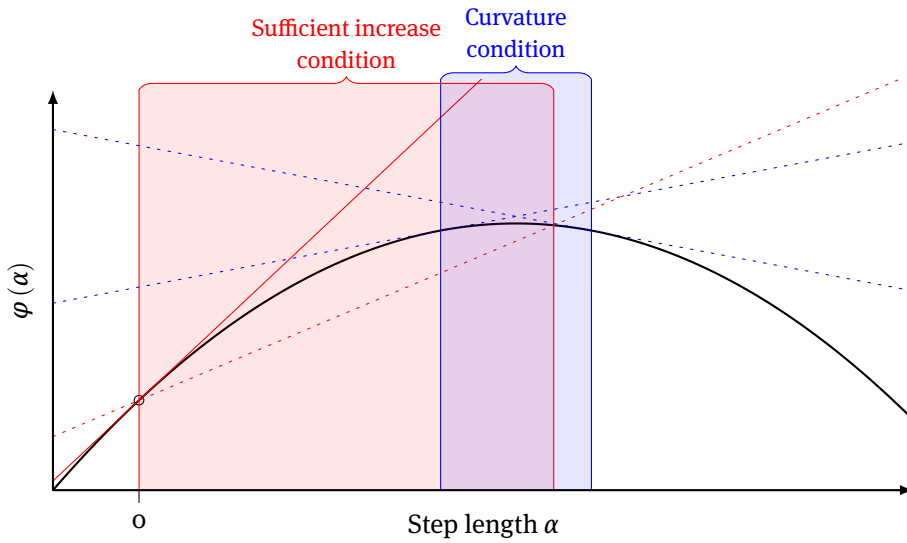


Figure 6.1: Conditions enforced by the Moré-Thuente line searching algorithm. The black curve gives the value of the auxiliary function  $\varphi(\alpha)$  at  $\alpha$ . The position  $\alpha = 0$  corresponds to the current parameter estimate, positive values of  $\alpha$  move from the current parameter estimate in direction of the gradient. The sufficient increase condition (6.10) is fulfilled in the red interval, defined by the intersections of the black curve  $\varphi(\alpha)$  and the red dashed line  $\varphi(0) + \mu\varphi'(0)\alpha$ . The curvature condition (6.11) is fulfilled in the blue interval, defined by the blue dashed lines which are tangents to  $\varphi(\alpha)$  at  $\alpha$  such that  $|\varphi'(\alpha)| = n|\varphi'(0)|$ .

ditions:

$$\text{Sufficient increase condition} \quad \varphi(\alpha) \geq \varphi(0) + \mu\varphi'(0)\alpha \quad (6.10)$$

$$\text{Curvature condition} \quad |\varphi'(\alpha)| \leq n|\varphi'(0)|, \quad (6.11)$$

where  $\mu$  and  $n$  are parameters of the algorithm and  $\varphi(\alpha)$  is an auxiliary function that gives the objective function value at a distance of  $\alpha$  from the current parameter estimate in the direction of the gradient,

$$\varphi(\alpha) = f(\mathbf{X}, \boldsymbol{\theta} + \alpha\nabla_{\boldsymbol{\theta}}f(\mathbf{X}, \boldsymbol{\theta})), \quad (6.12)$$

and  $\varphi'(\alpha)$  is the derivative of this function. Figure 6.1 illustrates these conditions.

## 6.3 Complexity

Convergence to a fixed point has been proven for the BW algorithm in the general case (Baum, Petrie, et al., 1970), but it is not clear how many iterations may be needed. Similarly, for gradient learning, supposing that line search finds a parameter with higher objective function value in a finite number of tries, or that gradient search terminates in case no suitable point can be found during line searching, a bounded objective function implies convergence of gradient learning in finite time.

While thus the number of iterations until convergence is unknown, we discuss below

the computational complexity of each iteration using variables defined in section 4.6.2.

**Re-estimation learning** First, as noted in section 4.4, computation of forward and backward matrices, of their scaled variants, and of the Viterbi path is done in  $\mathcal{O}(TE)$ . Thus, the computation for the E-step of the BW and Viterbi learning algorithms can be done in  $\mathcal{O}(TE)$ . Subsequently, the M-step is done in  $\mathcal{O}(E + F)$ . In summary, the runtime for one iteration of re-estimation learning is  $\mathcal{O}(TE + F)$ .

**Gradient learning** Gradient learning needs to compute the gradient of the objective function at least once per iteration. While the complexity of gradient calculations for arbitrary models can not be discussed in general, for HMMs we have seen that the likelihood gradient is computed in time  $\mathcal{O}(TE + F)$  for the untransformed probabilities, and in time  $\mathcal{O}(TE + EN + FM)$  for transformed probabilities, which is in both cases linear in the amount of data.

The line searching step of gradient learning requires several evaluations of the objective function until a suitable point has been found. Clearly, the choice of line searching procedure is of influence (see Moré and Thuente, 1994, for details). Thus, we here just relate our empirical experience that in the implementation of Discover, which uses the Moré-Thuente line search algorithm, typical iterations only need two function evaluations in addition to the gradient computation.



## **Part III**

# **Discriminative Learning for Probabilistic Sequence Analysis**



# Overview of discriminative learning methods

The following part collects the discriminative learning methods for probabilistic sequence analysis that constitute the main contributions of this thesis. Chapter 7 introduces the statistics used in the discriminative sequence analysis approaches presented here. Objective functions that measure association of motif occurrence with the conditions of a contrast are explained in chapter 8. In order to avoid clashing mathematical notation, we separated the description of classification probability based objective functions to chapter 9. Where simple expressions are available, chapters 8 and 9 also give expressions for the gradients of these objective functions in order to allow gradient-based optimization. Significance of association depends both on the number of parameters and the amount of data available, and chapter 10 proposes measures of significance of association for this purpose. An application of the objective functions to identify informative words is presented in chapter 11. Chapter 12 explains how generative and discriminative objectives may be used to jointly train binding site HMM parameters. Chapter 14 concludes this part with an overview of published DMD programs.



## Chapter 7

# Statistics for discriminative learning

Discriminative learning requires statistics of a feature of interest across a suitably chosen contrast between a set of signal and control data samples. This chapter first discusses suitable contrasts for MD by discriminative learning in section 7.1. For conditions of such contrasts section 7.2 then presents tables with counts of signal and control sequences that have at least one motif occurrence.

### 7.1 Contrasts

In many applications the data analysis practitioner is given data containing a signal of interest and is asked to characterize the signal. In order to elicit which parts of the data constitute signal, the discriminative learning approach suggests to compare what differentiates real data from suitably chosen control data. Naturally, the choice of control data strongly affects the chances of successfully discovering the signal.

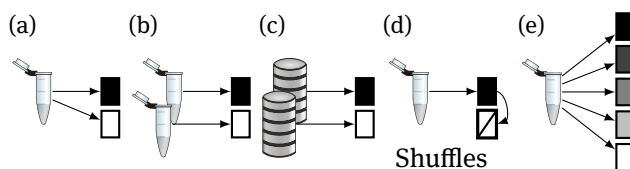


Figure 7.1: Contrasts for discriminative sequence analysis. Test tubes represent samples from different biological conditions, boxes are sets of sequences with blackness indicating the true positive rate in the set. (a) Binary contrast from a single biological sample, e.g. bound and not bound, or expressed and not expressed. (b) Binary contrast of different biological conditions, e.g. pull-down and mock, two different tissues, cell-types, or treatments. (c) Binary contrast of sequences from two different database searches. (d) Binary contrast in which the data of the contrasting condition is synthesized from the signal data by shuffling. (e) Contrast from grading the signal strength.

**Binary contrasts** For MD, a minimal contrast to which discriminative learning is applicable is that in which a single experiment yields evidence for binding to one set of sequences, and no evidence for another, see figure 7.1a. Another suitable binary contrast may be to compare the sequences that have more binding evidence in the first of a pair of experiments with those that have more evidence in the second, as depicted in figure 7.1b. Similarly, different database queries could provide signal and control sequences, see figure 7.1c. In case no suitable biological control is available it is possible to synthesize a control set of sequences, as illustrated in figure 7.1d. For example, this may be done by shuffling the signal sequences keeping word frequencies up to some order.

Notable ways of procuring biological control samples for binary contrasts comprise mock treatments, chromatin accessibility stratification for DBPs, and expression stratification for RBPs. The latter can for example be achieved by sampling sequences directly from RNA-Seq data of the same tissue.

**Contrasts with more than two conditions** Another possibility is that an experiment gives rise to a binding evidence rank order of the sequences. Then, the analyst may break the data into groups by their ranks, as indicated in figure 7.1e, and study which rank-grouping is useful to elicit the signal.

**Repeat experiments** Frequently data from repeat experiments is available. While in such cases multiple independent analyses may be applied to determine whether the results are consistent, it is also possible to analyze repeat experiments jointly so as to increase statistical power and sensitivity.

## 7.2 Contingency tables of features across contrasts

Differential binding affinity of a nucleic-acid binding factor to sets of bound and unbound sequences is a necessary precondition for an efficient recognition and selection of specific targets. It is hoped that the characteristics that distinguish sets of bound and unbound sequences may be elucidated by studying their statistical properties. Thus, upon performing an experiment that indicates that a set of sequences may potentially be bound by a nucleic-acid binding factor we still need to answer two critical questions.

The first question is which of these potentially bound sequences are truly bound and which are false positives of the experiment. A variant of this question is which sequences are directly bound, which are bound by co-factors, and which are just experimental artifacts.

The second question is related to the first, and answers to it help to tell true from false positives. What are the characteristics that separate true from false positives, or directly from indirectly bound sequences and from artifacts?

In order to address these questions we need to collect statistics of potentially relevant features. For this purpose the following two statistics may be considered. The first is about sequences that have at least one feature occurrence, the other about words that constitute feature occurrences. Reasoning that a sequence exhibiting a certain feature at least once

might be sufficient for explaining the recognition of the sequence, we use as feature the question whether a sequence has *at least one occurrence* of a given motif. Alternatively, because multiple occurrences of a motif in a sequence might induce a stronger regulatory response, also the *number of occurrences per sequence* may be a relevant feature. Here, we focus on the prior feature, but the methods presented here might be extended to the latter case too.

We tabulate in how many of the potentially bound sequences a given feature occurs, and in how many it is not observed. As features that are observed in most of the potentially bound sequences may also be frequent in unbound sequences, one can procure a second dataset of sequences for which no evidence for binding is available. When two such sets of sequences are given, we refer to these sets of sequences as signal and control, respectively.

For each potentially interesting feature  $\theta$  we then tabulate the number of sequences in signal and control that have the feature and those that do not, as illustrated in table 7.1.

Table 7.1:  $2 \times 2$  contingency table of number of sequences in datasets of two conditions that have or do not have at least one occurrence of a motif. *TP* and *FP* stand for true and false positives, *TN* and *FN* for true and false negatives, respectively.

Condition	Motif present	Motif absent
Signal	<i>TP</i>	<i>FN</i>
Control	<i>FP</i>	<i>TN</i>

$2 \times 2$  contingency tables as depicted in table 7.1 are applicable in binary classification problems on data representing contrasts involving a pair of positive and negative example sets.

Similarly, when there are more than two conditions, we may consider contingency tables as in table 7.2.

Table 7.2:  $k \times 2$  contingency table of number of sequences with and without a feature in datasets of  $k$  conditions.  $m_i$  gives the number of sequences with motif in the regulatory sequences of dataset  $i$ .  $n_i$  is the number of sequences in dataset  $i$ .

Condition	Motif present	Motif absent
1	$m_1$	$n_1 - m_1$
2	$m_2$	$n_2 - m_2$
$\vdots$	$\vdots$	$\vdots$
$k$	$m_k$	$n_k - m_k$

Note that for probabilistic motif models  $\theta$  these contingency tables will hold expected counts  $m_i(\theta)$  of sequences with motif occurrence in dataset  $X^i$ ,

$$m_i(\theta) = \mathbb{E} [\mathbb{P}(M|X^i, \theta)] = \sum_{\mathbf{X} \in X^i} \mathbb{P}(M|\mathbf{X}, \theta), \quad (7.1)$$

where the expectation and summation are over the sequences  $\mathbf{X}$  in dataset  $X^i$ , and the posterior probability of motif occurrence in one sequence  $\mathbf{X}$ ,  $\mathbb{P}(M|\mathbf{X}, \theta)$ , is computed as explained in section 5.2.

**Gradient** Due to the linearity of the expectation operator, the gradient of the expected counts of sequences with motif occurrences in dataset  $\mathbf{X}^i$  is trivially given by the sum of gradients of the probability of motif occurrence in the individual sequences  $\mathbf{X} \in \mathbf{X}^i$ ,

$$\nabla m_i(\boldsymbol{\theta}) = \nabla \mathbb{E} [\mathbb{P}(M|\mathbf{X}^i, \boldsymbol{\theta})] = \sum_{\mathbf{X} \in \mathbf{X}^i} \nabla \mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta}). \quad (7.2)$$

An expression for the gradient of the posterior probability of occurrence of a motif  $M$  in a sequence  $\mathbf{X}$  given the model  $\boldsymbol{\theta}$ ,  $\nabla \mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta})$ , is given in equation (5.4).



## Chapter 8

# Table-based discriminative objective functions

This chapter discusses discriminative objective functions based on contingency tables that are of potential interest for MD problems. Table 8.1 illustrates how some of these quantify the association of motif occurrences with conditions in several small hypothetical datasets. Some of these objective functions are directional, i.e. motifs that maximize them are not only differential but in fact enriched in the signal sequences. By considering motifs that minimize directional objective functions one can identify differential motifs that are depleted in the signal sequences. For non-directional objective functions it is possible to filter differential motifs for enrichment in the desired sample.

We first discuss measures of association computable from  $2 \times 2$  contingency tables, as in table 7.1. Then we will turn to  $k \times 2$  contingency table based measures, as in table 7.2. Where simple expressions are available, expressions for the gradients of the objective functions will also be given. The gradients of the objective functions given in this chapter are based on the gradient of the expected site counts,  $\nabla m_i(\boldsymbol{\theta}) = \nabla \mathbb{E} [\mathbb{P}(M|\mathbf{X}^i, \boldsymbol{\theta})]$ , in equation (7.2).

### 8.1 Difference of occurrence frequency

The first directional measure of association that we want to discuss here is the difference of relative frequency (DFREQ) of motif prevalence in the signal and control data. It is defined in terms of the number of true and false positives  $TP$  and  $FP$ , and of true and false negatives  $TN$ ,  $FN$ ,

$$\begin{aligned} \text{DFREQ} &= \frac{TP}{TP + FN} - \frac{FP}{FP + TN} \\ &= \frac{m_1(\boldsymbol{\theta})}{n_1} - \frac{m_2(\boldsymbol{\theta})}{n_2}, \end{aligned} \tag{8.1}$$

where the second expression uses the variables introduced in table 7.2, and we assume that conditions 1 and 2 represent signal and control, respectively. One may argue that, measuring enrichment rather than significance of enrichment, a possible shortcoming of

Table 8.1: Comparison of discriminative performance of four motifs on regulatory sequences in a contrast of two conditions. Motifs 1 and 4 are discriminative, with positive correlation of motif 1 with condition A, and equally strong anti-correlation in the case of motif 4. Motif 2 is weakly discriminative, and motif 3 is neutral with respect to occurrences in the two conditions. Seq: regulatory sequences, Occ: motif occurrences in the sequences, filled and empty circles denote (counts of) sequences with at least one motif occurrence. DFREQ: difference of relative frequency of sequences with motif occurrence between conditions,  $p$ -Fisher:  $p$ -value according to Fisher's exact test, MCC: Matthews correlation coefficient, MICO: Mutual information of condition and motif occurrence. Even for motifs 1 and 4 the  $p$ -values according to Fisher's exact test are not significant due to the small sizes of these hypothetical datasets.

Condition	Seq	Motif 1			Motif 2			Motif 3			Motif 4		
		Occ	●	○	Occ	●	○	Occ	●	○	Occ	●	○
A			4	1		5	0		2	3		1	4
B			1	4		4	1		2	3		4	1
	<b>DFREQ</b>	60 %			20 %			0 %			-60 %		
	<b><math>p</math>-Fisher</b>	0.2063			1			1			0.2063		
	<b>MCC</b>	0.6			0.3			0			-0.6		
	<b>MICO</b>	1.92 bit			0.31 bit			0 bit			1.92 bit		

this measure is the failure to assign higher relevance to qualitative differences between signal and control samples. To illustrate, consider the following two contingency tables,

$$T_1 = \begin{pmatrix} 1000 & 0 \\ 500 & 500 \end{pmatrix} \quad \text{and} \quad T_2 = \begin{pmatrix} 950 & 50 \\ 450 & 550 \end{pmatrix}.$$

For both  $T_1$  and  $T_2$  the relative frequency difference  $\text{DFREQ} = \frac{1}{2}$ , but there is a qualitative difference in that the data of  $T_1$  are consistent with the hypothesis  $FN = 0$ , while there is evidence contradicting this hypothesis for  $T_2$ .

**Gradient** The gradient of DFREQ of sequences with motifs between the signal and control is given by

$$\nabla \text{DFREQ} = \frac{\nabla m_1(\boldsymbol{\theta})}{n_1} - \frac{\nabla m_2(\boldsymbol{\theta})}{n_2}. \quad (8.2)$$

## 8.2 Matthews correlation coefficient

Another frequently used directional measure of association is Matthews correlation coefficient (MCC) (Matthews, 1975), defined as

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(FN + TN)(FP + TN)}}. \quad (8.3)$$

Adopting the notation of table 7.2 and assuming that conditions 1 and 2 respectively represent signal and control, we have

$$\text{MCC} = \frac{1}{\sqrt{n_1 n_2}} \cdot \frac{n_2 m_1(\boldsymbol{\theta}) - n_1 m_2(\boldsymbol{\theta})}{\sqrt{(m_1(\boldsymbol{\theta}) + m_2(\boldsymbol{\theta}))(N - m_1(\boldsymbol{\theta}) - m_2(\boldsymbol{\theta}))}}, \quad (8.4)$$

where  $N = n_1 + n_2$ .

As the name indicates, the MCC is a proper measure of correlation, i.e. it takes on values between -1 and 1, where a value of 1 indicates perfect correlation, a value of -1 an inverse perfect correlation, and a value of 0 statistical independence.

The MCC of the matrices  $T_1$  and  $T_2$  is 0.577 and 0.546, respectively, demonstrating that the MCC assigns higher relevance to the association observed in the case of  $T_1$ .

**Gradient** To derive an expression for the gradient of the MCC, we consider first the gradient of the numerator of the second term of equation (8.4),

$$\nabla (n_2 m_1(\boldsymbol{\theta}) - n_1 m_2(\boldsymbol{\theta})) = n_2 \nabla m_1(\boldsymbol{\theta}) - n_1 \nabla m_2(\boldsymbol{\theta}). \quad (8.5)$$

The gradient of the denominator of the second term of equation (8.4) is

$$\begin{aligned} \nabla \sqrt{(m_1(\boldsymbol{\theta}) + m_2(\boldsymbol{\theta}))(N - m_1(\boldsymbol{\theta}) - m_2(\boldsymbol{\theta}))} \\ = \frac{(\frac{N}{2} - m_1(\boldsymbol{\theta}) - m_2(\boldsymbol{\theta})) \cdot (\nabla m_1(\boldsymbol{\theta}) + \nabla m_2(\boldsymbol{\theta}))}{\sqrt{(m_1(\boldsymbol{\theta}) + m_2(\boldsymbol{\theta}))(N - m_1(\boldsymbol{\theta}) - m_2(\boldsymbol{\theta}))}}. \end{aligned} \quad (8.6)$$

Using these two expressions, the quotient rule for differentiation, and canceling a few terms, we have the following expression for the gradient of the MCC,

$$\begin{aligned} \nabla \text{MCC} = \frac{1}{\sqrt{n_1 n_2}} \frac{1}{\sqrt{(m_1(\boldsymbol{\theta}) + m_2(\boldsymbol{\theta}))(N - m_1(\boldsymbol{\theta}) - m_2(\boldsymbol{\theta}))}} \left( n_2 \nabla m_1(\boldsymbol{\theta}) - n_1 \nabla m_2(\boldsymbol{\theta}) \right. \\ \left. + \frac{(n_2 m_1(\boldsymbol{\theta}) - n_1 m_2(\boldsymbol{\theta})) (\frac{N}{2} - m_1(\boldsymbol{\theta}) - m_2(\boldsymbol{\theta}))}{(m_1(\boldsymbol{\theta}) + m_2(\boldsymbol{\theta}))(N - m_1(\boldsymbol{\theta}) - m_2(\boldsymbol{\theta}))} (\nabla m_1(\boldsymbol{\theta}) + \nabla m_2(\boldsymbol{\theta})) \right). \end{aligned} \quad (8.7)$$

### 8.3 Fisher's exact test

Another widely used measure of association on  $2 \times 2$  contingency tables is the exact test of Fisher (1922). It is based on the tail probabilities of the hypergeometric distribution. The hypergeometric distribution quantifies the likelihood of drawing in  $d$  draws without replacement  $w$  white balls from an urn containing a total of  $D$  balls, of which  $W$  are white,

$$\mathbb{P}(d, D, w, W) = \frac{\binom{W}{w} \binom{D-W}{d-w}}{\binom{D}{d}}, \quad (8.8)$$

where  $\binom{a}{b} = \frac{a!}{b!(a-b)!}$  is a binomial coefficient. Fisher's exact test amounts to summing the probabilities of all contingency tables that are as extreme or more so than the observed contingency table. It thus gives the probability of observing a contingency table as extreme or more so than the observed one based on a null model of independence of rows and columns under fixed marginals.

For  $T_1$  and  $T_2$  Fisher's exact test gives an odds ratio of  $\infty$  and 23.17, respectively, which both correspond to  $p$ -values less than the smallest representable positive floating point number. When adding one pseudo-count before computing Fisher's exact test, the resulting odds ratios are 984.59 and 22.73, respectively. There exist generalizations of Fisher's

exact test to contingency tables larger than  $2 \times 2$  (Mehta and Patel, 1983).

## 8.4 Normalized enrichment score

The first  $k \times 2$  contingency table based measure that we want to discuss here are normalized enrichment scores. They are related to the difference of relative frequency discussed above, and are usable when the contrast provides one set of positive example sequences and multiple sets of control sequences. Normalized enrichment scores divide the difference of relative frequency in the signal data and the mean of relative frequencies in the control by a standard deviation computed from the relative frequencies in the control datasets.

We refer to the relative frequency of sequences with motif occurrences in condition  $i$  as  $f_i = \frac{m_i}{n_i}$ . We assume that condition 1 is the signal dataset, and conditions 2 to  $k$  are control datasets. Then the normalized enrichment score  $z$  is given by

$$z = \frac{f_1 - \mu}{\sigma}, \quad (8.9)$$

where  $\mu$  is the mean relative frequency of sequences with motif occurrences in the control data, and  $\sigma$  the standard deviation of relative frequencies in the control datasets.

## 8.5 Pearson's $\chi^2$ test for independence

Perhaps the most widely used association measure is the  $\chi^2$  test for independence (K. Pearson, 1900). Given a  $n \times k$  contingency table with counts  $O_{ij}$  in the cell in row  $i$  and column  $j$ , and defining the row sums  $R_i = \sum_{j=1}^k O_{ij}$  and column sums  $C_j = \sum_{i=1}^n O_{ij}$ , as well as the expected counts under the independence hypothesis  $E_{ij} = \frac{R_i C_j}{N}$ , where  $N = \sum_{i=1}^n R_i = \sum_{j=1}^k C_j$ , then the  $X^2$  statistic is given by

$$X^2 = \sum_{i=1}^n \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \quad (8.10)$$

This statistic  $X^2$  is asymptotically distributed like  $\chi^2$  with  $(n-1) \cdot (k-1)$  degrees of freedom, and  $p$ -values are available for it.

## 8.6 Mutual information of condition and occurrence

Assume to be given a contingency table of numbers of sequences with or without at least one occurrences of a motif  $M$  across the conditions of a contrast  $C$ . *Mutual information* of condition  $C$  and feature occurrence  $M$ , is an unidirectional measure of association from information theory (Cover and Thomas, 2006; MacKay, 2003; Shannon, 1948), see appendix C,

$$\mathbb{I}(C; M) = \sum_{\substack{c \in \mathcal{C} \\ m \in \mathcal{M}}} \mathbb{P}(C = c, M = m) \log_2 \frac{\mathbb{P}(C = c, M = m)}{\mathbb{P}(C = c)\mathbb{P}(M = m)}, \quad (8.11)$$

where  $\mathbb{P}(C = c, M = m)$ ,  $\mathbb{P}(C = c)$ , and  $\mathbb{P}(M = m)$  are joint and marginal relative frequencies of contingency tables like the one depicted in table 7.2, and  $\mathbb{P}(C = c) = \sum_{m \in \mathcal{M}} \mathbb{P}(C = c, M = m)$ , and  $\mathbb{P}(M = m) = \sum_{c \in \mathcal{C}} \mathbb{P}(C = c, M = m)$ . In terms of the variables of table 7.2 this is

$$\begin{aligned} \mathbb{I}(C; M) = \log_2 N + \frac{1}{N} \left( \sum_{i=1}^k m_i(\boldsymbol{\theta}) \log_2 \frac{m_i(\boldsymbol{\theta})}{n_i \sum_j m_j(\boldsymbol{\theta})} \right. \\ \left. + \sum_{i=1}^k (n_i - m_i(\boldsymbol{\theta})) \log_2 \frac{n_i - m_i(\boldsymbol{\theta})}{n_i \sum_j (n_j - m_j(\boldsymbol{\theta}))} \right), \end{aligned} \quad (8.12)$$

where  $N = \sum_{j=1}^k n_j$ .

Being the KL divergence of the joint probability model and the independent probability model,  $\mathbb{I}(C; M) = D_{\text{KL}}(\mathbb{P}(C, M) \| \mathbb{P}(C)\mathbb{P}(M))$  (see appendix C), mutual information measures discrimination, or association of condition and motif occurrence. It quantifies how much information is conveyed about the distinction between signal and control, or generally the condition, by knowing the motif to be present or absent. Symmetrically it also quantifies how much information is conveyed about the motif presence by the signal/control distinction. Mutual information is measured in units of bits per sequence.

We define mutual information of condition and motif occurrence (MICO) by scaling mutual information by the number of sequences  $N$ , to yield

$$\text{MICO} = N \cdot \mathbb{I}(C; M), \quad (8.13)$$

in units of bits, not bits per sequence.

**Gradient** As detailed in appendix F, from equation (8.12) we derive the following expression for the gradient of the mutual information of condition and motif presence,

$$\nabla \mathbb{I}(C; M) = \frac{1}{N} \left( \sum_i (\nabla m_i(\boldsymbol{\theta})) \log_2 \frac{m_i(\boldsymbol{\theta})}{n_i - m_i(\boldsymbol{\theta})} - \left( \sum_i \nabla m_i(\boldsymbol{\theta}) \right) \log_2 \frac{\sum_i m_i(\boldsymbol{\theta})}{\sum_i (n_i - m_i(\boldsymbol{\theta}))} \right). \quad (8.14)$$

**Analyzing multiple contrasts** Frequently it is useful to jointly analyze multiple contrasts, such as when repeat experiments are available. In this case—as MICO is an additive measure—the MICO values of the individual contrasts can simply be added together.



## Chapter 9

# Probabilistic discriminative objective functions

This chapter discusses three related discriminative objective functions that are not based on contingency tables but utilize probabilistic formulations for classification. Section 9.1 discusses an approach which uses one probabilistic model. An approach that uses multiple probabilistic models is presented in section 9.2. And a composite-model-based approach is outlined in section 9.3.

### 9.1 Difference of log likelihood

A directional, discriminative objective function applicable to binary contrasts that is not based on contingency tables is the difference of log likelihood (DLOGL) between signal and control,

$$\begin{aligned} \text{DLOGL} &= \log \mathbb{L}(\boldsymbol{\theta} | \mathbf{X}_{\text{signal}}) - \log \mathbb{L}(\boldsymbol{\theta} | \mathbf{X}_{\text{control}}) \\ &= \log \mathbb{P}(\mathbf{X}_{\text{signal}} | \boldsymbol{\theta}) - \log \mathbb{P}(\mathbf{X}_{\text{control}} | \boldsymbol{\theta}) \\ &= \sum_{i \in \text{signal}} \log \mathbb{P}(X_i | \boldsymbol{\theta}) - \sum_{i \in \text{control}} \log \mathbb{P}(X_i | \boldsymbol{\theta}), \end{aligned} \tag{9.1}$$

where  $\mathbf{X}_{\text{signal}}$  and  $\mathbf{X}_{\text{control}}$  are the signal and control data, respectively, and  $\boldsymbol{\theta}$  are the parameters of a probabilistic model. In words, this objective functions identifies models for which the signal data appear as typical examples but simultaneously the control data appear as unlikely examples. Thus, data yielding high likelihood for a model selected by DLOGL tends to indicate signal data.

The choice of DLOGL as objective function for learning necessitates balancing of the sizes of signal and control data used for learning. In case the control dataset is considerably larger, any gain in likelihood for the signal data may be outweighed by a loss of likelihood for the control data. Dominating control data sizes may thus lead to parameters that are determined primarily by being bad generative models for the control data. Conversely, in case the control dataset is considerably smaller, this objective function is dominated by the signal likelihood and loses its discriminative character as DLOGL becomes practically equivalent to ML learning.

**Gradient** The gradient of the DLOGL is simply the difference of log likelihood gradients,

$$\nabla \text{DLOGL} = \sum_{i \in \text{signal}} \nabla \log \mathbb{P}(\mathbf{X}_i | \boldsymbol{\theta}) - \sum_{i \in \text{control}} \nabla \log \mathbb{P}(\mathbf{X}_i | \boldsymbol{\theta}). \quad (9.2)$$

For the case of an HMM  $\boldsymbol{\theta}$ , expressions for the log likelihood gradient  $\nabla \log \mathbb{P}(\mathbf{X} | \boldsymbol{\theta})$  are given in section 4.6.

## 9.2 Multiple model classification

Instead of learning just one model  $\boldsymbol{\theta}$ , it is also possible to use two or more models  $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots$ , which are identified with the conditions of the contrast. We think of the conditions as classes  $C_1, C_2, \dots$ . The conditional probability of classifying a sequence  $\mathbf{X}$  as belonging to class  $C_i$  is then

$$\begin{aligned} \mathbb{P}(C = C_i | \mathbf{X}) &= \frac{\mathbb{P}(C = C_i, \mathbf{X})}{\mathbb{P}(\mathbf{X})} \\ &= \frac{\mathbb{P}(C = C_i, \mathbf{X})}{\sum_k \mathbb{P}(C = C_k, \mathbf{X})} \\ &= \frac{\mathbb{P}(\mathbf{X} | C = C_i) \mathbb{P}(C = C_i)}{\sum_k \mathbb{P}(\mathbf{X} | C = C_k) \mathbb{P}(C = C_k)}. \end{aligned} \quad (9.3)$$

Here,  $\mathbb{P}(C = C_i)$  is the prior for class  $C_i$ . The likelihood of the data  $\mathbf{X}$  given by class  $C_i$  is given by a class-dependent model,

$$\mathbb{P}(\mathbf{X} | C = C_i) = \mathbb{P}(\mathbf{X} | \boldsymbol{\theta}_i). \quad (9.4)$$

Possible choices for the class-dependent models are the OOPS or ZOOPS models for signal data, and a pure background model for control data in the case of a binary contrast. Another possibility is to use for all classes ZOOPS models for the same motif but with mixing parameter  $\lambda_i$  depending on the class  $C_i$ .

This objective function is not implemented in Discover.

## 9.3 Probability of correct classification

We now turn to consider a different probabilistic model of classes. For this we assume that the data are given in form of paired sets of sequences  $\mathbf{X} = (\mathbf{X}_i)$  with corresponding classes  $\mathbf{c} = (c_i)$ . Maximum mutual information estimation (MMIE) considers the *probability of correctly classifying all samples*,

$$\mathbb{P}(\mathbf{C} = \mathbf{c} | \mathbf{X}, \boldsymbol{\theta}) = \prod_i \mathbb{P}(C_i = c_i | \mathbf{X}_i, \boldsymbol{\theta}), \quad (9.5)$$

or its logarithm,

$$\log \mathbb{P}(\mathbf{C} = \mathbf{c} | \mathbf{X}, \boldsymbol{\theta}) = \sum_i \log \mathbb{P}(C_i = c_i | \mathbf{X}_i, \boldsymbol{\theta}). \quad (9.6)$$

It may appear tempting to identify the classes with motif presence or absence, but this turns out to be problematic when the data includes mislabeled samples, in particular false



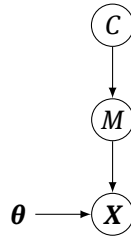


Figure 9.1: The graphical model of MMIE.  $C$  represents the class, or dataset,  $M$  the motif presence, and  $X$  the observed sequence. For an explanation of the graphical model notation see e.g. Koller and N. Friedman (2009).

positives. As both false positives and false negatives appear to be common in real biological data, it makes sense to consider alternatives. One possibility is to add a mixture model as follows.

The model comprises three random variables,  $X$  for the sequence, a binary variable  $M$  for the motif presence in a sequence, and a discrete variable  $C$  for the class of the sequence. Additionally, there are parameters  $\theta$  for a probabilistic (sub-) model that determines  $\mathbb{P}(X|\theta)$ . The structure of the model is as depicted in figure 9.1, which corresponds to the following factorization,

$$\mathbb{P}(C, M, X|\theta) = \mathbb{P}(X|M, \theta)\mathbb{P}(M|C)\mathbb{P}(C). \quad (9.7)$$

The conditional motif occurrence probabilities  $\mathbb{P}(M|C)$ , the class prior  $\mathbb{P}(C)$ , as well as the HMM parameters  $\theta$  are parameters of the MMIE model.  $\mathbb{P}(C)$  is a probability distribution over  $k = |C|$  classes, and thus represents  $k - 1$  free parameters.  $\mathbb{P}(M|C)$  is a table of conditional probabilities with  $k \times 2$  entries, and  $k$  free parameters. Given an expression for the likelihood  $\mathbb{P}(X|\theta)$  and for posterior probability of a feature occurrence  $\mathbb{P}(M|X, \theta)$ , we then have

$$\mathbb{P}(C, M, X|\theta) = \frac{\mathbb{P}(X, M|\theta)}{\mathbb{P}(M)} \mathbb{P}(M|C)\mathbb{P}(C) = \frac{\mathbb{P}(M|X, \theta)\mathbb{P}(X|\theta)}{\mathbb{P}(M)} \mathbb{P}(M|C)\mathbb{P}(C), \quad (9.8)$$

where  $\mathbb{P}(M) = \sum_{c \in C} \mathbb{P}(M, C = c) = \sum_{c \in C} \mathbb{P}(M|C = c)\mathbb{P}(C = c)$ . The likelihood of motif presence and class,  $\mathbb{P}(C, M|X, \theta)$ , is then given by

$$\mathbb{P}(C, M|X, \theta) = \frac{\mathbb{P}(C, M, X|\theta)}{\mathbb{P}(X|\theta)} = \frac{\mathbb{P}(M|X, \theta)}{\mathbb{P}(M)} \mathbb{P}(M|C)\mathbb{P}(C). \quad (9.9)$$

By summing over  $M \in \mathcal{M} = \{M, \neg M\}$  we can express the posterior probability of classify-

ing data  $\mathbf{X}$  as class  $C$  given the HMM parameters  $\boldsymbol{\theta}$ ,  $\mathbb{P}(C|\mathbf{X}, \boldsymbol{\theta})$ , as follows,

$$\mathbb{P}(C|\mathbf{X}, \boldsymbol{\theta}) = \mathbb{P}(C, M|\mathbf{X}, \boldsymbol{\theta}) + \mathbb{P}(C, \neg M|\mathbf{X}, \boldsymbol{\theta}) \quad (9.10)$$

$$= \mathbb{P}(C) \left( \frac{\mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta})}{\mathbb{P}(M)} \mathbb{P}(M|C) + \frac{\mathbb{P}(\neg M|\mathbf{X}, \boldsymbol{\theta})}{\mathbb{P}(\neg M)} \mathbb{P}(\neg M|C) \right) \quad (9.11)$$

$$= \mathbb{P}(C) \left( \frac{\mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta})}{\mathbb{P}(M)} \mathbb{P}(M|C) + \frac{1 - \mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta})}{1 - \mathbb{P}(M)} (1 - \mathbb{P}(M|C)) \right) \quad (9.12)$$

$$= \mathbb{P}(C) \left( \frac{\mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta})}{\mathbb{P}(M)} \mathbb{P}(M|C) - \frac{\mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta})}{1 - \mathbb{P}(M)} (1 - \mathbb{P}(M|C)) + \frac{1 - \mathbb{P}(M|C)}{1 - \mathbb{P}(M)} \right) \quad (9.13)$$

$$= \mathbb{P}(C) \left( \mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta}) \left( \frac{\mathbb{P}(M|C)}{\mathbb{P}(M)} - \frac{1 - \mathbb{P}(M|C)}{1 - \mathbb{P}(M)} \right) + \frac{1 - \mathbb{P}(M|C)}{1 - \mathbb{P}(M)} \right) \quad (9.14)$$

$$= \frac{\mathbb{P}(C)}{1 - \mathbb{P}(M)} \left( \mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta}) \left( \frac{\mathbb{P}(M|C)}{\mathbb{P}(M)} - 1 \right) + 1 - \mathbb{P}(M|C) \right). \quad (9.15)$$

In (9.11) we use (9.9). Step (9.12) uses the fact that  $M$  is a binary variable, and thus  $M$  and  $\neg M$  are complementary events, from which we have  $\mathbb{P}(M) = 1 - \mathbb{P}(\neg M)$ , and similarly  $\mathbb{P}(M|C) = 1 - \mathbb{P}(\neg M|C)$ , and  $\mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta}) = 1 - \mathbb{P}(\neg M|\mathbf{X}, \boldsymbol{\theta})$ . The steps (9.13), (9.14), and (9.15), are just rearranging and cancelling terms.

**Gradient** The MMIE learning routine of Discover uses gradient optimization for the HMM parameters  $\boldsymbol{\theta}$ . The other MMIE parameters, i.e. the class prior  $\mathbb{P}(C)$  and the conditional motif occurrence priors  $\mathbb{P}(M|C)$ , are simply re-estimated by Discover's MMIE learning routine. Thus, for MMIE, we only give the gradient of the classification probability  $\nabla \mathbb{P}(C|\mathbf{X}, \boldsymbol{\theta})$  with respect to the HMM parameters  $\boldsymbol{\theta}$ ,  $\nabla f(\boldsymbol{\theta}) = \left( \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_i} \right)_i$ ,

$$\nabla \mathbb{P}(C|\mathbf{X}, \boldsymbol{\theta}) = \frac{\mathbb{P}(C)}{1 - \mathbb{P}(M)} \left( \frac{\mathbb{P}(M|C)}{\mathbb{P}(M)} - 1 \right) \nabla \mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta}). \quad (9.16)$$

As the global MMIE objective (9.6) is given by the sum of log probabilities of correct classification of the individual sequences, we finally consider the gradient of the log probability of correctly classifying a sequence  $\mathbf{X}$  as belonging to class  $C$  given the model  $\boldsymbol{\theta}$ ,

$$\nabla \log \mathbb{P}(C|\mathbf{X}, \boldsymbol{\theta}) = \frac{\nabla \mathbb{P}(C|\mathbf{X}, \boldsymbol{\theta})}{\mathbb{P}(C|\mathbf{X}, \boldsymbol{\theta})} = \frac{\mathbb{P}(C)}{1 - \mathbb{P}(M)} \left( \frac{\mathbb{P}(M|C)}{\mathbb{P}(M)} - 1 \right) \frac{\nabla \mathbb{P}(M|\mathbf{X}, \boldsymbol{\theta})}{\mathbb{P}(C|\mathbf{X}, \boldsymbol{\theta})}. \quad (9.17)$$

## Chapter 10

# Significance of association

In this work we make use of the following connection between mutual information and the likelihood ratio test which allows to compute  $p$ -values for mutual information in contingency tables. This connection is explained in section 10.1. Section 10.2 details how  $p$ -values of association may be corrected for multiple testing (MT) to control the family-wise error rate (FWER). Finally, section 10.3 describes how MT-corrected  $p$ -values that do not meet significance criteria are filtered to increase MD performance.

### 10.1 Mutual information, likelihood ratio, and $\chi^2$ test

The value of mutual information  $I$  is related to the log likelihood ratio  $\log \Lambda$  of the hypothesis that the counts in rows and columns of a contingency table are distributed independently by

$$\log \Lambda = -I \cdot n \cdot \log 2, \quad (10.1)$$

where  $n$  is the total number of cases in the table. Wilks' theorem (Wilks, 1938) relates the log likelihood ratio to the  $\chi^2$  test. Specifically, for  $k \times 2$  contingency tables and for increasing sample sizes,  $-2 \log \Lambda$  is asymptotically distributed like  $\chi^2$  with  $k - 1$  degrees of freedom. Also see appendix D.

In MD frequently the problem arises to compare the discriminative performance of models with differing numbers of parameters. If two models are optimal for their respective motif spaces, and when additionally the larger of the two motif spaces comprises the smaller one, then discriminability must be greater or equal for the motif which is optimal over the larger motif space. Through the connection of mutual information to the  $\chi^2$  test via the likelihood ratio test we may determine  $p$ -values in both cases. By correcting  $p$ -values for motif space size, we propose to make comparable the discriminability of motifs with different numbers of parameters. To this end we correct  $p$ -values in a Bonferroni-style by multiplying with the motif space size, to control the FWER (Hochberg and Tamhane, 1987). This counteracts usage of overly long or degenerate motifs, for which the search space is large, by favoring short words or words with low degree of degeneracy, with a correspondingly smaller search space. Note that control of the FWER may be conservative (Hochberg, 1988), and that more powerful MT-correction methods are available,

e.g. those of Benjamini and Hochberg (1995), Benjamini and Yekutieli (2005), Hochberg and Benjamini (1990), and Storey and Tibshirani (2003).

We will next propose ways of calculating motif space sizes for the discrete case of IUPAC regexes and for the continuous case applicable to HMMs.

## 10.2 Multiple testing correction for motif discovery problems

Any MD method aims to find optimal sets of parameters according to some objective function. During this (exact, approximate, or heuristic) optimization many parameter values are tested and the best one is reported. Clearly, the larger the allowed motif space the higher the maximally achievable objective function is. It is thus desirable to account for the difference in number of parameters when comparing the values of the objective function on the same data for two parameter sets with differing numbers of parameters. There is no generally applicable way to account for difference in number of parameters for arbitrary objective functions. However, whenever the objective function represents a  $p$ -value  $P$ , we may, in a Bonferroni style, multiply the  $p$ -value with the size  $N$  of the motif space, to yield a corrected  $p$ -value  $P_{\text{corrected}}$ . In log-space we have then

$$\log P_{\text{corrected}} = \min(0, \log P + \log N). \quad (10.2)$$

Below, we give expressions for the motif space size for discrete, string-based motif representations, as well as a proposition for continuous matrix-based motif representations.

### 10.2.1 Discrete motif space sizes

We first consider the case when the motif space over which the optimization takes place is that of strings of length  $n$  over the nucleic acid alphabet. In this case the motif space has a size  $N = 4^n$ . Similarly, the motif space of all IUPAC regexes of length  $n$  is of size  $N = 15^n$ .

When considering IUPAC regexes up to some maximal degeneracy the motif space size is less than given above. In this case the computation of the motif space size is easily done by dynamic programming in the length of the motif and the degree of degeneracy.

For this define the matrix  $M$  with entries for each length  $n$  and degree of degeneracy  $d$ , whose entries are given by the following recursive equations,

$$M(n, d) = \begin{cases} 0 & \text{if } 3n < d \\ 1 & \text{if } n = 0 \text{ and } d = 0 \\ 4 \cdot M(n-1, d) & \text{if } n > 0 \text{ and } d = 0 \\ 4 \cdot M(n-1, d) + 6 \cdot M(n-1, d-1) & \text{if } n > 0 \text{ and } d = 1 \\ 4 \cdot M(n-1, d) + 6 \cdot M(n-1, d-1) + 4 \cdot M(n-1, d-2) & \text{if } n > 0 \text{ and } d = 2 \\ 4 \cdot M(n-1, d) + 6 \cdot M(n-1, d-1) + 4 \cdot M(n-1, d-2) \\ \quad + 1 \cdot M(n-1, d-3) & \text{if } 3 \leq d \leq 3n \end{cases} \quad (10.3)$$

Note that the maximal degeneracy of a motif of length  $n$  is  $3n$ , giving the first recursion anchor. Then, there is exactly one motif of length 0, the empty motif. Next, when considering motifs of length  $n$  with degeneracy zero, there are four times as many as motifs of length  $n-1$  with degeneracy zero, as each of these may be extended by appending one

of A, C, G, or T. Similarly, when considering motifs of length  $n$  with degree of degeneracy  $d = 1$ , in addition to extending motifs of length  $n - 1$  and  $d = 1$ , one may extend motifs of length  $n - 1$  and  $d = 0$  with any of the six singly-degenerate symbols S, W, M, K, R, and Y. Motifs of length  $n$  and degree of degeneracy  $d = 2$  additionally may extend motifs of length  $n - 1$  and  $d = 1$  with the four two-fold degenerate symbols B, D, H, and V. Finally, for motifs of length  $n$  and degree of degeneracy  $d \geq 3$ , the motifs of length  $n - 1$  and  $d = 2$  may be extended with the single three-fold degenerate symbol N.

From the matrix  $M$  we can then easily compute the number  $N(n, d)$  of IUPAC motifs of size  $n$  with degeneracy up to  $d$  as

$$N(n, d) = \sum_{k=0}^d M(n, k). \quad (10.4)$$

### 10.2.2 Continuous motif space sizes

For general matrix-based motif representations there are infinitely many values due to infinitely many distributions that exist already for individual positions. Some methods restrict the continuous matrix space to a lattice, and the number of lattice vertices naturally defines the size of the motif space in this case.

Here we propose to base the effective number of parameter in a continuous matrix on rank statistics. Considering individual positions, we assume that one to four nucleotide may be allowed. If one nucleotide is allowed, there are four possibilities of choosing this one nucleotide. When two nucleotides are allowed, e.g. nucleotides  $x$  and  $y$  then we may have  $x < y$ ,  $x = y$ ,  $x > y$ , indicating that nucleotide  $x$  is respectively less frequent, as frequent, or more frequent than nucleotide  $y$ . Thus, for two nucleotides, one can choose two of the four nucleotides and can have the two nucleotides in three relations, resulting in  $\binom{4}{2} \cdot 3 = 18$  possibilities. Following this logic, we have the following formula for the total number of rankings of up to 4 elements selected from the nucleic acid alphabet,

$$J = \sum_{i=1}^4 \binom{4}{i} \cdot K(i) = 4 \cdot 1 + 6 \cdot 3 + 4 \cdot 13 + 1 \cdot 75 = 149, \quad (10.5)$$

where  $K(i)$  is the number of total preorders of  $i$  elements (). Then, by multiplying this number across the positions, we have a motif space size of  $N = J^n$ .

## 10.3 Significance of association significance filtering

Rejecting models whose MT-corrected  $p$ -values are not significant reduces the number of falsely predicted models. Thus, Discover accepts or rejects the final, optimized parameterization, depending on whether the MT-corrected MICO-based  $p$ -value meets a given threshold. This discriminative significance filtering based on MICO is applied regardless of the objective function chosen for seeding and optimization.

Discover uses the motif space calculation as described in section 10.2.2. Regarding significance filtering in Plasma we experimented with the motif space size definition given

in section 10.2.1. While we generally found it to be useful (results not shown), we ultimately decided for the sake of comparability with Discover to also base significance calculation for Plasma on the continuous motif space size definition given in section 10.2.2.

## Chapter 11

# Discrete optimization of discriminative objectives

This chapter describes two methods for optimizing discriminative objectives over discrete spaces. Section 11.1 describes a progressive algorithm, named `corenmers`, that enumerates words over an alphabet in decreasing order of residual contribution to discrimination. The method that is used to determine seeds for subsequent HMM optimization is described in section 11.2.

Both methods are based on sets of sequences, and may utilize the non-probabilistic, table-based discriminative objective functions DFREQ, MCC, and MICO described in chapter 8. Naturally, when more than two sets of sequences are given, then the binary discriminative objective functions DFREQ and MCC are not applicable, and only MICO may be used.

### 11.1 Enumerating residually most discriminative words

The following method is implemented in the program `corenmers` that is part of the `Dis-crover` software package released with this thesis.

As input `corenmers` expects a set of sequence sets. The user may specify one length  $n$  or a length range  $n_1$  to  $n_2$ . The method then enumerates all  $n$ -mers of the desired lengths occurring in the sequences, and prints out the top scoring  $k$  words or all words in the sequences, in order of decreasing objective function.

Algorithm 4 gives pseudocode for `corenmers`.

The algorithm determines the  $k$  most relevant words in the desired length range on the data  $\mathbf{X}$  according to some objective function  $f$ . This is done by progressively identifying the most relevant word, and masking its occurrences in the data, before identifying further words. The algorithm is relatively simplistic and just considers all words occurring in the sequences, without allowing degenerate positions. This allows a runtime of  $\mathcal{O}(kdn)$  to determine the  $k$  most-relevant words, where  $d = n_2 - n_1 + 1$  and  $n$  is the total length of the sequences. Masking may either mask out the exact occurrence of the word, or discard sequences with occurrences of the word. Optionally, the algorithm also takes into account

**Algorithm 4** Core-mer analysis

**Input:** word length range  $n_1$  to  $n_2$ , number of words to determine  $k$ , data  $\mathbf{X}$ , objective function  $f$ , alphabet  $\mathcal{A}$

**Output:**  $(w_i)_{i=1,\dots,n}$  the  $k$  most relevant words of length  $n_1$  to  $n_2$

```

1: for  $i = 1 \rightarrow k$  do
2:    $w_i \leftarrow \operatorname{argmax}_{w \in \mathcal{A}^n, n_1 \leq n \leq n_2} f(w, \mathbf{X})$ 
3:    $\mathbf{X} \leftarrow \operatorname{mask}(\mathbf{X}, w_i)$ 
4: return  $(w_i)_{i=1,\dots,n}$ 

```

words on the reverse complementary strand to allow analysis of DBPs.

As mentioned in the introductory paragraph to this chapter, the objective functions must be based on discrete counts, and we typically employ MICO. In addition to the above-mentioned discriminative objective functions, also the MT-corrected  $p$ -value of association based on MICO (see chapter 10), and relative occurrence frequency in all sequences may be used as objective function.

As the objective function for a word may change after masking of occurrences of an overlapping word or due to co-occurring with previously identified words, we refer to the objective function value at which a word is selected as its residual objective value. The residual objective value is thus an estimate of the independent contribution a word conveys for discriminating sets of sequences after more relevant words have been accounted for.

## 11.2 Identifying discriminative words with degeneracy

The following method is implemented in the program Plasma that is part of the Discover software package released with this thesis. It is the routine that is used to find seeds for subsequent optimization of HMM parameters. To allow both manual experimentation and automation, Plasma is integrated into the HMM preprocessing, but can also be run separately.

The seed finding routine heuristically identifies motifs in the form of IUPAC regexes that score high according to the chosen objective function. The seed finding procedure presented here is similar to DREME (Bailey, 2011), but offers multiple, contingency table based objective functions to choose from, and uses a different heuristic to filter unpromising candidate motifs. Input to seed finding consists of sets of sequences among which discriminative motifs are suspected. Parameters include the choice of objective function and motif lengths to consider. The objective functions comprise relative occurrence frequency, DFREQ, MCC, and MICO. Optionally, discriminative motifs may be filtered for enrichment in specific samples.



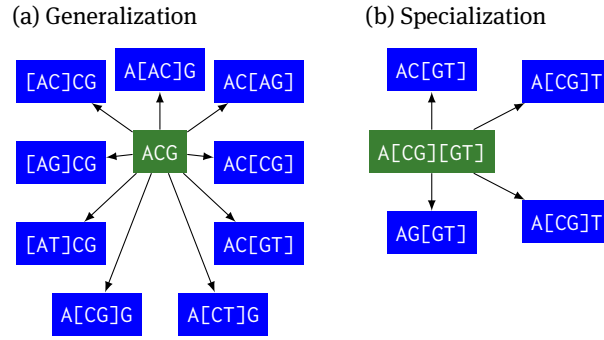


Figure 11.1: IUPAC generalizations of ACG (a) and specializations of A[CG][GT]=ASY (b).

---

**Algorithm 5** Most discriminative IUPAC motif

**Input:** word length  $k$ , maximal degeneracy  $D$ , data  $\mathbf{X}$ , objective function  $f$ , number of candidates to maintain  $n$

**Output:** Heuristically most discriminative IUPAC word  $w$  of maximal degeneracy  $D$

- 1:  $d \leftarrow 0$
  - 2:  $C \leftarrow \text{kmers}(\mathbf{X}, k)$
  - 3:  $C \leftarrow \text{top}_n(c \in C, f(c, \mathbf{X}))$
  - 4: **while**  $C \neq \emptyset$  **and**  $d \leq D$  **do**
  - 5:    $G \leftarrow \{g \in \text{generalizations}(c) \mid c \in C\}$
  - 6:    $C \leftarrow \{g \in G \mid f(g, \mathbf{X}) \geq \max_{s \in C \cap \text{specializations}(g)} f(s, \mathbf{X})\}$
  - 7:    $C \leftarrow \text{top}_n(c \in C, f(c, \mathbf{X}))$
  - 8:    $d \leftarrow d + 1$
  - 9: **return**  $\text{top}_1(c \in C, f(c, \mathbf{X}))$
- 

**Algorithm** Algorithm 5 gives pseudocode for Plasma. Initially, the algorithm considers all words of a given length occurring in the set of sequences, and for each occurring word  $w$  the number of sequences that contain  $w$  is determined for each of the sets of sequences. From these counts, and the number of sequences in each set, the objective function is evaluated for each word, and the words are sorted according to it. Then, only the top  $n$  words are retained as candidates, where  $n$  is a parameter whose default value is 100.

Each retained word  $w$  is generalized (line 5) by generating all IUPAC generalization of  $w$  that differ from  $w$  by allowing one additional nucleotide at any position. E.g. the word ACG may be generalized to [AC]CG, [AG]CG, [AT]CG, A[AC]G, A[CG]G, A[CT]G, AC[AG], AC[CG], AC[GT], see figure 11.1(a). By scanning over the sets of sequences, occurrence statistics are procured for each of the generalizations. Subsequently, the objective function is computed from the statistics of the generalizations. Generalizations with a score less than any of their generating specializations (see figure 11.1(b)) are dropped (line 6), and the resulting top  $n$  generalizations are kept (line 7) for further rounds of adding degeneracy, scoring, and retaining the top.

This scheme is iteratively continued until no further generalizations are available either because all have been dropped, or because the maximal degeneracy was reached. The user may limit the maximally allowed degeneracy with an absolute or relative limit<sup>1</sup>.

<sup>1</sup>For words of length  $k$  the maximal degeneracy is  $3k$ ; a relative limit  $g$  on degeneracy would allow up to  $\lfloor 3gk \rfloor$  degrees of degeneracy. For example, a relative degeneracy of  $g = 0.2$  for a motif of length 8 allows up to  $\lfloor 3 \cdot 0.2 \cdot 8 \rfloor = 4$  degrees of degeneracy.

Optionally, the algorithm also takes into account word occurrences on the reverse complementary strand to allow analysis of DBPs.

**Multiple seeds** When multiple seeds are desired, the most relevant one is identified according to algorithm 5. Subsequently, all occurrences of this motif are masked from the sequences, and further seeds may be sought. Alternatively, instead of masking just occurrences, the sequences containing occurrences may be discarded for the identification of further seeds.

**Runtime complexity and implementation details** Plasma uses a suffix array based word counting procedure, that finds occurrences of words without degenerate positions in time linear in the number of occurrences. For each occurrence, it is then necessary to determine the sequence that the position is contained in. For each position, this is done in constant time by means of a table that stores the index of the sequence that each position occurs in. The sequence indices are then reduced to unique sequence indices, incurring a runtime cost of  $\mathcal{O}(n \log n)$  due to sorting<sup>2</sup>, where  $n$  is the number of involved sequences. Then, for each sequence, the index of the set of sequences that the sequence is part of is determined in constant time by look-up in a sequence index to set index table.

However, for words that include positions where multiple symbols are allowed look up time is larger than for non-degenerate words. The runtime increase for words with degenerate positions depends both on the data and on where in the word degeneracy is allowed. In view of this, we do not give a formulation for the theoretical runtime cost of the algorithm used, but refer the reader to the example execution wall clock time given in the next paragraph.

**Example execution** Figure 11.2 lists the output of running Plasma on the cross-linking centered regions of PUM2 data of Hafner et al. (2010), analyzed in chapter 18. As is visible, execution of Plasma took under 5 seconds, demonstrating feasibility of optimizing a discriminative objective over IUPAC regexes on regular workstations.

---

<sup>2</sup>This cost could be reduced by using a hash-based routine to determine unique sequence indices. However, sorting the sequences indices is not the main contribution to runtime.

```
$ plasma PUM:pum.ccr.signal.fa control:pum.ccr.control.fa -m PUM:8
IUPAC representation      tgtanhwh
Regular expression        tgta.[act][at][act]
Length                    8
Information content [bit]  9.83007
Information content [bit / pos] 1.22876
Degeneracy                8
Objective:                PUM:mutual information
Score                     0.134184
Delta frequency           0.413516
Matthew's correlation coefficient 0.423151
Mutual information [bit]  0.13426
Mutual information (pscmt) [bit] 0.134184
Expected mutual information [bit] 0.134248
Variance mutual information [bit] 1.49677e-05
Sd mutual information [bit] 0.00386882
Z mutual information [bit] 34.7001
G-test                   2913.06
Uncorrected log-P        -1460.74
Corrected log-P          -1439.95
Occurrence statistics    pum.ccr.signal.fa 4702 / 7828 = 0.600664
Occurrence statistics    pum.ccr.control.fa 1465 / 7828 = 0.187149
User time = 4.596 sec
System time = 0.026 sec
CPU time = 4.622 sec
Elapsed time = 0.873674 sec
529.03% CPU
```

Figure 11.2: Example run of Plasma on sequences of PAR-CLIP cross-linking centered regions of PUM2 data by Hafner et al. (2010) optimizing MICO over the space of IUPAC regexes. The dataset comprises 7828 signal sequences, and equally many control sequences. The command line options instruct Plasma to find a motif of length 8 that is discriminative for the two specified FASTA files. The command was run on a workstation with an Intel® Core™i7-4770K CPU @ 3.50GHz, which has 8 cores.



## Chapter 12

# Hybrid learning

### 12.1 Signal and context parameters

Binding site HMMs are composite models of the cognate motifs as well as the surrounding sequence context. Some parameters of binding site HMMs, in particular the emission probabilities of the motif chain states, pertain to signal features, and we refer to these as *signal parameters*. The other parameters are referred to as *context parameters*, and comprise the emission probabilities of the background and all transition probabilities, including the prior occurrence probabilities of the motifs, realized as transition probabilities from other states to the beginning of the respective chain of motif states.

We assume that only the presence of signal features differs between signal and control sequences, while the surrounding sequence context is shared. Thus we propose to employ discriminative learning principles to learn signal parameters by contrasting signal and control sequences. However, to leverage HMM learning methods, we must specify a complete set of HMM parameters, including the context parameters. When context parameters are learned uninformed of signal features, they may erroneously incorporate properties of the signal features. There is thus a mutual dependence of the learning problems of signal and context parameters. In order to resolve it we propose the following procedure.

### 12.2 Learning scheme

We associate objective functions to the signal and context parameters. The signal parameters will use a discriminative objective, and the context parameters a generative one. We then employ a hybrid learning scheme which aims to jointly optimize both objectives over their respective parameters. The scheme consists of alternately updating the signal and context parameter classes, using suitable updating methods for their respective objective function. Hybrid learning is finished when termination criteria of both updating procedures are simultaneously fulfilled.

The natural choice for the generative objective function is the likelihood. Thus, updates for the context parameters are performed using iterations of the Baum-Welch algo-

rithm. Signal parameters may be optimized for any of the implemented discriminative objective functions by performing iterations of gradient search.

**Choice of learning scheme and alternatives** This hybrid learning scheme in which only the motif emissions are optimized by discriminative objectives, and all other parameters are optimized by the Baum-Welch algorithm is used by default in Discover. It is not guaranteed to terminate in the general case of arbitrary data and arbitrary choices of generative and discriminative objective functions. Yet, in our experience such problems are rare. In any case, to practically address the termination problem, the user may specify a maximal number of iterations to perform.

Aside from the hybrid learning scheme, Discover also allows the user to train all parameters by one objective function, or only the motif emissions and leave other parameters unmodified. Should the hybrid learning scheme fail for some data, these alternative, single-objective learning schemes are expected to optimize more robustly.

**Sequence sets for learning context parameters** By definition, the occurrence frequency of discriminative motifs differs between sets of sequences of suitable contrasts. Thus it matters which set of sequences the occurrence prior is learned from. By default all sequence sets are used to train the context parameters, but the user may specify a subset of the sequence sets to train the context parameters on. E.g., for contrasts of signal and scrambled sequences context parameters may be learned from the signal data only.

**Learning MMIE parameters** Differently from the other objectives, the optimization of the MMIE objective requires learning of class priors  $P(C)$  and conditional motif occurrence probabilities in the classes  $P(M|C)$  in addition to the HMM parameters. It has been suggested to optimize these separately from the other parameters (Krogh, 1994). Accordingly, Discover separately re-estimates these class and conditional occurrence probabilities during each iteration of learning, after the HMM parameters have been updated.

## 12.3 Multi-objective learning

Next, we will formulate the above-described procedure more abstractly to arrive at a more general understanding of its properties.

The method proposed here partitions the parameters to be estimated in  $n$  classes,  $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i \in \mathbb{Z}_n}$ . With each class  $i$  is associated an objective function  $f_i(\mathbf{X}, \boldsymbol{\theta})$ . Note that the objective function of each class may depend on the values of parameters in all classes. For notational simplicity in what follows we will not make explicit the dependence on the data  $\mathbf{X}$ ,  $f_i(\mathbf{X}, \boldsymbol{\theta}) = f_i(\boldsymbol{\theta})$ . Denote by  $\boldsymbol{\theta}_{-i}$  all parameters that are not in class  $i$ , i.e.  $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_j)_{j \neq i}$  so that each class  $i$  induces a partition of  $\boldsymbol{\theta}$  as  $\boldsymbol{\theta} = \boldsymbol{\theta}_i \times \boldsymbol{\theta}_{-i}$ .

The proposed optimization problem is that of determining parameters  $\boldsymbol{\theta}^*$  such that it holds simultaneously for all classes  $i$

$$\boldsymbol{\theta}_i^* = \operatorname{argmax}_{\boldsymbol{\theta}_i} f_i(\boldsymbol{\theta}_i \times \boldsymbol{\theta}_{-i}^*). \quad (12.1)$$

In words this criterion states that the parameters in each class  $i$  should be chosen so as to maximize the associated objective functions over all assignments of values to parameters in class  $i$  keeping the other parameters constant.

Clearly, such parameter values are not unique, and different parameter assignments may fulfill this criterion. Also, the achieved maximum of an objective function  $f_i$  may only be a local one. As an illustration of this point consider that in the general case each objective function  $i$  is assumed to depend on all parameters  $\boldsymbol{\theta}$  (and the data  $\mathbf{X}$ ), not just on those in class  $i$ . Thus choosing some values for the complement of class  $i$ , i.e. for  $\boldsymbol{\theta}_{-i}$ , the achievable value of the constrained optimization over the parameters of class  $i$  is less or equal to the achievable value of global optimization over all parameters,  $\max_{\boldsymbol{\theta}_i} f_i(\boldsymbol{\theta}_i \times \boldsymbol{\theta}_{-i}) \leq \max_{\boldsymbol{\theta}} f_i(\boldsymbol{\theta})$ .

While this criterion neither identifies unique parameters nor guarantees each objective to be globally optimized, it does guarantee local optimality in the sense that by modifying the parameters in class  $i$  leaving other parameters fixed no better value of the objective function  $f_i$  may be achieved. The criterion therefore is helpful in identifying meaningful parameters in case multi objectives are to be optimized and if it is clear which parameters should tune which objective.

The derivative of a function at a maximum vanishes. The converse is not true however, as also minima and saddle points of a function have vanishing derivative. Thus, for any parameter  $\boldsymbol{\theta}^*$  that is valid according to the above criterion, we simultaneously have for each class  $i$

$$\left. \frac{\partial f_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} \right|_{\boldsymbol{\theta}^*} = 0. \quad (12.2)$$

Using gradient optimization or re-estimation-based methods it is possible to perform a modification to the parameters of each class  $i$  that increases the value of objective function  $f_i$ . Thus, a simple procedure to find parameters that fulfill the above requirement is to alternately update the parameters of each class until convergence is reached when no update to any class changes the parameters more than some limit.

Unfortunately, convergence properties of such a procedure are quite unclear in the general case.





## Chapter 13

# Discovering multiple motifs

### 13.1 Measures of conditional association

Discover makes use of conditional mutual information (cMI) (Cover and Thomas, 2006), see equation (C.13) in appendix C.2.6. In particular, it is used to define cMI of conditions of a contrast  $\mathcal{C}$  and occurrence of motif  $\mathcal{A}$  given occurrences of motif  $\mathcal{B}$  (cMICO),

$$\text{cMICO}(\mathcal{C}; \mathcal{A}|\mathcal{B}) = \mathbb{I}(\mathcal{C}; \mathcal{A}|\mathcal{B}), \quad (13.1)$$

as well as to define motif pair cMI of occurrences of two motifs  $\mathcal{A}$  and  $\mathcal{B}$  given conditions of a contrast  $\mathcal{C}$ ,

$$\text{motif pair cMI}(\mathcal{A}; \mathcal{B}|\mathcal{C}) = \mathbb{I}(\mathcal{A}; \mathcal{B}|\mathcal{C}). \quad (13.2)$$

cMICO of a contrast  $\mathcal{C}$  and a motif  $\mathcal{A}$  given a motif  $\mathcal{B}$  measures the discriminatory contribution of  $\mathcal{A}$  across the contrast  $\mathcal{C}$  after accounting for the discriminatory contribution of  $\mathcal{B}$ . Motif pair cMI of  $\mathcal{A}$  and  $\mathcal{B}$  across  $\mathcal{C}$  quantifies how strongly occurrences of  $\mathcal{A}$  and  $\mathcal{B}$  are associated throughout the contrast.

#### 13.1.1 Motif pair MI and motif pair cMI

Our usage of cMICO and motif pair cMI for filtering (see section 13.2) is motivated by FIRE (Elemento, Slonim, and Tavazoie, 2007). Unlike our criterion, however, FIRE uses the non-conditional motif pair mutual information (MI) in place of the motif pair cMI. In our opinion motif pair cMI improves over motif pair MI, as illustrated by the two cases in figure 13.1 and figure 13.2. In the first case, two motifs that independently occur within each condition are found as associated by MI, but not by cMI. Conversely, in the second case two motifs that are dependently occurring within each condition are only found as associated according to cMI, but not according to MI.

In other words, usage of (non-conditional) motif-pair MI may lead to the conclusion that independently occurring motifs are occurring dependently, and conversely that dependent motifs are occurring independently, while cMI does not have this problem. These cases are of course instances of Simpson's paradox (Simpson, 1951).

	(a) Condition 1		(b) Condition 2		(c) Marginal (1+2)			
	A	¬A		A	¬A		A	¬A
B	1	9	B	81	9	B	82	18
¬B	9	81	¬B	9	1	¬B	18	82

Figure 13.1: Two motifs occur independently in condition 1, and independently in condition 2, but their marginal distribution appears dependent. In this case motif pair cMI yields 0.44 bit, while motif pair MI yields 61 bit (calculations done after adding a pseudo-count of 1).

	(a) Condition 1		(b) Condition 2		(c) Marginal (1+2)			
	A	¬A		A	¬A		A	¬A
B	40	0	B	0	60	B	40	60
¬B	0	60	¬B	40	0	¬B	40	60

Figure 13.2: Two motifs are dependently occurring in condition 1, and dependently in condition 2, but their marginal distribution appears independent. In this case motif pair cMI yields 167 bit, while motif pair MI yields 0 bit (calculations done after adding a pseudo-count of 1).

## 13.2 Discovering multiple motifs

Figure 13.3 illustrates the first part of the multiple motif discovery mode of Discover. First, seeds are discovered using Plasma. For each seed an HMM is initialized and independently optimized by Discover. The HMM achieving the best MICO based  $p$ -value is accepted. In a second part further motifs are then added to this HMM as described below and illustrated in figure 13.4.

In turn, the single HMM motifs are added to the accepted HMM, forming candidate HMMs. The candidate HMMs are then filtered, ensuring that newly added motifs provide sufficient additional discrimination and are not redundant with previously accepted motifs. This is done by comparing in each candidate HMM the new motif first pairwise against each previously accepted motif, and then jointly against all previously accepted motifs. Candidate HMMs and corresponding single-motif HMMs are discarded when the filtering criteria—outlined below—are not met for the newly added motif, whether in any of the pairwise comparisons or in the joint comparison.

Filtering is based on cMI, calculated in two ways: (I) cMI of conditions of the contrast and occurrences of the newly added motif given occurrences of previously accepted motifs (cMICO), and (II) cMI between occurrences of newly added and previously accepted motifs given the conditions of the contrast (motif pair cMI). cMICO quantifies the discriminatory contribution of the new motif after accounting for previous ones, while motif pair cMI quantifies association between occurrences of the newly added and previously accepted motifs. See section 13.1 for definitions of cMICO and motif pair cMI. In order to concentrate on motifs with a large residual explanatory contribution relative to their association with previous motifs, HMMs are discarded if at least one of two criteria is fulfilled: (a) the ratio

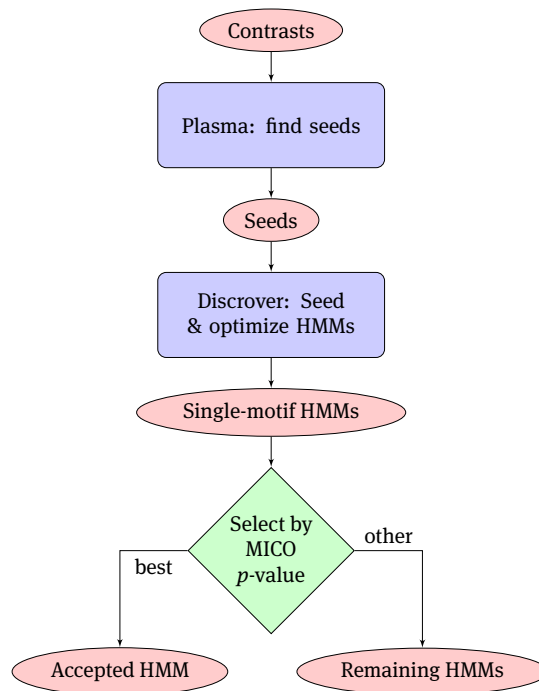


Figure 13.3: Flow chart of the first part of multiple motif discovery. The resulting accepted HMM and the set of remaining single-motif HMMs are the inputs for the second part, depicted in figure 13.4. See text for description.

of cMICO over motif pair cMI does not meet a threshold<sup>1</sup>, or (b) the cMICO based  $p$ -value is not significant.

As mentioned above, these criteria are first checked pairwise for the newly added motif and each previously accepted motif, and subsequently for the new motif and jointly all previously accepted motifs. In the joint comparison, an occurrence for the previously accepted motif is counted whenever any of the previously accepted motifs occurs.

Among the candidate HMMs that pass the filtering steps, we select the one whose newly added motif achieves the best cMICO based  $p$ -value. This HMM is then re-trained to optimize MICO for the feature of sequences having at least one occurrence of any of its motifs. If, after retraining, the MICO based  $p$ -value improves over the previously accepted one's, it is accepted, and further motifs may be added. Otherwise, or if all candidate motifs have been discarded, the last accepted HMM is reported.

### 13.2.1 Note on filtering

Note that usage of the ratio of cMICO over motif pair cMI is quite related to the usage of cMICO over motif pair MI as is done by FIRE (Elemento, Slonim, and Tavazoie, 2007). However, as already mentioned in section 13.1.1, unlike FIRE we use the motif pair cMI instead of motif pair MI. The intention behind this choice is to avoid pitfalls as illustrated

<sup>1</sup>We use the same threshold value of 5.0 as FIRE (Elemento, Slonim, and Tavazoie, 2007). As noted by Elemento et al., the user may want to experiment with this value, as it serves as a redundancy trade-off parameter. High values yield fewer motifs, low values yield more redundant motifs.

by the two cases in figure 13.1 and figure 13.2. While the illustrated cases may be said to be extreme, we found the underlying issue to be real. As motif pair cMI is used in the denominator of the ratio, the filtering step is rather sensitive to misjudgement of motif pair association. It is, in our view, therefore crucial to avoid quantitatively misjudging motif pair association as may frequently happen when using non-conditional motif pair MI.

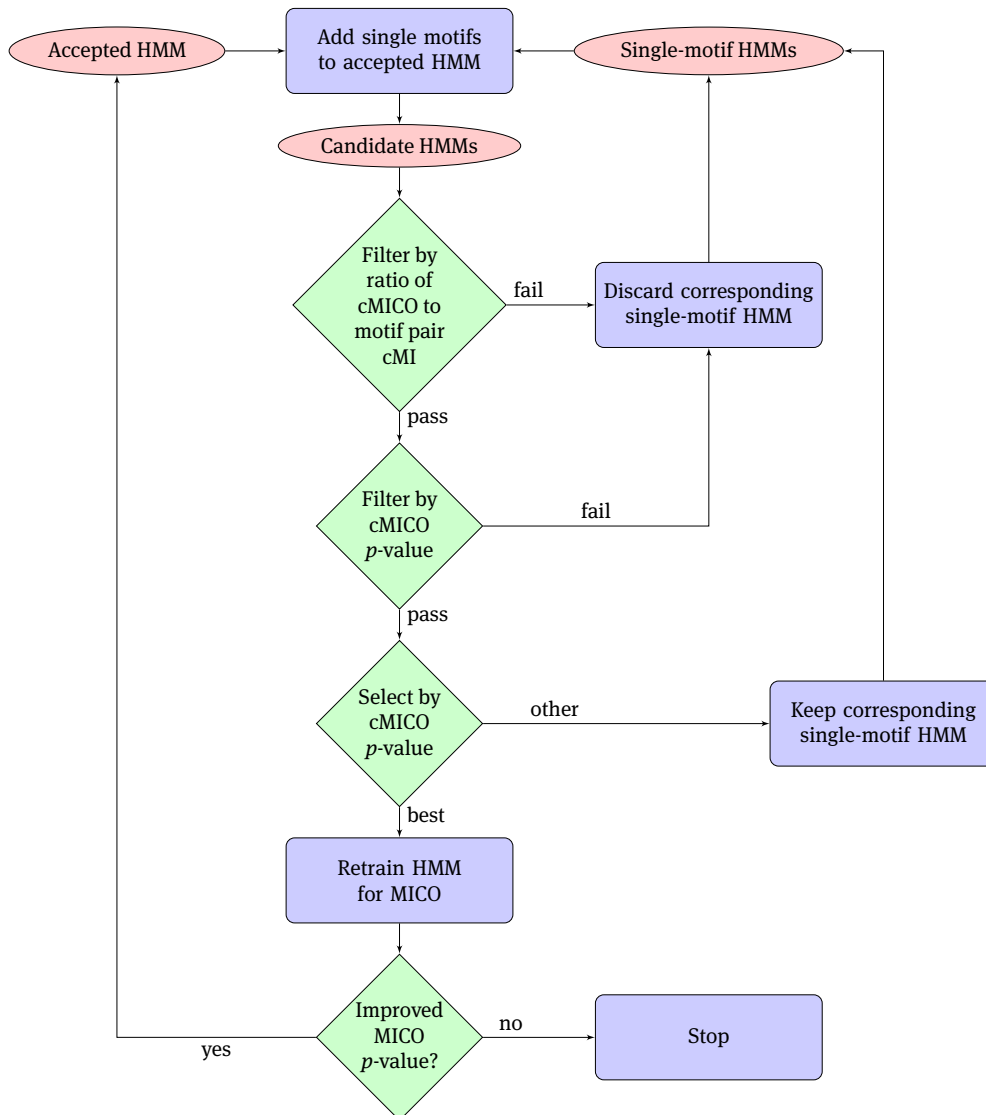


Figure 13.4: Flow chart of the second part of multiple motif discovery. Inputs are the accepted HMM, and the set of remaining single-motif HMMs resulting from the first part, depicted in figure 13.3. This part is executed until all single-motif HMMs have been accepted or discarded, or until the MICO  $p$ -value is not improved after retraining. See text for description.



# Chapter 14

## Related work

This chapter reviews previously published DMD algorithms. It first gives an overview in section 14.1. The individual methods are then reviewed in sections 14.2 to 14.11. Finally, section 14.12 briefly mentions some further discriminative tools.

### 14.1 Overview

DMD tools can be classified according to many different characteristics, among them most importantly discriminative objective function, and employed motif models. Table 14.1 gives an overview of the methods reviewed below.

### 14.2 DREME: Discriminative Regular Expression Motif Elicitation

DREME, published by Bailey (2011), is a regex-based DMD tool. The objective function of DREME is Fisher’s exact test (Fisher, 1922) applied to contingency tables of numbers of sequences with at least one occurrence of a given motif.

Table 14.1: An overview of published methods for discriminative motif discovery.

Method	Model	Objective	References
DREME	RegEx	Fisher’s exact test	Bailey (2011)
YMF	RegEx	z-score	Sinha and Tompa (2000, 2002, 2003)
CMF	PWM	z-score, min-FDR	Mason, Plath, and Zhou (2010)
DME	PWM	DLOGL	A. D. Smith, Sumazin, Das, et al. (2005) and A. D. Smith, Sumazin, and M. Q. Zhang (2005)
DIPS	Site set	DFREQ	Sinha (2006)
DECOD	PWM	DFREQ	Huggins et al. (2011)
DEME	PWM	MMIE	Redhead and Bailey (2007)
MoAn	PWM	MMIE	Valen et al. (2009)
Dispom	PWM	MMIE	Keilwagen et al. (2011)
FIRE	RegEx	MICO	Elemento, Slonim, and Tavazoie (2007) and Lieber, Elemento, and Tavazoie (2010)
Discoverer	HMM	Multiple	Maaskola and Rajewsky (2014)

**Initialization** DREME considers all  $n$ -mers occurring in the sequences for  $4 \leq n \leq 8$ , and determines for each the number of sequences that have at least one occurrence of it. From these statistics it computes the value of the objective function for each  $n$ -mer, and keeps as seeds the 100  $n$ -mers that are most significantly enriched.

**Discrete optimization** In a second stage, seed generalizations that are more significant than the seeds are found by heuristic search (see below). DREME uses regexes based on the IUPAC code for nucleic acids. In each iteration all generalizations of a regex are considered that differ from it at one position by allowing one additional nucleotide at that position. After improving the regexes until no further improvement is possible, the most significant regex is selected and accepted. In order to find multiple motifs, occurrences of the accepted motif are removed from the set of sequences by replacing them with special symbols.

**Heuristic search** DREME performs a heuristic search that keeps a pool of the 100 most significant regexes, of which generalizations are respected. In order to save runtime while determining significant generalizations in the second phase of the algorithm, DREME at first only estimates the enrichment of generalizing regexes. Only after finding the top 100 generalizations according to the estimated significance does DREME determine the actual significance of the top 100 generalized regexes.

**Termination** Finally, when significance of the regex can not be increased anymore by further generalizations, DREME determines a PSCM from the sequence occurrences that match the regex.

### 14.3 YMF: Yeast Motif Finder

Sinha and Tompa (2000, 2002, 2003) utilize in their approach YMF, Yeast Motif Finder, as alphabet a subset of the IUPAC code, namely  $\{A, C, G, T, R, Y, S, W\}$ , with optional spacer positions of N in the middle, while only allowing up to two positions of the degenerate letters R, Y, S, W. YMF requires a signal dataset of sequences of identical length. Before discovering motifs, a 3<sup>rd</sup> order Markov chain is extracted from a suitable set of genomic loci, e.g. from promoter sequences. The Markov chain is used to generate multiple control sequence sets of equally many sequences as in the signal dataset, with sequence lengths identical to those of the signal dataset.

**Objective function** YMF identifies motifs with the  $z$ -score of occurrence frequency as objective function. The  $z$ -score of motif  $m$  is defined as

$$z_m = \frac{N_m - \mu_m}{\sigma_m}, \quad (14.1)$$

where  $N_m$  is the number of occurrences of  $m$  in the signal sequences,  $\mu_m = \mathbb{E}[N_m]$  is the expected occurrence frequency of  $m$  in the control sequences, and  $\sigma_m$  is the standard de-



viation of the occurrence frequency of  $m$  in the control datasets. YMF then reports the motifs in the order of decreasing  $z$ -score.

## 14.4 CMF: Contrast Motif Finder

A related method that also uses a standardized enrichment score is the Contrast Motif Finder (CMF), published by Mason, Plath, and Zhou (2010). The objective function used by CMF is the difference of proportions test, and is given by

$$z(m) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{L_1} + \frac{1}{L_2}\right)}}, \quad (14.2)$$

where  $\hat{p}_1 = \frac{C_1}{L_1}$ ,  $\hat{p}_2 = \frac{C_2}{L_2}$ , and  $\hat{p} = \frac{C_1+C_2}{L_1+L_2}$ , and  $C_1$  and  $C_2$  are the occurrence counts of the  $w$ -mer  $m$  in the signal and control datasets  $\mathbf{S}_1$  and  $\mathbf{S}_2$ , and  $L_1$  and  $L_2$  are the number of positions at which  $w$ -mers can occur in signal and control, respectively.

CMF consists of two phases: seed finding and subsequent matrix optimization.

**Seed finding** The strategy used by CMF to determine seeds consists of finding sub-neighborhoods of words that are enriched in the signal data as compared to the control. For this first all words of length  $w$  occurring in the sequences are enumerated and their  $z$ -score is calculated. Then, every word is considered as the center of a neighborhood where two words are considered neighbors if they have Hamming distance at most  $m$ . Mason, Plath, and Zhou use  $w = 7$  and  $m = 2$  in their publication. The neighborhood is then constrained in two ways. First, only those neighbors are included that are overrepresented in the same set of sequences as the center of the neighborhood. The second restriction considers sub-neighborhoods that differ from the center of the neighborhood on the same set of positions  $J$ , where  $J$  is a size  $m$  subset of  $\{1, \dots, w\}$ . For example, for the string TTCGCGC the strings aTCGCGC, cTCGCGC, and gTCGCGC are part of the sub-neighborhood for mismatch position 1. CMF thus determines for each seed  $w$ -mer the sub-neighborhood that yields the highest  $z$ -score with counts  $C_1$  and  $C_2$  being the total number of occurrences of the members of the sub-neighborhood in signal and control.

**Iterative matrix optimization** In the second phase, CMF iteratively updates matrix motifs. For this, the sub-neighborhoods of the initial phase are summarized in  $w \times 4$  count matrices  $\mathbf{N}_1^{(1)}$  and  $\mathbf{N}_2^{(1)}$ , for signal and control respectively. In general,  $\mathbf{N}_1^{(t)}$  will contain the nucleotide counts at each position of the predicted sites in the signal sequences in iteration  $t$ , and analogously for  $\mathbf{N}_2^{(t)}$  and the predicted sites in the control sequences.

After rescaling  $\mathbf{N}_2^{(t)}$  by  $\frac{L_2}{L_1}$ , Mason, Plath, and Zhou regard  $\mathbf{N}_2^{(t)}$  as a matrix of expected nucleotide frequencies in false-positive sites. Thus, to mitigate for the presence of false-positive predictions in the signal sequences, CMF computes in each iteration  $\mathbf{N}^{(t)}$  as

$$\mathbf{N}^{(t)} = \left( N_{ij}^{(t)} \right)_{w \times 4} = \max \left( \mathbf{N}_1^{(t)} - \mathbf{N}_2^{(t)}, \mathbf{0} \right). \quad (14.3)$$

$$A = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 0.5 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 0 \\ 0.5 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 0 \\ 0 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 0 \\ 0.5 \\ 0.5 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0.5 \\ 0.5 \end{bmatrix} \right\}$$

$$B = A \cup \left\{ \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33 \\ 0.33 \\ 0 \\ 0.33 \end{bmatrix}, \begin{bmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \end{bmatrix}, \begin{bmatrix} 0 \\ 0.33 \\ 0.33 \\ 0.33 \end{bmatrix} \right\}$$

Figure 14.1: Column types of discrete matrices in DME. Column type  $B$  is used by DME for motifs of width  $w \leq 10$ , column type  $A$  for motifs of width  $w \geq 11$ .

After adding 5% pseudo-counts to this matrix, the PSFM  $\theta^{(t)}$  is obtained by renormalization. A PSFM  $\theta^{(t)}$  is used to identify predicted sites by computing the likelihood ratio of a sequence  $\mathbf{X} = X_1 \dots, X_w, X_i \in \{A, C, G, T\}$ ,

$$\text{LR}(\mathbf{X}) = \frac{\prod_{i=1}^w \theta_{i,X_i}^{(t)}}{\prod_{i=1}^w \theta_0(X_{i-1}, X_i)}, \quad (14.4)$$

where the background model  $\theta_0$  is assumed to be a 1<sup>st</sup> order Markov chain, i.e.  $\theta_0(X_{i-1}, X_i)$  is the transition probability from  $X_{i-1}$  to  $X_i$ .

During iterative updating CMF predicts sites whose likelihood ratio exceeds a threshold  $\tau^{(t)}$ , which is determined as the lowest  $\tau$  such that  $\text{FDR}(\tau) < \delta$ , where  $\delta$  is a user-specified upper bound. The FDR is estimated by  $\text{FDR}(\tau) = \frac{C_2 L_1}{C_1 L_2}$ , where  $C_1$  is the number of sites in the signal sequences with a likelihood ratio exceeding  $\tau$ , and  $C_2$  the corresponding number in the control sequences. As above,  $L_1$  and  $L_2$  are the number of positions where sites can occur.

Then the following algorithm is executed for  $t = 1, 2, \dots$ :

1. Update  $\theta^{(t)}$  using  $\mathbf{N}_1^{(t)}$  and  $\mathbf{N}_2^{(t)}$ , as described above.
2. Scan  $S_1$  and  $S_2$  with  $\theta^{(t)}$  and determine  $\tau^{(t)}$ .
3. Use sites with  $\text{LR}(\mathbf{X}) > \tau^{(t)}$  to create  $\mathbf{N}_1^{(t+1)}$  and  $\mathbf{N}_2^{(t+1)}$ .

**Termination** CMF terminates when  $d^{(t)}$ , the maximal absolute value of parameter differences between iterations  $t$  and  $t + 1$ ,

$$d^{(t)} = \max_{i,j} \left| \theta_{ij}^{(t+1)} - \theta_{ij}^{(t)} \right| < \varepsilon,$$

falls below a specified value  $\varepsilon$ . Mason, Plath, and Zhou choose  $\varepsilon = 0.01$ .

## 14.5 DME: Discriminating Matrix Enumerator

DME, the Discriminating Matrix Enumerator, published by A. D. Smith, Sumazin, Xuan, et al. (2006) and A. D. Smith, Sumazin, and M. Q. Zhang (2005), uses a discriminative

objective function similar to DLOGL, and optimizes this objective function over a discrete matrix space.

**Objective function** Unlike our definition of DLOGL in section 9.1 based on HMMs, DME does not define the likelihood  $\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})$  of a sequence  $\mathbf{X}$  in terms of smoothing over all possible parses  $\mathbf{q}$ ,  $\mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{q}} \mathbb{P}(\mathbf{X}, \mathbf{q}|\boldsymbol{\theta})$ , but rather in terms of the maximum likelihood parse,

$$\mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) = \max_{\mathbf{q}} \mathbb{P}(\mathbf{X}, \mathbf{q}|\boldsymbol{\theta}). \quad (14.5)$$

In other words, DME computes the Viterbi parse  $\mathbf{q}^*$ , and uses the Viterbi path's likelihood as the likelihood, i.e.  $\mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) = \mathbb{P}(\mathbf{X}, \mathbf{q}^*|\boldsymbol{\theta})$ , where  $\mathbf{q}^* = \operatorname{argmax}_{\mathbf{q}} \mathbb{P}(\mathbf{X}, \mathbf{q}|\boldsymbol{\theta})$  is the Viterbi parse.

A. D. Smith, Sumazin, Das, et al. consider the difference of log-likelihoods of the signal and control as function of the parameters  $\boldsymbol{\theta}$ ,

$$\log \lambda(\boldsymbol{\theta}) = \sum_{\mathbf{X} \in \mathbf{X}_{\text{signal}}} \log \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}) - \sum_{\mathbf{X} \in \mathbf{X}_{\text{control}}} \log \mathbb{P}(\mathbf{X}|\boldsymbol{\theta}). \quad (14.6)$$

But instead of maximizing  $\log \lambda(\boldsymbol{\theta})$ , DME maximizes an approximation  $\log \tilde{\lambda}(\boldsymbol{\theta})$  to it, that only considers the likelihood contribution due to motif occurrence in signal and control sequences, but not the likelihood contribution due to non-motif positions. As A. D. Smith, Sumazin, and M. Q. Zhang write, the well-foundedness of this approximation rests on some assumptions. First, that motif occurrence are not too frequent, i.e. that the number of non-motif occurrence positions is a lot larger than the number of motif occurrence positions in both signal and control sequences. And second, that the base composition of non-motif occurrence positions is close to equal between signal and control sequences. Although unmentioned by the authors, a further assumption appears to be that the signal and control dataset sizes are balanced, as mentioned in section 9.1.

Then, DME defines  $\boldsymbol{\theta}^*$  so as to maximize  $\log \tilde{\lambda}(\boldsymbol{\theta})$ ,

$$\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} \log \tilde{\lambda}(\boldsymbol{\theta}). \quad (14.7)$$

**Initial search** As mentioned above, DME optimizes over a discrete matrix space, where matrices are formed as products of discrete column spaces. A. D. Smith, Sumazin, and M. Q. Zhang state two intended goals in the choice of allowed column types. One is to cover matrix column space uniformly in the sense of a grid that spans the space it is embedded in uniformly. Thus, DME allows different column types depending on the motif length, see figure 14.1. Second, that the discrete column spaces are combined in a manner that enforces a minimal matrix-wise IC, depending on the motif length, see table 14.2.

DME then enumerates all matrices in this discrete space, and computes the objective function (14.7) for each. The matrix that maximizes the objective function is then subjected to the local search procedure described below for further optimization.

**Local search** DME terms the matrix resulting from the initial search the original matrix. In the second phase, DME performs a local search in the vicinity of the original matrix

Table 14.2: Default parameters for DME depending on the width of the motif. Column types given in figure 14.1; IC: information content in bit;  $g$ : parameter for local search.

Width	Column types	IC per column	min $g$
6	B	1.90	0.01
7	B	1.90	0.01
8	B	1.80	0.01
9	B	1.65	0.05
10	B	1.60	0.05
11	A	1.55	0.10
12	A	1.50	0.10

that achieves a higher objective function value. A parameter  $g \in (0, 1)$  is involved that determines the distance in which candidate matrices are tried, and which is increasingly constrained during local search. Given a matrix  $M$  and a value for  $g$ , A. D. Smith, Sumazin, and M. Q. Zhang define a  $g$ -neighborhood  $N_g(M_i)$  for each column  $M_i$  as follows,

$$N_g(M_i) = \{X | d_1(M_i, X) \in \{0, 2g\}, \forall j : |M_{ij} - X_j| \in \{0, g\}\}, \quad (14.8)$$

where  $d_1(X, Y) = \|X - Y\|_1 = \sum_i |X_i - Y_i|$ . The  $g$ -neighborhood of  $M$  then consists of all matrices  $M'$  such that  $M'_i \in N_g(M_i)$ . There are  $2^{\binom{4}{2}} + 1 = 13$   $g$ -neighbor columns for each column, thus for a length  $w$  matrix there are  $13^w - 1$   $g$ -neighbor matrices.

DME computes the objective function for each of the  $g$ -neighbor matrices. If any of the neighbors achieves a higher score than the original matrix, it is used in place of the original matrix,  $g$  is halved, and the procedure continues until  $g$  is below a given cutoff. The cutoff for  $g$  depends on the motif width, see table 14.2.

**Termination and further motifs** After local search has finished, DME accepts the found matrix, and masks all of its occurrences in the sequences. Subsequently, further motifs are sought by repeating the procedure on the masked sequences.

**Extension** Related to DME is DME-X (A. D. Smith, Sumazin, Das, et al., 2005) which extends DME by incorporating localization information.

## 14.6 DIPS: Discriminative PWM Search

The motif finder DIPS, Discriminative PWM Search, by Sinha (2006) uses as binding motif model a set of subsequences of equal length  $m$  of the signal sequences. These subsequences define in a unique way a PSCM, which together with a (fixed) background model, gives rise to a unique PWM. As described below, DIPS uses the PWMs in a way similar to HMMs.

**Objective function** DIPS computes the relative occurrence frequencies of a motif  $M$  per position in the signal and control sequences,  $f_S(M)$  and  $f_C(M)$ , and considers their differ-

ence as objective function.

$$f(M|S, C, \boldsymbol{\theta}) = f_S(M|\boldsymbol{\theta}) - f_C(M|\boldsymbol{\theta}), \quad (14.9)$$

where the relative frequency of motif occurrence of motif  $M$  among the signal sequences is defined as

$$f_S(M|\boldsymbol{\theta}) = \frac{\sum_{\mathbf{X} \in \mathcal{X}_S} \frac{\mathbb{E}[M|\mathbf{X}, \boldsymbol{\theta}]}{|\mathbf{X}|}}{|\mathcal{S}|}, \quad (14.10)$$

where  $|\mathcal{S}|$  is the number of signal sequences, the summation is over all signal sequences  $\mathbf{X}$ ,  $|\mathbf{X}|$  is the length of sequence  $\mathbf{X}$ , and  $\mathbb{E}[M|\mathbf{X}, \boldsymbol{\theta}]$  is the expected number of occurrences of  $M$  in the sequence  $\mathbf{X}$  given the model parameters  $\boldsymbol{\theta}$ . This is

$$\mathbb{E}[M|\mathbf{X}, \boldsymbol{\theta}] = \sum_{\mathbf{q}} c(M|\mathbf{q})\mathbb{P}(\mathbf{q}|\mathbf{X}, \boldsymbol{\theta}), \quad (14.11)$$

where the summation is over all parses  $\mathbf{q}$  of the model  $\boldsymbol{\theta}$  over the sequence  $\mathbf{X}$ ,  $c(M|\mathbf{q})$  gives the number of occurrences of  $M$  in the parse  $\mathbf{q}$ , and  $\mathbb{P}(\mathbf{q}|\mathbf{X}, \boldsymbol{\theta})$  is the posterior probability of parse  $\mathbf{q}$  given the sequence  $\mathbf{X}$  and parameters  $\boldsymbol{\theta}$ . Sinha (2006) computes  $\mathbb{E}[M|\mathbf{X}, \boldsymbol{\theta}]$  using the forward-backward algorithm from the theory of HMMs, described in chapter 4.

$f_C(M)$  is defined analogously to  $f_S(M)$ .

**Optimization** As written above, DIPS operates on sets of subsequences of length  $m$  of the signal sequences. Initially, a random set of subsequences is chosen. DIPS then iteratively improves the set such that the objective function of the PWM  $M$  resulting from the set of subsequences increases. In each iteration a number of improvements is tried, as described next.

Modifications to the set of subsequences are performed by first deleting one of the subsequences from the set. For this, the objective function is determined for each of  $n$  sets of subsequences that result from deleting one of the subsequences from the current set of  $n$  subsequences. Then, the subsequence is deleted for which the objective function of the resulting set of  $n - 1$  subsequences is maximal.

Subsequently, a new subsequence is selected that is to be added to the set. Not all possible such subsequences are tried, but rather selection is stopped once a subsequence has been found whose resulting objective function is higher than the objective function prior to the removal of this iteration's removed subsequence.

DIPS uses gradient information to order the subsequences that are probed for addition so as to try first promising subsequences.

**Termination** Iterations of modifications to the set of subsequences are performed until in an iteration 500 modifications to the set have been tried without yielding an improved score.

**Further capabilities** As part of the model  $\boldsymbol{\theta}$ , in addition to the to-be-optimized PWM  $M$ , DIPS allows the specification of a set of previously known motifs.

DIPS also offers another discriminative objective function. Before counting motif occurrences, this alternative objective function transforms PWM scores using the logistic function into soft classifiers. Sinha (2006) reports that out of 20 replicates of test data 9 times the above defined objective function outperformed the logistic variant, while the logistic variant was better in 3 cases, with the remaining 8 being inconclusive.

## 14.7 DECOD: Deconvolved Discriminative Motif Discovery

DECOD, short for Deconvolved Discriminative Motif Discovery, by Huggins et al. (2011) uses DFREQ as discriminative objective function, and is thus limited to the analysis of binary contrasts. Unlike the definition of DFREQ given in chapter 7, the feature used by DECOD is not sequences with at least one occurrence of a motif, but relative occurrence frequency of motifs in the sequences. DECOD models binding motifs with PSFMs.

**Count table based approach** DECOD extracts two tables of  $k$ -mer counts,  $\mathbf{A}$  and  $\mathbf{B}$  for the signal and control data, respectively. All subsequent operations are based on these count tables. A significant benefit offered by using count tables is fast look-up times. However,  $k$ -mers that overlap only partially with motif occurrences are also enriched in signal sequences, and could be mistaken for the true motif. To overcome this difficulty, DECOD uses a deconvolution approach that explicitly models the convolved  $k$ -mer occurrences in the signal sequences.

**Mixture model** DECOD uses a mixture model  $\theta$  for  $k$ -mers, composed of a motif component  $M$  and a background component  $B$ , with a mixture parameter  $p = \mathbb{P}(M) = 1 - \mathbb{P}(B)$ . Thus the likelihood of a  $k$ -mer  $\mathbf{X} \in \Sigma^k$  is given by

$$\mathbb{P}(\mathbf{X}|\text{DECOD}) = p\mathbb{P}(\mathbf{X}|M) + (1 - p)\mathbb{P}(\mathbf{X}|B). \quad (14.12)$$

The background component's likelihood  $\mathbb{P}(\mathbf{X}|B)$  is modeled by a simple  $0^{\text{th}}$  order Markov model.

**Convolved motif component** In order to account for convolution of partially overlapping  $k$ -mers, DECOD specifies the motif component's likelihood by

$$\begin{aligned} \mathbb{P}(\mathbf{X}|M) = \frac{1}{2k-1} & \left( \mathbb{P}(\mathbf{X}|B_{k-2}M^{k-1}) + \mathbb{P}(\mathbf{X}|B_{k-3}M^{k-2}) + \dots \right. \\ & + \mathbb{P}(\mathbf{X}|B_0M^1) + \mathbb{P}(\mathbf{X}|M^0) + \mathbb{P}(\mathbf{X}|M_{k-2}B^{k-1}) + \\ & \left. \dots + \mathbb{P}(\mathbf{X}|M_0B^1) \right), \end{aligned} \quad (14.13)$$

where  $M_i$  and  $B_i$  signify that positions 0 to  $i$  of  $\mathbf{X}$  are generated by the last  $i$  positions of  $M$  or  $B$  respectively; similarly,  $M^i$  and  $B^i$  denote that positions  $i$  to  $k-1$  of  $\mathbf{X}$  are generated by the first  $k-i+1$  positions of  $M$  or  $B$ , respectively. Note that in this notation  $\mathbb{P}(\mathbf{X}|M^0) = \mathbb{P}(\mathbf{X}|M_{k-1})$ .

**Expected difference of motif occurrence frequency** Using the count tables  $\mathbf{A}$  and  $\mathbf{B}$ , we can look up the number of times  $k$ -mer  $\mathbf{X}$  is occurring in the signal and control sequences,  $A_{\mathbf{X}}$  and  $B_{\mathbf{X}}$ . From this, the expected difference of occurrences of the motif  $M$  between the signal and control sequences is then given by

$$\begin{aligned} \mathbb{E}_A[M] - \mathbb{E}_B[M] &= \sum_{\mathbf{X} \in \Sigma^k} \mathbb{P}(M|\mathbf{X}) (A_{\mathbf{X}} - B_{\mathbf{X}}) \\ &= \sum_{\mathbf{X} \in \Sigma^k} \frac{p\mathbb{P}(\mathbf{X}|M)}{p\mathbb{P}(\mathbf{X}|M) + (1-p)\mathbb{P}(\mathbf{X}|B)} (A_{\mathbf{X}} - B_{\mathbf{X}}). \end{aligned} \quad (14.14)$$

**Optimization and speed-up** A discretized heuristic hill-climbing method is employed by DECOD to optimize the objective function (14.14).

While look-up times for the number of occurrences of a  $k$ -mer take constant time with precomputed count tables, the number of  $k$ -mers over which equation (14.14) is summing still grows exponentially<sup>1</sup> in  $k$ . For this reason, DECOD implements a speedup variant, that computes an approximation to the partial derivative of (14.14). The approximation replaces the sum over all  $k$ -mers by performing two sums. The first is over all  $k$ -mers that exhibit a large occurrence frequency difference between signal and control sequences. The second sum then uses the current motif estimate  $M$  to include all those  $k$ -mers that are similar to  $M$ .

## 14.8 MoAn: Motif Annealer

MoAn, the Motif Annealer, by Valen et al. (2009), is a DMD tool whose objective function is based on the conditional probability of the class label  $C$  given the sequence  $\mathbf{X}$  and the model parameters  $\theta$ ,  $\mathbb{L}(\theta; C, \mathbf{X}) = \mathbb{P}(C|\mathbf{X}, \theta)$ .

**Objective function** Assuming, independence of the labeled samples  $D = \{(C, \mathbf{X})\}$ , the objective function of MoAn is the joint log likelihood of correct classification of all samples,

$$\log_2 \mathbb{L}(\theta; D) = \sum_{(C, \mathbf{X}) \in D} \log_2 \mathbb{P}(C|\theta, \mathbf{X}). \quad (14.15)$$

The conditional probability of the class label  $C$ ,  $\mathbb{P}(C|\mathbf{X}, \theta)$ , is computed via Bayes theorem from

$$\mathbb{P}(C|\mathbf{X}, \theta) = \frac{\mathbb{P}(\mathbf{X}, C|\theta)}{\mathbb{P}(\mathbf{X}|\theta)} \quad (14.16)$$

$$= \frac{\mathbb{P}(\mathbf{X}|C, \theta)\mathbb{P}(C)}{\sum_{C'} \mathbb{P}(\mathbf{X}|C', \theta)\mathbb{P}(C')}. \quad (14.17)$$

In MoAn, the prior of class  $C$  is estimated as the relative frequency of sequences of this class among all sequences,  $\mathbb{P}(C) = \frac{n_C}{\sum_{C'} n_{C'}}$ , where  $n_C$  is the number of sequences of class

<sup>1</sup>Of course, the exponential growth of the number of  $k$ -mers is still limited by the linear length of all sequences.

C. The other terms in this expression,  $\mathbb{P}(\mathbf{X}|C, \boldsymbol{\theta})$ , the likelihood of sequence  $\mathbf{X}$  in class  $C$ , are specified by the probabilistic model described next.

**Probabilistic model** MoAn's likelihood function assumes two motifs, which can either have or not have one binding site in a sequence. Thus the likelihood function is a sum over four terms that represent no binding site, one binding site of motif 1, one binding site of motif 2, or one binding site of either motif,

$$\begin{aligned} \mathbb{P}(\mathbf{X}|C, \boldsymbol{\theta}) &= \mathbb{P}(\text{no BS}|C)\mathbb{P}(\mathbf{X}|\text{no BS}, \boldsymbol{\theta}) \\ &\quad + \mathbb{P}(\text{BS}_1|C)\mathbb{P}(\mathbf{X}|\text{BS}_1, \boldsymbol{\theta}) \\ &\quad + \mathbb{P}(\text{BS}_2|C)\mathbb{P}(\mathbf{X}|\text{BS}_2, \boldsymbol{\theta}) \\ &\quad + \mathbb{P}(\text{BS}_1, \text{BS}_2|C)\mathbb{P}(\mathbf{X}|\text{BS}_1, \text{BS}_2, \boldsymbol{\theta}). \end{aligned} \tag{14.18}$$

The terms  $\mathbb{P}(\text{no BS}|C)$ ,  $\mathbb{P}(\text{BS}_1|C)$ ,  $\mathbb{P}(\text{BS}_2|C)$ , and  $\mathbb{P}(\text{BS}_1, \text{BS}_2|C)$ , which represent the prior probabilities of the joint or single (non-)occurrence of motifs 1 and 2 in class  $C$ , are parameters of MoAn.

When MoAn is executed in a single motif mode, then  $\mathbb{P}(\text{BS}_2|C) = \mathbb{P}(\text{BS}_1, \text{BS}_2|C) = 0$ . For  $C = \text{signal}$ , in single motif mode the default values for the remaining parameters are,  $\mathbb{P}(\text{no BS}|\text{signal}) = 0.01$ ,  $\mathbb{P}(\text{BS}_1|\text{signal}) = 0.99$ . For  $C = \text{control}$ , the default values are,  $\mathbb{P}(\text{no BS}|\text{control}) = 0.8$ ,  $\mathbb{P}(\text{BS}_1|\text{control}) = 0.2$ .

MoAn uses a  $o^{\text{th}}$  order background model for  $\mathbb{P}(\mathbf{X}|\text{no BS}, \boldsymbol{\theta})$ ,

$$\mathbb{P}(\mathbf{X}|\text{no BS}, \boldsymbol{\theta}) = \prod_{i=0}^{|\mathbf{X}|-1} \mathbb{P}(X_i|\text{BG}, \boldsymbol{\theta}), \tag{14.19}$$

where  $|\mathbf{X}|$  is the length of sequence  $\mathbf{X}$ , and  $\mathbb{P}(X_i|\text{BG}, \boldsymbol{\theta})$  is the probability of emitting  $X_i$ , the  $i$ -th symbol of  $\mathbf{X}$ , in the background model.

Defining the log-odds score matrix  $\mathbf{W} = (W_{i,b})_{\substack{0 \leq i < w \\ b \in \mathcal{A}}}$ , where  $w$  is the length of  $\mathbf{W}$ , for position  $i$  and symbol  $b$  by

$$W_{i,b} = \log_2 \frac{\mathbb{P}(b|\text{Motif}_1(i), \boldsymbol{\theta})}{\mathbb{P}(b|\text{BG}, \boldsymbol{\theta})}, \tag{14.20}$$

where  $\mathbb{P}(b|\text{Motif}_1(i), \boldsymbol{\theta})$  is the emission frequency of symbol  $b$  at the  $i$ -th position of motif 1, Valen et al. compute the log odds score of a motif occurrence at position  $k$  in sequence  $\mathbf{X}$  with

$$S(\mathbf{X}, k, \mathbf{W}) = \log_2 \mathbb{P}(\mathbf{X}_{k, \dots, k+w-1}|\text{Motif}_1, \boldsymbol{\theta}) - \log_2 \mathbb{P}(\mathbf{X}_{k, \dots, k+w-1}|\text{BG}, \boldsymbol{\theta}) = \sum_{i=0}^{w-1} W_{i, X_{k+i}}. \tag{14.21}$$



The model used by MoAn for the case where one binding site of motif 1 is occurring is

$$\begin{aligned} \mathbb{P}(\mathbf{X}|\text{BS}_1, \boldsymbol{\theta}) &= \frac{1}{|\mathbf{X}| - w + 1} \sum_{k=0}^{|\mathbf{X}|-w} H(S(\mathbf{X}, k, w) - c) 2^{S(\mathbf{X}, k, w)} \prod_{i=0}^{|\mathbf{X}|-1} \mathbb{P}(X_i|\text{BG}, \boldsymbol{\theta}) \\ &= \frac{1}{|\mathbf{X}| - w + 1} \sum_{k=0}^{|\mathbf{X}|-w} H(S(\mathbf{X}, k, w) - c) \frac{\mathbb{P}(\mathbf{X}_{k, \dots, k+w-1}|\text{Motif}_1, \boldsymbol{\theta})}{\prod_{i=k}^{k+w-1} \mathbb{P}(X_i|\text{BG}, \boldsymbol{\theta})} \prod_{i=0}^{|\mathbf{X}|-1} \mathbb{P}(X_i|\text{BG}, \boldsymbol{\theta}), \quad (14.22) \end{aligned}$$

where,  $H(x)$  is the Heaviside step function,  $H(x) = 0$  for  $x < 0$  and  $H = 1$  for  $x \geq 0$ , and  $c$  is a cutoff on the PWM score of motif 1.

The likelihood  $\mathbb{P}(\mathbf{X}|\text{BS}_2, \boldsymbol{\theta})$  is defined analogously to  $\mathbb{P}(\mathbf{X}|\text{BS}_1, \boldsymbol{\theta})$ ; for the definition of  $\mathbb{P}(\mathbf{X}|\text{BS}_1, \text{BS}_2, \boldsymbol{\theta})$  see Valen et al. (2009).

**Optimization** MoAn uses simulated annealing to sample PWM space, so it directly optimizes matrices for its discriminative objective function. The advantage of this is that it performs a global optimization, and does not rely on a specific choice of initial value. Conversely, the large number of iterations necessary to sufficiently explore parameter space can make the method prohibitively slow for large dataset sizes.

## 14.9 DEME: Discriminatively Enhanced Motif Elicitation

DEME, Discriminatively Enhanced Motif Elicitation, by Redhead and Bailey (2007) uses, like MoAn, the conditional log-likelihood of the class labels as objective function, see equation (14.15). The probabilistic model for the classes used by DEME is nearly equivalent to the single motif model of MoAn, described in section 14.8. There are only two small differences. First, DEME does not threshold on the PWM score, thus instead of the Heaviside function in equation (14.22) DEME has a constant factor of unity. Second, DEME assumes that there are no binding sites in the control data, i.e. for  $C = \text{control}$ ,  $\mathbb{P}(\text{no BS}|\text{control}) = 1$  and  $\mathbb{P}(\text{BS}|\text{control}) = 0$  (see section 14.8). For  $C = \text{signal}$ , DEME learns a parameter  $\lambda$  such that  $\mathbb{P}(\text{BS}|\text{signal}) = \lambda$  and  $\mathbb{P}(\text{no BS}|\text{signal}) = 1 - \lambda$ .

Differently from MoAn which operates exclusively by sampling parameters of the probabilistic model, DEME first performs a substring search, followed by a branching search on regexes, and finally turns to gradient-based matrix optimization.

**Initial search** Substring search considers each substring of length  $w$  occurring in the sequences, and maps it to a PSFM by mixing with a uniform distribution, where the mixing is determined by a parameter specified by the user. The PSFMs corresponding to each occurring substring are evaluated using the objective function and the top  $n$  are retained.

During branching search, for each of the top  $n$  substrings all possible ways of allowing one additional nucleotide at any position are considered. For each of these, a PSFM is formed as above and their corresponding objective function value is determined. Again, the top  $n$  are retained. This procedure is repeated for a number of iterations that is given by the user.

**Local search** The regex that achieved the highest objective function value during the initial search is used as a seed, and its corresponding PSFM is used as starting point for local search. Optimization of the PSFM and the parameter  $\lambda$  is done by DEME using conjugate gradient (Press et al., 1995).

Instead of directly optimizing the PSFM and the parameter  $\lambda$ , DEME transforms both in a way similar to what is done in Discover to transform the constrained into an unconstrained optimization problem (see Redhead and Bailey, 2007, for details).

## 14.10 Dispom

Dispom, published by Keilwagen et al. (2011), is a discriminative motif learning approach that incorporates modeling of positional preference. Dispom uses as binding site model a PWM that is assumed to be embedded in a  $o^{\text{th}}$  order Markov chain background.

**Probabilistic model** The binding site model is related to the ZOOPS model pioneered by MEME (Bailey and Elkan, 1995a,b; Bailey, Williams, et al., 2006), and described in equation (3.20). The positional preference model,  $\mathbb{P}(M_i|\boldsymbol{\theta})$  in equation (3.19), of Dispom is related to those of Improbizer (Ao et al., 2004) and A-GLAM (N. Kim et al., 2008). Like these approaches Dispom learns the positional preference simultaneously with that for the sequence motif. Specifically, Dispom uses a mixture of a skew normal and a uniform distribution to model the positional distribution.

A consequence of learning a positional model is that Dispom is limited to datasets with sequences of identical length. Note that actually only a weaker constraint would be necessary: the sequences have to be aligned to a common reference point. This could for example be the position of the TSS in case of promoter sequences.

**Objective function** The objective function of Dispom is based on a supervised classification posterior for the class  $C = c$  of a sequence  $\mathbf{X}$ ,

$$\begin{aligned} \mathbb{P}(C = c|\mathbf{X}, \boldsymbol{\theta}) &= \frac{\mathbb{P}(C = c, \mathbf{X}|\boldsymbol{\theta})}{\mathbb{P}(\mathbf{X}|\boldsymbol{\theta})} \\ &= \frac{\mathbb{P}(C = c, \mathbf{X}|\boldsymbol{\theta})}{\sum_{\tilde{c} \in \mathcal{C}} \mathbb{P}(C = \tilde{c}, \mathbf{X}|\boldsymbol{\theta})} \\ &= \frac{\mathbb{P}(C = c|\boldsymbol{\theta})\mathbb{P}(\mathbf{X}|C = c, \boldsymbol{\theta})}{\sum_{\tilde{c} \in \mathcal{C}} \mathbb{P}(\tilde{c}|\boldsymbol{\theta})\mathbb{P}(\mathbf{X}|C = \tilde{c}, \boldsymbol{\theta})}. \end{aligned} \tag{14.23}$$

For signal sequences, Dispom assumes the likelihood  $\mathbb{P}(\mathbf{X}|C = \text{signal}, \boldsymbol{\theta})$  to be given by the ZOOPS model of equation (3.20) extended by the positional distribution model, as mentioned above. The likelihood of control sequences,  $\mathbb{P}(\mathbf{X}|C = \text{control}, \boldsymbol{\theta})$ , is assumed by Dispom to be given by a  $o^{\text{th}}$  order Markov model.

Finally, the objective function of Dispom<sup>2</sup> is given by maximizing  $\boldsymbol{\theta}$  over the sum of log

<sup>2</sup>Note that there is an error in equation 5 of Keilwagen et al. (2011). I notified the authors, and they promptly issued a formal correction. See also appendix L.1.

supervised classification posteriors for all sequences  $\mathbf{X}_i$  with classes  $c_i$ ,

$$\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_i \log \mathbb{P}(C = c_i | \mathbf{X}_i, \boldsymbol{\theta}) + \log Q(\boldsymbol{\theta} | \boldsymbol{\alpha}), \quad (14.24)$$

where  $Q(\boldsymbol{\theta} | \boldsymbol{\alpha})$  is some choice of prior for the parameters.

**Optimization** Dispom optimizes the objective function by gradient optimization. A set of heuristic steps aims to avoid getting stuck in local minima due to locking in on phase shifted variants of a motif. These heuristics consists in testing extensions and shortenings of the motif during optimization. Also Dispom tries several different starting point for the gradient optimization.

## 14.11 FIRE: Finding Informative Regulatory Elements

Another example of a word-based DMD method is FIRE (Elemento, Slonim, and Tavazoie, 2007; Lieber, Elemento, and Tavazoie, 2010). The acronym FIRE stands for Finding Informative Regulatory Elements. FIRE uses MICO as discriminative objective function.

**Initialization** The method starts off by considering all 7-mers occurring in the sequences, which are sorted by MICO. The highest scoring ones are kept as seeds for a subsequent optimization phase.

**Randomized greedy optimization** For the discrete optimization phase, the 7-mers are embedded in the middle of 9-mers by extending both ends with Ns. Then, modifications to the motif are tested to find variants that yield higher MICO. In particular, a position is chosen randomly among the 9 positions, and all degenerate symbols consistent with the present symbol at that position are considered. Consistency is here defined such that only degenerate codes that include the present symbol are allowed. All consistent replacements for the randomly chosen position are tested for discrimination, and if a variant increases discrimination sufficiently, it replaces the previous motif. This greedy procedure is repeated until no further improvement is possible, and the whole procedure is repeated ten times for each seed.

**Additional features** In addition to finding motifs that discriminate between signal and control sequences, Elemento, Slonim, and Tavazoie also use MICO to identify motifs that exhibit positional bias, or that exhibit strand bias. This is done by means of Perl scripts that break the sequence sets into groups according to position or strand, and evaluating MICO on the thus-defined conditions. Similarly, FIRE uses MICO to find motifs associated with expression (changes) by binning continuous expression values into quantized expression groups.

## 14.12 Further motif discovery tools

There exist further MD tools based on discriminative learning techniques.

These include a method based on the hyper-geometric test for enrichment (Barash, Bejerano, and N. Friedman, 2001). This unnamed method is apparently the first published DMD method. No source code or executables are available for this method.

ALSE (acronym for: all sequences) by Leung and Chin (2006) apparently uses a probabilistic model for enrichment. Unfortunately, the description of the objective function in the paper is fairly inscrutable. In addition, the available source code fails to compile with contemporaneous compilers, see appendix L.1.

The ChIP-Seq data analysis suite HOMER (Heinz et al., 2010) includes a method for *de-novo* MD in ChIP-Seq data. Unfortunately, no publication describing the MD method is available, and information about it is limited to a web page that describes usage of the program. This web page has a dead link to another page that purportedly describes the method. The method seemingly computes a  $p$ -value for enrichment of motif occurrences in a set of signal sequences. The calculation is based on the binomial distribution, and HOMER automatically determines a set of GC-content matched background sequences.

Another MD method that uses discriminative learning is DISPARE (Piedade, M. E. Tang, and Elemento, 2009). Unfortunately, the description of Piedade, M. E. Tang, and Elemento was not clear enough to allow the author of this thesis to understand the objective function that the method tries to optimize.

## **Part IV**

# **Empirical Study of Motif Discovery Methodology**



# Overview of applications

The following part is an empirical study of MD methodology. It will showcase applications of discriminative and non-discriminative learning methods for probabilistic sequence analysis in MD problems, including synthetic data, data of RNA-binding proteins (RBPs), and of transcription factors (TFs).

Chapter 15 describes how synthetic data is generated that is used for supervised performance evaluation. Metrics to evaluate MD performance are introduced in chapter 16. Then, chapter 17 presents an evaluation of MD performance for various methods implemented in Discover and in published DMD tools.

Chapter 18 presents an application of DMD to the PUF RBP family. The data come from various publications generated with two different technologies: RIP-Chip and PAR-CLIP data. The results confirm previous findings, proving applicability of Discover for the analysis of real biological datasets.

Chapter 19 studies the alternative splicing regulator RBM10. DMD is applied to PAR-CLIP data. This reveals motifs which have been implicated in splicing regulation but that have not previously been connected to RBM10.

The part concludes in chapter 20 with an application to numerous mouse embryonic stem cell (ESC) TFs. ChIP-Seq data of two publications will be analyzed.





## Chapter 15

# Synthetic test data

Testing performance of MD methods on synthetic data offers several advantages, as well as some disadvantages. Most importantly, generating synthetic data allows to evaluate performance in a supervised manner, as true and false site occurrences are intrinsically defined in the data generation process. Therefore control over the parameters determining the synthetic data generating process allows for controlled experiments of the sensitivity of MD methods with respect to important variables.

Regarding disadvantages of performance testing with synthetic data, it is trivially true for any specific choice of a synthetic data generation procedure that not all possible effects can be tested. Thus, the central disadvantage of performance evaluation on synthetic data is the blindness for sensitivity to effects not realized in the synthetic data.

In general, it is of course unknown whether all important effects that exist in real data are accounted for in the generation of the synthetic test data. For this reason, diligence is due in choosing which parameters to test.

### 15.1 Generation of synthetic data

Synthetic data was generated according to algorithm 6 and as described below.

Three sets of experiments were generated, respectively referred to as basic, 3'UTR, and decoy experiments. The parameters varied in the experiments are summarized in table 15.1, and include length and number of sequences, information content (IC) and implantation frequency of signal (and decoy) motifs.

For each data generation parameter setting a pair of signal and control sequence sets is generated. A signal motif with a specific IC is generated by choosing a random PWM and polarizing (exponentiating component-wise and renormalizing) so as to achieve the desired IC. Each sequence is generated according to the background model of the set of experiments it is part of. Then, motifs are implanted into the signal sequences: signal sequences are selected with a given implantation probability, and for each selected sequence one signal motif occurrence is generated from the PWM and inserted into the sequence at a random position.

All motifs are inserted on the sense strand, simulating an RNA MD experiment. The

**Algorithm 6** Synthetic sequence data generation

**Input:** Motif length  $w$ , signal and decoy motif IC, length  $l$  and number  $n$  of sequences, signal and control motif implantation probabilities  $\mathbb{P}_{\text{signal}}$  and  $\mathbb{P}_{\text{decoy}}$ , background model

**Output:**  $n$  signal and decoy sequences

```

1: generate signal PWM of length  $w$  with signal motif IC
2: optionally generate decoy PWM of length  $w$  with decoy IC
3: generate  $n$  signal sequences of length  $l$  using the background model
4: generate  $n$  control sequences of length  $l$  using the background model
5: for  $i \leftarrow 1 \dots n$  do
6:   with probability  $\mathbb{P}_{\text{decoy}}$  do
7:     generate decoy motif instance from decoy PWM
8:     implant decoy motif instance at random position of signal sequence  $i$ 
9:   with probability  $\mathbb{P}_{\text{decoy}}$  do
10:    generate decoy motif instance from decoy PWM
11:    implant decoy motif instance at random position of control sequence  $i$ 
12: for  $i \leftarrow 1 \dots n$  do
13:   with probability  $\mathbb{P}_{\text{signal}}$  do
14:    generate signal motif instance from signal PWM
15:    implant signal motif instance at random position of signal sequence  $i$ 
16: return signal and decoy sequences

```

Table 15.1: Parameters for the generation of synthetic sequence data.

Parameter	Basic	3'UTR	Decoy
Sequence background	Uniform, $0^{\text{th}}$ -order MC	Human 3'UTR	Uniform, $0^{\text{th}}$ -order MC
Sequence length [nt]	20, 50, 100, 200, 500, 1000	20, 50, 100, 200, 500, 1000	100
Sequence number	100, 1000, 10 000	100, 1000, 10 000	10 000
Motif length [nt]	8	8	8
Signal per sequence	0 or 1	0 or 1	0 or 1
Signal probability [%]	1, 2, 5, 10, 20, 50, 100	1, 2, 5, 10, 20, 50, 100	10
Signal motif IC [bit]	0, 2, 4, 6, 8, 10, 12, 14, 16	0, 2, 4, 6, 8, 10, 12, 14, 16	0, 2, 4, 6, 8, 10, 12, 14, 16
Decoy per sequence	0	0	0 or 1
Decoy probability [%]	0	0	1, 2, 5, 10, 20, 50, 100
Decoy motif IC [bit]	NA	NA	0, 2, 4, 6, 8, 10, 12, 14, 16
Strandedness	Single-stranded	Single-stranded	Single-stranded
Total experiments	1134	1134	567

three sets of experiments differ by the choice of the sequence context into which motifs are inserted. The basic and decoy experiments use a uniform,  $0^{\text{th}}$ -order Markov chain to generate synthetic sequences, while the 3'UTR experiments use sequences sampled from human 3'UTRs. In the decoy experiments, before implanting the signal motifs, occurrences of decoy motifs are implanted both into signal and control sequences.

## 15.2 Recognizability

How well motifs can be discovered depends on the difficulty of recognizing the motif when it is already known. As a reference we evaluate motif recognizability, which we define as the predictive performance of the true model. As an approximation to the true model we use HMMs comprising a background state and a motif chain with emission probabilities equal to the implanted PWM. The transition probability from the background state to the motif state is chosen such that the expected number of motifs per sequence is equal to the

implantation frequency of the experiment. I.e. if implantation frequency is 10% and if the sequence length is 100 then the per-position probability of transiting from the background state to the motif chain is set to  $0.1 \times 0.01 = 0.001$ . Background emissions of order zero are used. The background emission probabilities are set to uniform distributions for the basic and decoy experiments, but for the 3'UTR experiments are fit to the data with the Baum-Welch algorithm prior to evaluation.



# Chapter 16

## Performance metrics

Because the construction of synthetic data for performance evaluation, described in chapter 15, intrinsically defines true binding sites, it is possible to use supervised metrics to measure MD performance. This chapter presents a choice of supervised performance metrics in section 16.1. Subsequently, summarization techniques will be discussed in section 16.2.

### 16.1 Supervised performance metrics

Supervised performance metrics for MD require, in addition to the predicted binding sites, an exhaustive specification of all true binding sites in a synthetic sequence dataset. Evaluation may be done either on nucleotide level with metrics presented in section 16.1.1, or on the binding site level using metrics presented in section 16.1.2.

#### 16.1.1 Nucleotide level performance metrics

Given a set of implanted and predicted motif coordinates, individual nucleotide positions can be classified as true or false positives, and true or false negatives, as illustrated in figure 16.1.

**Basic nucleotide level statistics** The following definitions for nucleotide-level statistics can be found in Tompa et al. (2005).

*nTP* Number of nucleotides part of a site correctly predicted

*nTN* Number of background nucleotides correctly predicted

*nFP* Number of background nucleotides predicted to be part of a motif

*nFN* Number of nucleotides part of a site predicted as background

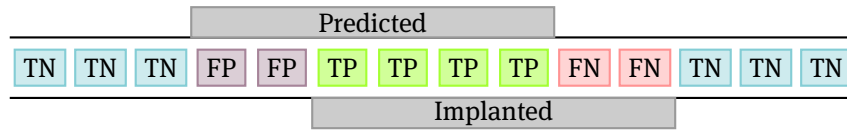


Figure 16.1: Classification of nucleotide positions as true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), based on the overlap with implanted and predicted motifs.

**Nucleotide correlation coefficient** The nucleotide correlation coefficient (nCC), first defined by Burset and Guigó (1996), and later also used by Tompa et al. (2005), is the MCC applied on the nucleotide-level statistics. The nCC is given by

$$\text{nCC} = \frac{nTP \cdot nTN - nFP \cdot nFN}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}}. \quad (16.1)$$

The nCC, like the general MCC, assumes values between -1 and +1, where a coefficient of 1 implies perfect prediction, and -1 a perfect inverse prediction. A coefficient of 0 implies that performance is equivalent to that of a random prediction.

For the limit of any of the factors of the product under the square root in the denominator approaching zero, the limiting value of the MCC is zero, see appendix E.

### 16.1.2 Site level performance metrics

**Basic site level statistics** Valen et al. (2009) give the following definitions for basic binding site level statistics, see figure 16.2.

*sTP* Number of real sites that share at least 50% of their nucleotides with a predicted site

*sFP* Number of predicted sites that share less than 50% of their nucleotides with a real site

*sFN* Number of real sites that share less than 50% of their nucleotides with a predicted site

These definitions are more strict than the site-level statistics given by Tompa et al. (2005), which consider a single overlapping base as sufficient. Using these basic statistics, one may define the following site-level performance metrics. First, the site sensitivity (sSn) is defined as

$$\text{sSn} = \frac{sTP}{sTP + sFN}. \quad (16.2)$$

Next, the site positive predictive value (sPPV) is defined as

$$\text{sPPV} = \frac{sTP}{sTP + sFP}. \quad (16.3)$$

Sensitivity is also known as recall, while another name for positive predictive value is precision.

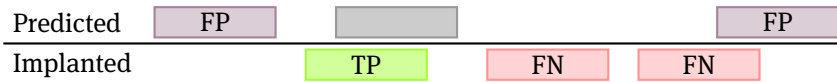


Figure 16.2: Classification of implanted binding sites as true positives (TP) and false negatives (FN), based on  $\geq 50\%$  overlap with predicted sites. Classification of predicted sites as false positives (FP) based on lack of  $\geq 50\%$  overlap with implanted sites.

**Average site performance** The average site performance (sAP) is the arithmetic mean of site sensitivity sSn and of site positive predictive value sPPV:

$$sAP = \frac{sSn + sPPV}{2}. \quad (16.4)$$

As both sSn and sPPV represent relative frequencies, they take on values between 0 and 1, where in both cases 0 signifies the worst performance. For the sSn a value of 1 denotes that each true site shares at least 50% of its nucleotides with a predicted site, which is equivalent to there being no false negative sites. In case of the sPPV a value of 1 denotes that none of the predicted sites is a false positive. Clearly, as the sAP is the average of these two measures, its values are confined to the same range.

**Site  $F_1$  score** Another choice of site-level statistic is the site  $F_1$  score ( $sF_1$ ). It is related to the average site performance, as it is the harmonic mean of site sensitivity and site positive predictive value:

$$sF_1 = 2 \frac{sSn \cdot sPPV}{sSn + sPPV}. \quad (16.5)$$

Like the sAP, the  $sF_1$  is symmetrical in the underlying statistics, assumes values between 0 and 1, and assumes its maximal value exactly when both underlying statistics achieve their maxima. However, unlike the sAP it is zero whenever either underlying statistic is zero. Whenever sSn is not equal to sPPV, the  $sF_1$  is less than the sAP, otherwise they are equal.

## 16.2 Summarization

Another topic worth discussion in the context of performance metrics is that of summarization. It is in principle possible to compute the nCC and sAP, and then study the distribution of these performance values over some variate of interest. However, discussion is eased by applying one of various ways of summarization.

Tompa et al. (2005) consider three summarization methods. These are ‘average’, ‘normalized’, and ‘combined’.

**Average** This method summarizes by computing the average value of the performance metric of interest for a variate of interest.

**Normalized** This method standardizes the performance of each experiment by subtracting the mean performance of all methods applied to this experiment and dividing

by the standard deviation of all methods' performance values. For summarization, these standardized scores are then averaged.

**Combined** This method unites the underlying statistics of the experiments to yield super-experiment statistics. From these super-experiment statistics the final performance values are computed.

Tompa et al. (2005) report few qualitative differences among these three methods of summarizing, except for that averaging of nCC and sAP tends to reward methods that make no prediction on many datasets. In light of this, in this work only the 'combined' method is used for summarization.



## Chapter 17

# Results on synthetic data

This chapter presents the results of the supervised performance evaluation of MD methods on the synthetic datasets. It first explains in section 17.1 which MD tools are included in the MD performance analysis. Then, it gives a high level summary of the performance of the considered MD methods by summarizing across all experiments in the three sets of experiments in section 17.2. Subsequently, section 17.3 analyzes more closely the performance of representative signal-only and discriminative learning methods as functions of the parameters varied within each set of experiments. Section 17.4 discusses the influence of significance filtering.

### 17.1 Evaluated motif discovery tools

In general, all MD methods were used to discover the single-best motif of length 8 nt. For methods, that do not allow to discover only one motif, the best motif was selected according to the score reported by the method.

As the synthetic datasets simulate RNA MD experiments, methods that allow it were instructed to consider only occurrences on the forward strand.

Motif occurrences were based on the methods' predictions, where available. Some of the published methods based on IUPAC regex motifs do not report motif occurrences. For these methods, motif occurrences were found by matching the regexes against the sequences.

**Discriminative motif discovery with Discover** Discover was run in single-motif discovery mode. The method consists of three parts: seed finding with Plasma (see section 11.2), HMM optimization by hybrid learning (see chapter 12), and significance filtering (see chapter 10). The following five discriminative objective functions implemented in Discover were considered:

1. Log likelihood difference (DLOGL)
2. Relative occurrence frequency difference (DFREQ)
3. Matthews correlation coefficient (MCC)

4. Maximum mutual information estimation (MMIE)
5. Mutual information of condition and motif occurrence (MICO)

Measures DFREQ, MCC, and MICO depend on statistics of counts of sequences that have at least one occurrence of the motif. Where during discrete optimization integer counts of sequences are used, optimization of HMMs involves expected counts of sequences. Thus DFREQ, MCC, and MICO can be used both for seeding and subsequent HMM optimization. DLOGL and MMIE may only be used in HMM optimization, as they are not applicable to the non-probabilistic nature of the seed finding method. However, DFREQ appears as a suitable objective for seeding HMMs to be optimized by DLOGL<sup>1</sup>. Similarly, for MMIE optimization of HMM parameters, MICO is used for seed finding.

**Published discriminative motif discovery tools** We intended to include in the supervised MD performance analysis all published DMD tools described in chapter 14. However, some tools (DEME, DIPS, and Dispom) were impractically slow on large datasets (see table H.1). Other tools (YMF, DISPARE, and the method of Barash, Bejerano, and N. Friedman (2001)) are not publicly available. These were excluded from analysis. This left the DMD methods CMF, DECOD, DME, DREME, FIRE, and MoAn, which were included in the performance comparison.

The methods CMF, DME, and MDscan do not support RNA motif analysis, and were thus run in DNA mode.

We found MoAn to be running relatively long (see table H.1), and therefore reduced to a tenth the number of iterations that MoAn performs.

**Signal-only motif discovery methods** We evaluated MD performance of Discover using signal-only learning with the Baum-Welch algorithm using as seeds the IUPAC regex motifs with degeneracy less or equal to 2 that are most frequent in the signal sequences. Additionally, we evaluated the performance of signal-only learning using the Baum-Welch algorithm after choosing discriminative seeds with MICO on signal and control data. These two approaches are respectively referred to as BW and BW-MICO.

The MD performance analysis also includes two published signal-only MD tools: Bio-Prospector (X. Liu, Brutlag, and J. S. Liu, 2001) and MDscan (X. S. Liu, Brutlag, and J. S. Liu, 2002).

**Command line options** The command line options and parameters used to run the tools are listed in appendix G.

## 17.2 Summary performance

High-level performance summary statistics were computed by summarizing across datasets (see section 16.2) the true and false site predictions, as well as true and false non-predictions. Figure 17.1 depicts the summarized nCC for the different MD tools in the three

---

<sup>1</sup>For HMM optimization with DLOGL we also tried MICO seeding; the results were similar.

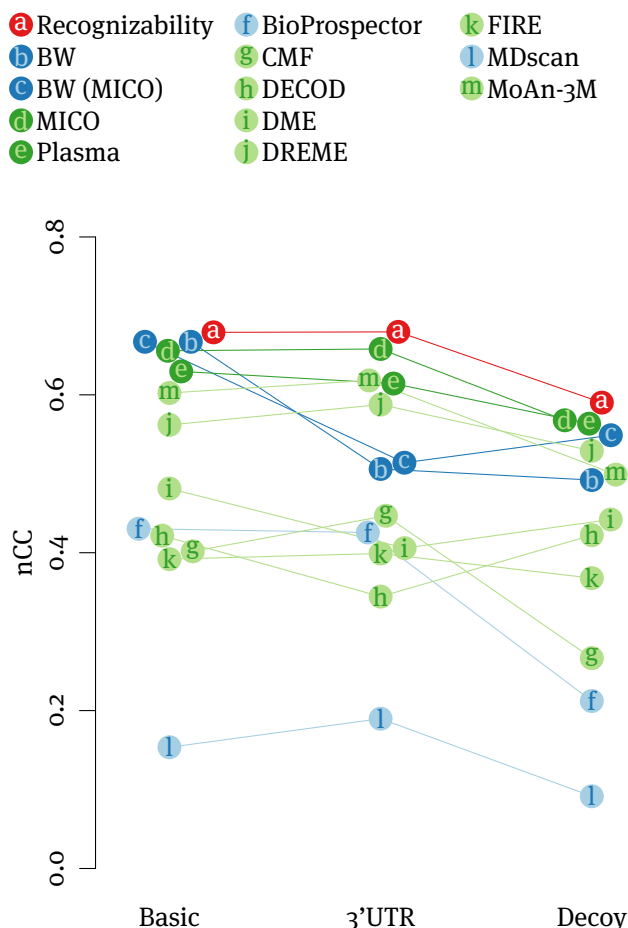


Figure 17.1: Summarized motif discovery performance of various methods on three synthetic datasets measured by the nucleotide-level Matthews correlation coefficient (nCC). Recognizability (red) serves as reference. Blue denotes signal-only motif learning methods, while green denotes discriminative motif discovery methods. Dark letters and light background denote published motif finding methods, light letters and dark background denote motif finding with objective functions implemented in Discover. BW: Baum-Welch training of HMMs seeded with the most frequent IUPAC motifs of degeneracy maximally 2, BW-MICO: Baum-Welch training of HMMs seeded with IUPAC motifs maximizing MICO. Plasma: IUPAC regex motif based seeding method of Discover, optimizing MICO as objective function. MoAn-3M: MoAn run with  $3 \times 10^6$  iterations.

sets of experiments, as well as the motif recognizability (see section 15.2 for the definition of recognizability). Table H.2 gives the corresponding numbers. Figure H.1 presents additional performance metrics, including sSn, sPPV, and sAP. More MD performance analyses for the same methods but with different parameters are shown in figure H.2.

Recognizability nCC indeed is larger than all methods' nCC in all three sets of experiments. The MD performance of Discover with MICO as objective function surpassed that of all other methods and is only marginally below recognizability in all three sets of experiments. It consistently achieved nCC of greater 96 % of recognizability (table H.2).

Plasma was the second-best performing MD method. At 90–96% of motif recognizability (table H.2) its MD performance was lower than when subsequently optimizing HMM parameters with Discover but surpasses that of most previously published DMD methods.

Other methods are close to recognizability in some but not all of the sets of experiments, including signal-only learning with the BW algorithm in the basic set of experiments.

To separate the influence of objective function choice in seeding from that in HMM parameter optimization, we used MICO to determine seeds that were used for signal-only learning of HMM parameters with the BW algorithm (BW-MICO in figure 17.1). This alleviated some of the problems that signal-only learning with signal-only seeding (BW in figure 17.1) has on the decoy experiments, but did not remedy those on the 3'UTR experiments.

**Motif discovery performance of published methods** DREME and MoAn were the best-performing previously published DMD methods and respectively achieved nCC of 83–90% and 85–91% relative to recognizability (table H.2). Note that the MoAn MD performance shown in figure 17.1 is based on the reduced number of iterations. We also evaluated MoAn's performance with the default number of iterations on the basic and 3'UTR experiments, and this further increased the performance of MoAn (figure H.2), yet not to the same level as that of Discover. Other published MD methods achieved lower performance. The performance of DME and DECOD is intermediate of the published discriminative methods. DME and DECOD show similar performance profiles across the experiments, with DME performing slightly better than DECOD due to yielding higher sPPV.

The low performance of CMF seems to be caused by overly eagerly accepting binding sites. This is evidenced by the highest average sSn over the three sets of experiments, and simultaneously the lowest average sPPV. Perhaps stringent filtering of results might help solve this problem.

**Strandedness and motif analysis** It should be noted that CMF, DME, and MDscan currently only support double-stranded DNA mode, and might exhibit higher MD performance on these experiments by accounting for the larger RNA motif space<sup>2</sup>. However, for DREME we made an observation contradicting this expectation. An initial evaluation

<sup>2</sup>DNA motif models treat non-palindromic sequences as equivalent to their reverse complement, while RNA motif models distinguish between them; thus, the space of DNA motifs of size  $n$  is only roughly half as large as the space of RNA motifs of size  $n$ .

of MD performance used the then-current version 4.7.0 of DREME, which did not support single-strand MD that is appropriate for the analysis of RNA datasets as simulated in the synthetic data. The current version 4.9.0 fixed this, but MD performance did not improve much (figure H.2 and table H.2).

**Runtime of motif finding methods** There is considerable variation in the time that the different methods need to analyze the synthetic datasets. The runtimes of the methods included in the analysis varied more than 360 fold between the fastest and slowest methods (figure H.3).

The fastest method was Plasma, the IUPAC regex based seeding method of Discoverer. It required 2.6 hours to process all synthetic data experiments.

The slowest method was MoAn, which achieved—together with DREME—the best MD performance of previously published methods. In spite of the reduced number of iterations, MoAn still required about 960 hours to analyze all synthetic datasets.

DREME needed a total of 8.9 hours, and Discoverer 16.9 hours. With one exception, DME, the other published MD methods had higher runtimes.

As mentioned, three methods (Dispom, DIPS, and DEME) had to be excluded as their runtime prohibited a full evaluation of the synthetic datasets. Instead, we exemplarily determined their runtime on one single pair of signal and control sequence sets (table H.1). This dataset was analyzed in about two minutes by Discoverer. Dispom needed 40 hours, DIPS more than 600 hours, or more than 25 days, and DEME did not finish in 74 days.

## 17.3 Comparing signal-only and discriminative learning

To illustrate the benefits of using negative examples in MD, we next compare the relative merits of discriminative learning and of signal-only, generative learning. Thus, we analyze the motif discoverability of representative objective functions, MICO and likelihood, as a function of the variables controlled in the experiments, and compare to motif recognizability. To emphasize differences between signal-only and discriminative learning, we use as seeds for HMM initialization for signal-only, generative learning the most frequent words of length 8 in the signal sequences, allowing up to two alternative nucleotides at any position (e.g. ACGTMSGT = ACGT[AC][CG]GT), i.e. BW in figure 17.1. For discriminative learning seeds are chosen by heuristic optimization of MICO over the space of IUPAC regexes as implemented in Plasma and described in section 11.2. Models found by both learning approaches are subjected to discriminative significance filtering based on MICO.

Figure 17.2 displays the nCC based on statistics broken down by variates, computed by summing over the other variates.

**Basic experiments** As figure 17.2a shows, recognizability decreases in the basic experiments with increasing sequence context size, with decreasing implantation frequency, and with decreasing IC. Throughout most combinations of the varied parameters, the MD

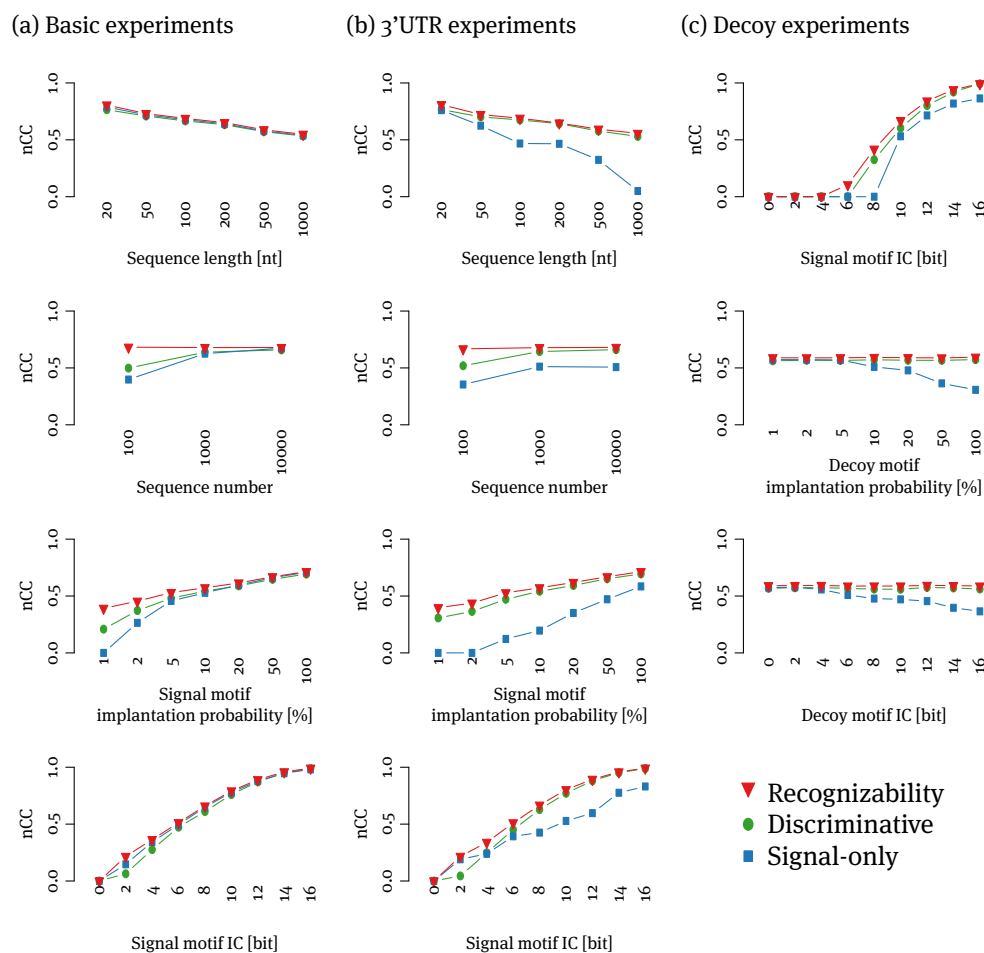


Figure 17.2: Motif recognizability and discovery performance on synthetic data in the (a) basic, (b) 3'UTR, and (c) decoy experiments. Recognizability and discovery performance are measured by the nucleotide-level Matthews correlation coefficient (nCC) as a function of different variates, summarized over the remaining variates. Recognizability (red) serves as reference. Signal-only learning (blue, BW in figure 17.1) was performed with the Baum-Welch algorithm on the signal data only, and used as seeds the 8mers of degeneracy at most 2 that are most frequent in the signal data. Discriminative learning (green, MICO in figure 17.1) used MICO as objective function for seed finding and HMM parameter optimization. See figures H.4 to H.6 for site-level average performance, sensitivity, and positive predictive value.

performance of both signal-only and discriminative learning is very close to motif recognizability. Where the points in the graphs are overlapping, motif recognizability is always (albeit at times only minutely) larger than the methods' motif discoverability. While the differences in MD performance are small between signal-only and discriminative learning on the basic experiments, they are slightly in favor of signal-only learning. The largest reductions relative to motif recognizability are seen when data is limited to 100 sequences. Increasing sequence numbers to 1000 yields a MD nCC close to motif recognizability, which increases further when 10 000 sequences are available. Some deficits relative to motif recognizability are also seen for motifs of very low implantation frequency or very low IC.

**3'UTR experiments** Motif recognizability and MD performance of discriminative learning in the experiments based on real human 3'UTR (figure 17.2b) are virtually identical to those in the basic experiments. The MD performance of signal-only learning is, however, negatively impacted in nearly all variate combinations by the higher complexity sequence background. Even 1000 or more sequences are not sufficient to yield nCC close to motif recognizability for signal-only learning. That nCC does not further approximate motif recognizability when 10 000 sequences are available demonstrates that signal-only learning is genuinely confused by the characteristics of real 3'UTR sequences.

**Decoy experiments** In the decoy experiments (figure 17.2c), (signal) motif recognizability varies in response to signal motif IC but is unaffected by variation of decoy motif implantation frequency or IC. MD performance of discriminative learning is also unaffected by increasing implantation frequency or IC of decoy motifs, and generally close to motif recognizability. MD performance of signal-only learning is deteriorating in response to the increasing potential likelihood contribution of decoy motifs. As implantation frequency of the signal motif was fixed to 10% in these experiments, a phase transition is visible at decoy motif implantation frequency 10% between little or no, and strong negative influence of the decoy motif on the discovery performance of signal-only learning. Similarly, decoy motifs of low IC do not strongly affect signal-only learning's MD performance, while higher IC decoy motifs lead to reduced MD performance.

**Sensitivity and positive predictive value** Figures H.4 to H.6 give sAP, sSn, and sPPV summarized in the same manner as nCC in figure 17.2. Note that the reason for missing values for sPPV and for sAP in some variate combinations is that sPPV is not defined when not a single motif occurrence is predicted. It is apparent that sPPV is generally higher than sSn. sPPV of Discoverer with MICO as objective function is frequently higher than sPPV of recognizability, while the converse holds for sSn. Another feature that is apparent from figures H.5a and H.5b is that motif recognizability sSn exhibits a sigmoidal response to changes in IC in the basic and 3'UTR experiments<sup>3</sup>.

<sup>3</sup>The same observation also holds for sSn of signal-only learning in the basic experiments, as well as for sSn of discriminative learning in the basic and 3'UTR experiments

## 17.4 Discriminative filtering

Significance filtering has a large effect on MD performance by reducing the number of predictions of motif occurrences when insufficient evidence is available. In section 17.3, discriminative significance filtering was applied for all objective functions implemented in Discover, including the signal-only learning approaches. To illustrate the effect of significance filtering, figure H.7 compares MD performance metrics with and without applying significance filtering. As is evident, non-discriminative MD is much improved by discriminative significance filtering based on MICO.



## Chapter 18

# PUF RNA-binding protein family

The PUF<sup>1</sup> proteins are a widespread, conserved family of eukaryotic RBPs that bind 3'UTRs and modulate expression of mRNAs (Wickens et al., 2002).

In *S. cerevisiae*, the Puf proteins control aging, mitochondrial function and mating-type switching (Kennedy et al., 1997; Olivas and Parker, 2000; Tadauchi et al., 2001). In early *D. melanogaster* embryos, Pumilio controls anterior-posterior development by repressing hunchback mRNA (Barker et al., 1992; Murata and Wharton, 1995). In *C. elegans*, the developmental switch from male to female is regulated by the translational repression of FEM-3 resulting from the binding of the PUF homolog FBF-1<sup>2</sup> to the 3'UTR of FEM-3 (B. Zhang et al., 1997). Double knock-out of FBF-1 and FBF-2 leads to loss of germline stem cells via de-repression of the meiosis-promoting GLD-1 (Crittenden et al., 2002).

Two PUF homologs exist in mammalian genomes. Consistent with roles in regulating germline biology, knockout of the mouse Pum2 gene results in smaller testis, although no reduction in fertility is observed (E. Y. Xu et al., 2007). The human PUM2 interacts with RBPs that function in early germline stem cells, such as deleted in azoospermia (DAZ), DAZ-like (DAZL) proteins, and the meiotic regulator BOULE (BOL) (M. Fox, Urano, and Reijo Pera, 2005; Moore et al., 2003; Urano, M. S. Fox, and Reijo Pera, 2005). Among the RIP-Chip-determined targets of human PUM1 and PUM2 are many DBPs acting as transcriptional regulators of cell cycle control and RBPs acting as post-transcriptional regulators (Galgano et al., 2008; Morris, Mukherjee, and Keene, 2008). The PUM1 and PUM2 targets also include PUM1 and PUM2 themselves (Galgano et al., 2008; Morris, Mukherjee, and Keene, 2008). Binding sites of the human PUF proteins are enriched in the vicinity of miRNA binding sites, indicating that PUF proteins and miRNAs interact in post-transcriptional regulation (Galgano et al., 2008; P. Jiang, M. Singh, and Collier, 2013; Kedde et al., 2010; Leibovich, Mandel-Gutfreund, and Yakhini, 2010; Miles et al., 2012; Triboulet and Gregory, 2010).

In this chapter, we apply DMD methods to datasets of the PUF family of RBPs in different species and coming from different technologies. In section 18.1 the analyzed data is presented, including data sources, which sequences are used, and how contrasts for discriminative learning are set up. Subsequently, section 18.2 outlines the analyses applied

---

<sup>1</sup>Named after the first studied members of the family: Pumilio and FBF-1.

<sup>2</sup>Fem-3 binding factor 1

to the PUF RBP family datasets. Section 18.3 discusses the motifs found by discriminative learning. We compare the results of applying MICO, MMIE, and ML learning principles in section 18.4. Section 18.5 presents the results of a dilution analysis. The chapter concludes with a word-based analysis in section 18.5.2.

## 18.1 Materials

**Data sources** We analyze various dataset for the PUF family of RBPs in different species, including *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *H. sapiens*. RIP-Chip was used to define targets of Puf1p, Puf2p, Puf3p, Puf4p, and Puf5p in yeast (Gerber, Herschlag, and Brown, 2004), of Pumilio in adult fly ovaries (Gerber, Luschnig, et al., 2006), of the worm homolog FBF-1 (Kershner and Kimble, 2010), and of human PUM1 (Galgano et al., 2008; Morris, Mukherjee, and Keene, 2008) and PUM2 (Galgano et al., 2008) from HeLa S3 cells. Additionally, we analyze PAR-CLIP data of human PUM2 (Hafner et al., 2010) from HEK293 cells.

**Sequence definition** Due to the lack of finer spatial resolution entire 3'UTR sequences are used for the RIP-Chip data. When probe sets map to multiple transcripts, the longest 3'UTR of the sequences is used. Yeast 3'UTRs are based on Nagalakshmi et al. (2008). Fly 3'UTRs are retrieved from FlyBase, version 5.48. Worm 3'UTRs are retrieved from WormMart, version WS220. Human 3'UTRs are retrieved from Ensembl release 70 (GRCh37.p10) via Ensembl's Biomart and from RefSeq NCBI36.1 / hg18 via the UCSC genome browser. Human 3'UTRs for the RIP-Chip data are retrieved by Ensembl transcripts IDs and RefSeq transcript ID (Galgano et al., 2008), or only by RefSeq transcript ID (Morris, Mukherjee, and Keene, 2008).

For PAR-CLIP data, read covered regions are used, as available from the Dorina data base (Anders et al., 2012).

**Contrasts for discriminative learning** For the yeast analysis, binary contrasts are considered, where the 3'UTR sequences of each Puf protein's target genes serve as signal set, and the controls comprise all yeast 3'UTRs not part of the signal set.

For the worm FBF-1 we retrieved a table of bound target genes from the supplementary material of Kershner and Kimble (2010). We mapped WormBase gene IDs to transcript IDs, and retrieved the 3'UTR sequences of all transcripts associated to each target gene, keeping only the longest for each. To set up contrasts for the FBF-1 data, we follow the analysis of Kershner and Kimble (2010) who split the data into 15 approximately equally sized rank groups. Thus, we split up the 3294 3'UTR sequences of target genes by rank into 14 groups of 220 and one group of size 214.

Target genes of the fly RBP Pumilio with estimated FDR < 1% are tabulated in the supplementary material of Gerber, Luschnig, et al. (2006). We translated the FlyBase gene IDs of the target genes in this table to transcript IDs. As signal sequences, for each target gene we choose the longest available 3'UTR sequence over all associated transcripts. As control sequences we choose the longest 3'UTR sequence for each non-target gene.

The supplementary material of Morris, Mukherjee, and Keene (2008) provides a table with LOD scores for binding of PUM1 to genes. Multiple RefSeq transcript IDs may be associated with each gene locus, and we use the longest associated 3'UTR sequence for each gene locus. Differently from the other RIP-Chip data that we use, this table also includes gene loci for which there is no evidence of binding. Thus, following Morris, Mukherjee, and Keene (2008) we use as signal data the 3'UTR sequences of gene loci with an LOD greater zero, and the remainder as control.

The supplementary materials of Galgano et al. (2008) provide tables for targets bound by PUM1 and PUM2. For each table entry multiple Ensembl and RefSeq transcript IDs may be given. We retrieved the corresponding 3'UTR sequences, and use the longest one for each entry of the table. We procured control data, by taking the set of Ensembl 3'UTR sequences complementary to those mentioned in the bound targets tables.

For the PAR-CLIP data of Hafner et al. (2010) we consider dinucleotide distribution conserving shuffles of the signal sequences as controls.

## 18.2 Methods

**Discriminative motif discovery** We perform DMD with MICO as objective function on the PUF RBP family datasets. As detailed in section 18.1, 3'UTR sequences are used for the RIP-Chip data, and covered regions for PAR-CLIP. Discriminative analysis employs shuffled sequences as control in the case of PAR-CLIP data, and 3'UTR sequences of non-target genes for the RIP-Chip data. For each dataset the most discriminative IUPAC word according to MICO is determined for each length of 7–12 nt using algorithm 5 implemented in Plasma, described in section 11.2. HMMs are seeded on each of these and parameters optimized, maximizing MICO. For each RBP, we select the HMM motif with the best corrected  $p$ -value.

**Dilution analysis of PUM2 PAR-CLIP data** We perform a dilution analysis of the PUM2 PAR-CLIP data by embedding the real sequences in increasingly larger, synthetically generated sequence context. Specifically, for each signal dataset, we extract dinucleotide frequencies, and generate for each signal sequence shorter than the desired length flanks of the appropriate size, as well as a control sequence of equal total size. We thus analyze the original PUM2 data (mean length 35.0 nt), as well as variants embedded to minimum lengths of 64, 128, 256, 512, and 1024 nt. Then, for the objective functions MICO and MMIE, we optimize parameters of HMMs seeded on NNUGUANAUANN.

**Word-based discriminative analysis** We perform two word-based discriminative analyses for the human members of the PUF RBP family PUM1 and PUM2 datasets of Galgano et al. (2008), Hafner et al. (2010), and Morris, Mukherjee, and Keene (2008). We consider scatterplots of relative occurrence frequencies of sequences that have at least one occurrence of a given word of length 8. Also, we apply algorithm 4 implemented in corenmers, described in section 11.1, in order to determine the top 50 words of length 8 according to residual MICO.

Table 18.1: Discriminative motif analysis of the PUF RBPs family with MICO as objective function. Shown are motifs of 7–12 nt with lowest corrected  $p$ -value. Data sources: <sup>A</sup> Gerber, Herschlag, and Brown (2004), <sup>B</sup> Kershner and Kimble (2010), <sup>C</sup> Gerber, Luschnig, et al. (2006), <sup>D</sup> Morris, Mukherjee, and Keene (2008), <sup>E</sup> Galgano et al. (2008), <sup>F</sup> Hafner et al. (2010).

Protein	Species	Technology	MICO motif
Puf1p <sup>A</sup>	<i>S. cerevisiae</i>	RIP-Chip	UAAU <u>AAU</u>
Puf2p <sup>A</sup>	<i>S. cerevisiae</i>	RIP-Chip	UAAU <u>AAU</u>
Puf3p <sup>A</sup>	<i>S. cerevisiae</i>	RIP-Chip	C <u>UGUA</u> AAUA
Puf4p <sup>A</sup>	<i>S. cerevisiae</i>	RIP-Chip	UGUA <u>AAUA</u>
Puf5p <sup>A</sup>	<i>S. cerevisiae</i>	RIP-Chip	UGUA <u>AAUAUA</u>
FBF-1 <sup>B</sup>	<i>C. elegans</i>	RIP-Chip	UGUA <u>AAU</u>
Pum <sup>C</sup>	<i>D. melanogaster</i>	RIP-Chip	UGUA <u>AAUA</u>
PUM1 <sup>D</sup>	<i>H. sapiens</i>	RIP-Chip	UGUA <u>AAUA</u>
PUM1 <sup>E</sup>	<i>H. sapiens</i>	RIP-Chip	UGUA <u>AAUA</u>
PUM2 <sup>E</sup>	<i>H. sapiens</i>	RIP-Chip	UGUA <u>AAUA</u>
PUM2 <sup>F</sup>	<i>H. sapiens</i>	PAR-CLIP	UGUA <u>AAUA</u>

### 18.3 Discriminative motifs in PUF RBP data

The motifs resulting from discriminative analysis are summarized in table 18.1, and details are presented in table 18.2. All identified motifs resemble published motifs of the respective factors and are, except for the cases of Puf1p and Puf2p, similar to the known PUF recognition element, UGUAAUA (PRE) (Galgano et al., 2008; Hafner et al., 2010; Morris, Mukherjee, and Keene, 2008; X. Wang, McLachlan, et al., 2002). Most conserved is the specificity of the first four positions, but variability is seen in the second part, and in the context. Puf1p and Puf2p exhibit a motif very unlike that of the other RBPs analyzed here, and the found discriminative motifs, consisting of two UAAU separated by three less defined positions, are consistent with previous reports by Gerber, Herschlag, and Brown (2004) and Yosefzon et al. (2011). Puf3p exhibits preference for a C two positions upstream. Puf4p appears to favor a 9 nt variant, and Puf5p a 10 nt variant. Also, FBF-1 appears to favor a 9 nt variant of the motif.

As is visible from the details in table 18.2a, all but the motif for Puf1p are deemed significantly discriminative according to the corrected  $p$ -value that we calculate. In the case of Puf1p, there is a considerable relative enrichment of 40.8% signal over 0.7% control sequences that have at least one motif occurrence but there are only 32 signal sequences in this dataset. Given the relatively large search space size of motifs length 7–12 nt, the observed enrichment is insufficient to meet the threshold for significance. Nevertheless, the identified motif is the same as what has previously been reported (Gerber, Herschlag, and Brown, 2004; Yosefzon et al., 2011).

Discriminative analysis of the datasets of the fly *Pumilio* and of human PUM1 and PUM2 RIP-Chip data uniformly yield UGUAAUA as most discriminative motif, while analysis of the PUM2 PAR-CLIP data yields a more diffuse affinity towards A/U on positions 7 to 10 (counting from the beginning of UGUAAWHH).

Table 18.2: Motif discovery results for RIP-Chip and PAR-CLIP data of PUF RBP family members.  $N_S$  and  $N_C$ : number of sequences in signal and control sequence sets.  $\log-L$ : log-likelihood of signal data.  $S$  and  $C$ : relative frequency of signal and control sequences with at least one motif occurrence.  $\log-p$ : MICO-based MT-corrected  $\log-p$  value. MMIE: log probability of correctly classifying all samples (equations (9.6) and (9.15)). (a): discriminative motif analysis with MICO. MICO is used to find seeds of length 7–12 nt, and to optimize HMM parameters. Motifs selected by  $\log-p$ . (b): discriminative motif analysis with MMIE. MICO is used to find seeds, and HMM parameters are optimized by MMIE. Motifs selected by MMIE score. (c): Baum-Welch algorithm applied to the seed NNUGUA-NAUANN. Data sources: <sup>A</sup> Gerber, Herschlag, and Brown (2004), <sup>B</sup> Kershner and Kimble (2010), <sup>C</sup> Gerber, Luschnig, et al. (2006), <sup>D</sup> Morris, Mukherjee, and Keene (2008), <sup>E</sup> Galgano et al. (2008), <sup>F</sup> Hafner et al. (2010).

(a) MICO

Protein	$N_S$	$N_C$	Motif	$\log-L$	$S$ [%]	$C$ [%]	MICO [bit]	$\log-p$
Puf1p <sup>A</sup>	32	5180	UAAU~UAAU	-6862	40.8	0.7	64.3	0
Puf2p <sup>A</sup>	124	5088	UAAU~UAAU	-28 993	33.9	1.0	150.6	-47.3
Puf3p <sup>A</sup>	68	5144	CUGUA~AUA	-10 184	52.1	3.5	103.7	-24.6
Puf4p <sup>A</sup>	184	5028	UGUA~AUA	-32 049	47.8	4.2	207.4	-101.9
Puf5p <sup>A</sup>	156	5056	UGUA~AUA	-32 416	35.7	2.3	146.6	-49.5
FBF-1 <sup>B</sup>	3294	10 096	UGU~AU	-970 238	20.9	5.2	462.4	-264.0
Pum <sup>C</sup>	834	12 135	UGUA~AUA	-780 326	50.7	12.9	448.6	-274.4
PUM1 <sup>D</sup>	1401	18 651	UGUA~AUA	-3 515 930	49.8	5.3	1375.8	-917.7
PUM1 <sup>E</sup>	836	6320	UGUA~AUA	-2 094 000	61.6	13.1	638.3	-406.1
PUM2 <sup>E</sup>	565	19 535	UGUA~AUA	-1 372 030	56.1	6.2	687.9	-440.5
PUM2 <sup>F</sup>	6916	6916	UGUA~AUA	-327 370	54.1	10.8	2269.7	-1517.7

(b) MMIE

Protein	$N_S$	$N_C$	Motif	$\log-L$	$S$ [%]	$C$ [%]	MICO [bit]	$\log-p$	MMIE
Puf1p <sup>A</sup>	32	5180	UAAUA~UAAU	-6876	38.7	0.8	58.3	0	-145
Puf2p <sup>A</sup>	124	5088	UAAU~UAAU	-29 003	36.2	2.0	130.5	-33.2	-464
Puf3p <sup>A</sup>	68	5144	CUGUAAUA	-10 221	49.3	2.8	104.4	-15.0	-276
Puf4p <sup>A</sup>	184	5028	UGUA~AUA	-32 077	45.6	3.5	208.7	-87.7	-611
Puf5p <sup>A</sup>	156	5056	UGUA~AUA	-32 421	41.4	4.2	139.6	-39.6	-564
FBF-1 <sup>B</sup>	3294	10 096	UGU~AU	-970 083	38.9	17.9	415.9	-246.7	-7005
Pum <sup>C</sup>	834	12 135	UGUA~AUA	-780 320	54.7	15.7	441.9	-249.7	-2695
PUM1 <sup>D</sup>	1401	18 651	UGUA~AUA	-3 515 800	59.1	9.0	1386.8	-905.3	-3776
PUM1 <sup>E</sup>	836	6320	UGUA~AUA	-2 093 960	65.8	16.4	618.2	-372.2	-1998
PUM2 <sup>E</sup>	565	19 535	UGUA~AUA	-1 372 080	58.3	7.9	646.1	-396.5	-1975
PUM2 <sup>F</sup>	6916	6916	UGUA~AUA	-326 888	60.5	16.6	2127.8	-1419.3	-7626

(c) Baum-Welch

Protein	$N_S$	$N_C$	Motif	$\log-L$	$S$ [%]	$C$ [%]	MICO [bit]	$\log-p$
Puf1p <sup>A</sup>	32	5180	UAAU~UAAU	-6771	88.9	82.4	0.3	0
Puf2p <sup>A</sup>	124	5088	UAAU~UAAU	-28 733	93.1	89.5	0.9	0
Puf3p <sup>A</sup>	68	5144	UUA~AUA	-10 127	85.3	68.7	6.4	0
Puf4p <sup>A</sup>	184	5028	UUA~AUA	-31 993	81.3	67.7	11.3	0
Puf5p <sup>A</sup>	156	5056	UUA~AUA	-32 377	35.3	38.1	0.3	0
FBF-1 <sup>B</sup>	3294	10 096	AA	-963 728	94.0	90.0	36.2	0
Pum <sup>C</sup>	834	12 135	UUA~AUA	-774 029	99.2	97.8	5.3	0
PUM1 <sup>D</sup>	1401	18 651	UUA~AUA	-3 489 370	99.2	96.2	32.2	0
PUM1 <sup>E</sup>	836	6320	UUA~AUA	-2 070 780	99.6	96.8	18.6	0
PUM2 <sup>E</sup>	565	19 535	UUA~AUA	-1 359 480	95.6	75.0	121.8	-27.2
PUM2 <sup>F</sup>	6916	6916	UUA~AUA	-324 488	82.0	53.8	934.2	-591.4

## 18.4 Comparing MICO, MMIE, and ML learning

To investigate the disparity of conclusions regarding the second half of the PUM2 motif between our analyses of RIP-Chip and PAR-CLIP data, we investigate the influence of the choice of objective function. We thus repeat the analysis of the PUF RBP family data using MMIE as objective function. We initialize HMMs with seeds of length 7–12 nt determined by MICO, optimize the HMM parameters for MMIE, and select the HMM with highest MMIE score. We treat the FBF-1 data slightly differently in this section: it is split only in 2 groups, rather than 15 groups as in the previous section<sup>3</sup>. The results are tabulated in table 18.2b.

We find that MMIE identifies longer variants of the MICO motifs which include positions with relatively low IC. This is because for MMIE we lack a comparable significance correction for motif length as is available for MICO.

Note that for one dataset, the PUM1 data of Morris, Mukherjee, and Keene (2008), optimization of MMIE yields a motif with higher MICO than that of the MICO motif of the same dataset. This is caused by selecting MICO motifs based on the MICO-based MT-corrected *p*-value rather than on MICO; however, the MICO motif's *p*-value is lower than that of the MMIE motif for this dataset due to the shorter length of the MICO motif.

The motifs identified by MMIE frequently have slightly lower IC than the MICO motifs (results not shown), and occur more frequently than those of MICO. This observation is in line with our findings on the synthetic datasets that MMIE yields higher sSn and lower sPPV than MICO (see section 17.2). Also for MMIE we observe a disparity between the RIP-Chip and PAR-CLIP analysis results regarding the second half of the motif.

To see what generative learning might yield, we apply the Baum-Welch algorithm to optimize the IUPAC seed sequence NNUGUANAUANN on the full PUF RBP family signal datasets. While this successfully determines models with higher likelihood than those for MICO or MMIE, it does not yield useful motifs, see table 18.2c. Although we use a seed similar to the motifs discovered by discriminative analysis in all but the Puf1p and Puf2p datasets, the PRE is mostly replaced by unspecific sequence. Only in the cases of Puf3p and Puf4p RIP-Chip data, and PUM2 PAR-CLIP data do some traces of the motif remain. However, the motifs incorporate so much background characteristic that for all except for the Puf5p datasets they occur in more than half of the respective control sequences, which are either random shuffles or all non-target 3'UTRs. For the Puf5p data, the motif is more frequent in the control data than in the signal data.

## 18.5 Dilution and word-based analyses










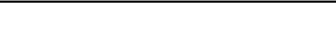
### 18.5.1 Dilution analysis

Next, because the covered regions of PAR-CLIP data are much shorter than the 3'UTR sequences used for the RIP-Chip data (figure I.1), we perform a dilution analysis of the PUM2 PAR-CLIP data by embedding the real sequences in increasingly larger, synthetically gen-











<sup>3</sup>We do this to give all three methods the same signal data for the purpose of this comparison; in particular, generative learning is supposed to only be applied to one signal dataset. Thus we chose to consider a binary contrast for the two discriminative methods.

Table 18.3: Dilution analysis of human PUM1 and PUM2 data for motifs found by optimizing (a) MICO and (b) MMIE. Sequences were embedded in increasing amounts of random sequences, varying from top to bottom. In the length column, numbers marked by  $\dagger$  give the average length of the original sequences, other numbers give the length to which the sequences were padded. HMMs are seeded on the IUPAC word NNUGUANAUANN and parameters optimized for MICO and MMIE, respectively. Data sources: <sup>A</sup> Hafner et al. (2010), <sup>B</sup> Galgano et al. (2008), <sup>C</sup> Morris, Mukherjee, and Keene (2008).

(a) MICO

Protein	Technology	Length [nt]	Motif
PUM2 <sup>A</sup>	PAR-CLIP	35.0 $\dagger$	
PUM2 <sup>A</sup>	PAR-CLIP	64	
PUM2 <sup>A</sup>	PAR-CLIP	128	
PUM2 <sup>A</sup>	PAR-CLIP	256	
PUM2 <sup>A</sup>	PAR-CLIP	512	
PUM2 <sup>A</sup>	PAR-CLIP	1024	
PUM2 <sup>A</sup>	PAR-CLIP	2048	
PUM2 <sup>B</sup>	RIP-Chip	1785.4 $\dagger$	
PUM1 <sup>B</sup>	RIP-Chip	1842.5 $\dagger$	
PUM1 <sup>C</sup>	RIP-Chip	1833.9 $\dagger$	

(b) MMIE

Protein	Technology	Length [nt]	Motif
PUM2 <sup>A</sup>	PAR-CLIP	35.0 $\dagger$	
PUM2 <sup>A</sup>	PAR-CLIP	64	
PUM2 <sup>A</sup>	PAR-CLIP	128	
PUM2 <sup>A</sup>	PAR-CLIP	256	
PUM2 <sup>A</sup>	PAR-CLIP	512	
PUM2 <sup>A</sup>	PAR-CLIP	1024	
PUM2 <sup>A</sup>	PAR-CLIP	2048	
PUM2 <sup>B</sup>	RIP-Chip	1785.4 $\dagger$	
PUM1 <sup>B</sup>	RIP-Chip	1842.5 $\dagger$	
PUM1 <sup>C</sup>	RIP-Chip	1833.9 $\dagger$	



erated sequence context. We thus analyze the original PUM2 data (mean length 35.0 nt), as well as variants padded to minimum lengths of 64, 128, 256, 512, and 1024 nt. Then, for the objective functions MICO and MMIE, we optimize parameters of HMMs seeded on NNUGUANAUANN, and the results are shown in table 18.3. We find that with increasing sequence context size the discriminative motif analyses of PAR-CLIP data embedded in random sequence become more alike to the results of our RIP-Chip data analyses. The dilution has the effect of yielding higher IC, particularly on the second half of the motif.

### 18.5.2 Word-based analyses

As another line of evidence we turn to word count analysis.

**Scatter plots of frequency of sequences with a given word** Scatter plots of frequencies of sequences that have a given word in figures I.3 to I.6 reveal that the longer sequence sizes of the RIP-Chip data compress the variability of word frequencies between signal and control sequences.

**Core-nmer analysis** In order to separate independent contributions due to central and neighboring words we employ a simple, progressive algorithm. We determine the top 50 words of length 8 on the human PUM1 and PUM2 datasets of Galgano et al. (2008), Hafner et al. (2010), and Morris, Mukherjee, and Keene (2008) according to residual MICO. Figure 18.1 shows these in the order produced by the Core-nmer algorithm, with the bars indicating how many of the sequences in signal (blue) and control (red) have at least one occurrence of the word. The light parts of the bars indicate which of the sequences have occurrences of any words with higher residual MICO, and thus are potentially already explained by them. It is thus the dark portions of the bars that indicate the novel explanatory contribution of a word when accepting words in decreasing order of residual MICO.

As is visible from figure 18.1, there is greater variety of UGUANNNN conforming words in the PAR-CLIP data among the 50 8mers with highest residual MICO. In contrast, only UGUAAAUA, UGUUAUA, and UGUACAUA appear to be strongly differential in the RIP-Chip data.

**Differences of explanatory contribution of UGUHAUA between technologies** The IUPAC motif UGUHAUA is occurring in more than half of the PUM1 and PUM2 RIP-Chip data signal sequences, and in less than 15% of corresponding control sequences (figure I.2). In contrast, for the PAR-CLIP data, this motif is only present in 19.7% and 2.3% of signal and control sequences, respectively.

### 18.5.3 Conclusion

Given the higher variety of weaker variants that independently contribute to MICO in the PAR-CLIP data, and that by diluting the PAR-CLIP data the results of DMD agree with those of the RIP-Chip data, we surmise that the smear is not observed in the RIP-Chip data due to the large length of 3'UTR sequences precluding discovery of weakly affine variants. It is likely that this shadowing of the weaker variants in the RIP-Chip data is a consequence of



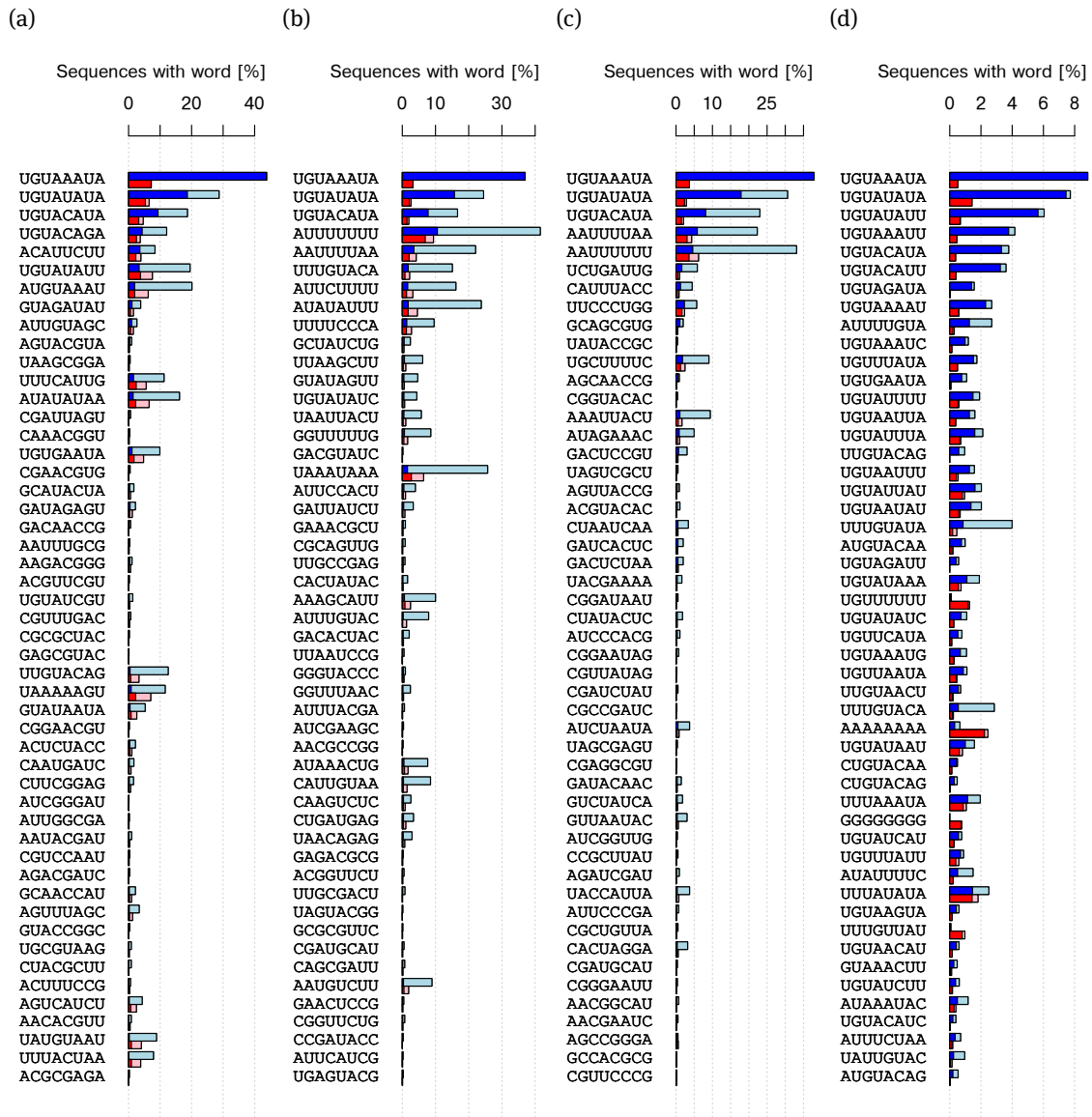


Figure 18.1: Discriminative word analysis of human RBP datasets of PUM1 Morris, Mukherjee, and Keene (2008) (a) and Galgano et al. (2008) (b), and PUM2 Galgano et al. (2008) (c), and Hafner et al. (2010) (d). Top 50 words of length 8 according to residual MICO as determined by algorithm 4. Bars indicate how many of the sequences in signal (blue) and control (red) have at least one occurrence of the word. Light parts of the bars indicate which of the sequences have occurrences of any words with higher MICO, and thus are potentially explained by them. It is thus the dark portions of the bars that indicate the novel explanatory contribution of a word when accepting words in decreasing order of MICO.

our choice of feature, that a sequence is considered a target if it has at least one occurrence of a motif.

## Chapter 19

# Alternative splicing regulator

## RBM10

RBM10 is an alternative splicing regulator (Y. Wang et al., 2013, that I co-authored) (Bechara et al., 2013; Inoue et al., 2014). Mutations of this gene are known to cause TARP syndrome<sup>1</sup> (Gripp et al., 2011; Johnston et al., 2010), and are also frequently found in lung adenocarcinomas (Imielinski et al., 2012).

RBM10 is present in splicing complexes (Rappsilber et al., 2002): the spliceosome A (Agafonov et al., 2011; Behzadnia et al., 2007; A. N. Kuhn et al., 2009) and B complexes (Agafonov et al., 2011; Deckert et al., 2006; A. N. Kuhn et al., 2009). It is a core protein of these complexes (Hegele et al., 2012), and interacts with spliceosome A complex proteins SF1 (branch-point-binding protein) and SF4, as well as with the U2-related proteins SR140 and DEAH helicase hPRP43 (Hegele et al., 2012).

RBM10 has four RNA-binding domains: two RRM domains, and RanBP2- and C2H2-type zinc finger domains. It may thus exhibit complex RNA-binding properties. SELEX<sup>2</sup>, fluorescence anisotropy, and NMR analysis showed that GST-fusion proteins carrying the RanBP2-type zinc finger domain of RBM10 bind *in vitro* to single-stranded RNA with the motif aGGUaa<sup>3</sup> (Loughlin et al., 2009; Nguyen et al., 2011). This sequence is almost identical to the conserved consensus sequence of metazoan 5' splice sites (Ast, 2004; M. Q. Zhang, 1998).

This chapter presents an application of the DMD methodology of Discover to study the *in vivo* sequence binding specificity of RBM10. Section 19.1 describes which data are used, section 19.2 how motif analysis is performed and presents the discovered motifs, and in section 19.3 it is shown that most previously published motifs for RBM10 are not corroborated by the data analyzed here.






---

<sup>1</sup>Talipes equinovarus, atrial septal defect, Robin sequence and persistent left superior vena cava, MIM #311900, an X-linked inherited disorder leading to multiple organ malformation in affected males

<sup>2</sup>Systematic evolution of ligands by exponential enrichment (Ellington and Szostak, 1990; Oliphant, Brandl, and Struhl, 1989; Tuerk and Gold, 1990)

<sup>3</sup>Lower-case letters denote less conserved positions.

Table 19.1: MICO motifs in RBM10 PAR-CLIP data from discriminative analysis versus shuffles.  $N_1$  and  $N_2$ : number of signal sequences in dataset 1 and 2, respectively. IC: information content. S and C: expected relative frequency of signal and control sequences with at least one motif occurrence. MICO: mutual information of condition and motif occurrence.  $\log-p$ : MICO-based  $\log-p$  value, corrected for motif length.

Sequences	$N_1$	$N_2$	Motif	PAR-CLIP dataset 1				PAR-CLIP dataset 2			
				IC [bit]	S [%]	C [%]	MICO [bit]	$\log-p$	S [%]	C [%]	MICO [bit]
Exonic	7469	22836		10.4	25.7	11.2	385.8–225.8	22.2	10.7	816.2–524.5	
				8.6	21.7	11.8	190.9–100.3	18.2	10.9	359.8–217.7	
				9.6	3.4	1.9	23.7	0.0	4.0	2.1	103.6 –29.5
Intronic	5908	21764		10.2	12.5	6.9	76.8	-5.8	12.6	6.9	289.3–153.7
				13.3	5.6	1.3	128.5	-41.9	3.5	1.0	222.7–107.4


## 19.1 Materials

GRCh37/hg19 coordinates were retrieved for binding sites of two RBM10 PAR-CLIP datasets of Y. Wang et al. (2013) via GEO<sup>4</sup>. These are defined as the positions with the highest number of cross-linking events within each of the PAR-CLIP clusters. All binding sites are considered that have at least 10 PAR-CLIP cross-linking events. Sequences of 41 nt are retrieved for each binding site by adding 20 nt flanks on each side. The sequences are split into two groups: those whose central position lies in exons, and among the rest those whose central position lies in introns. Exons and introns of RefSeq protein coding genes are considered.

## 19.2 Motif discovery for RBM10 reveals splicing-relevant motifs

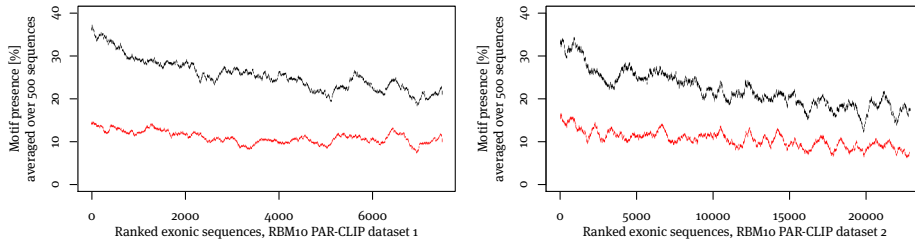
For RBM10 DMD is performed by jointly maximizing MICO across the contrasts given by the two datasets and their respective shuffles. Exonic and intronic sequences are independently analyzed. Using Plasma, the three most discriminative IUPAC regexes are identified for each length of 5–10 nt. Discover is run in multiple motif discovery mode, using the seeds reported by Plasma, as well as 1-nt-shifted variants of each. In total,  $6 \times 3 \times 3 = 54$  seeds are considered for each analysis.

**Discovered motifs** The analysis of the RBM10 PAR-CLIP data reveals three motifs for the exonic sequences, and two for the intronic ones, see table 19.1.

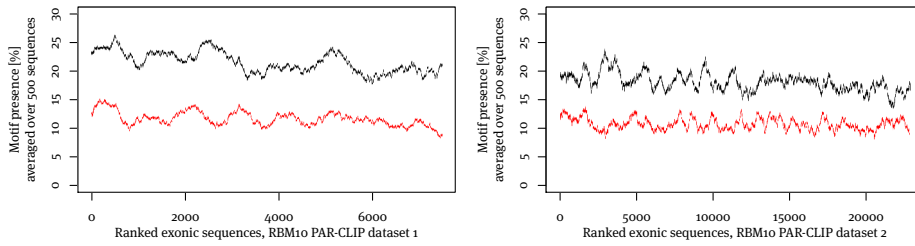
The motif , a known exonic splicing enhancer (ESE) signal (Caputi et al., 1994; Fairbrother, Yeh, et al., 2002), is the most differential one within the exonic sequences. This motif is bound by SFRS1 (Ramchatesingh et al., 1995; Sanford et al., 2009; Tacke and Manley, 1995; X. Wang, Juan, et al., 2011) and by eIF4AIII (Saulière et al., 2012). Occurrence of the motif is positively correlated with the number of PAR-CLIP conversions in the sequence (figure 19.1a), with the motif occurring in  $\geq 30\%$  of the sequences with most conversions.

<sup>4</sup>Datasets available by GEO accession GSM1095142 and GSM1095143.

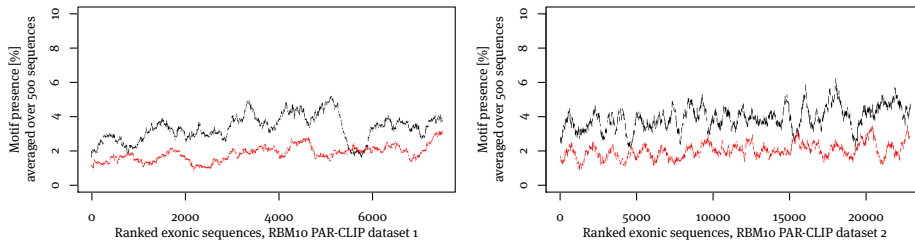
(a) RBM10 exonic motif **GAAGA** across exonic sequences



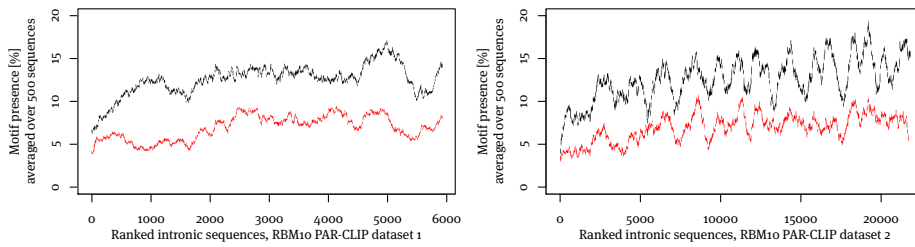
(b) RBM10 exonic motif **UGGA** across exonic sequences



(c) RBM10 exonic motif **CUCC** across exonic sequences



(d) RBM10 intronic motif **UUU** across intronic sequences



(e) RBM10 intronic motif **CACUG** across intronic sequences

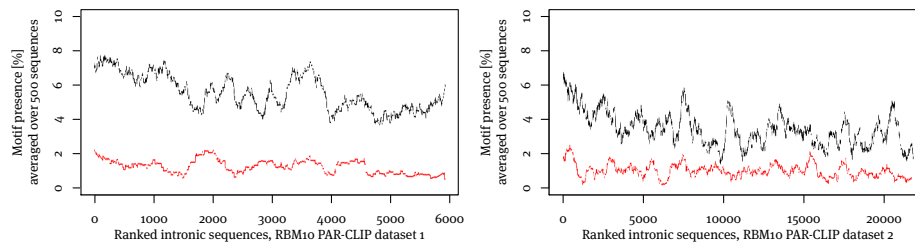



Figure 19.1: Occurrences of RBM10 motifs across ranked sequences. (a)–(c) Exonic motifs in exonic sequences. (d),(e) Intronic motifs in intronic sequences. Sequences are ranked by the number of PAR-CLIP conversions (conversions decreasing left to right).

Table 19.2: Summary statistics for enrichment in the PAR-CLIP data of Y. Wang et al. (2013) of the 94 CLIP-Seq-based RBM10 motifs reported by Bechara et al. (2013). See tables J.1 and J.2 for details. Absolute and relative numbers of RBM10 motifs that are less or equally frequent ( $\leq$ ), more frequent ( $>$ ), or much more frequent ( $\gg$ ) in the signal sequences compared to shuffled sequences. Motifs are counted as much more frequent if their relative frequency is higher in signal than control and the MICO-based  $\log$ - $p$  value is less than or equal to  $-10$ .

Data set	$S \leq C$		$S > C$		$S \gg C$	
		[%]		[%]		[%]
Exonic 1	45	47.9	49	52.1	9	9.6
Exonic 2	49	52.1	45	47.9	14	14.9
Intronic 1	49	52.1	45	47.9	2	2.1
Intronic 2	54	57.4	40	42.6	10	10.6

The most differential motif in the intronic sequences, , resembles the signal of the pyrimidine-rich tract. Unlike the ESE motif in the exonic sequences, however, this motif is negatively correlated with PAR-CLIP conversions (figure 19.1d).

The third exonic motif is remarkably reverse-complementary to the ESE motif. Like the one found in the intronic sequences, it is pyrimidine-rich and also negatively correlated with PAR-CLIP conversions (figure 19.1c). Its occurrence frequency in the exonic sequences is lower than that of the intronic pyrimidine-rich motif in the intronic sequences.

The second motif discovered in the intronic sequences is a palindromic 9mer with consensus CCACNGUGG which has not previously been described, and whose occurrence is positively correlated with PAR-CLIP conversions (figure 19.1e).

### 19.3 Previously reported RBM10 motifs not corroborated

**Most RBM10 motifs of Bechara et al. (2013) unsupported by PAR-CLIP data** A recent publication performed CLIP-Seq of RBM10 (Bechara et al., 2013). Their analysis of words enriched in the CLIP-Seq data yielded a set of 9 groups of 5mers, 94 words in total<sup>5</sup>. For each of these words, we counted the number of sequences of the PAR-CLIP data of Y. Wang et al. (2013) having at least one occurrence and determined MICO and the MICO-based  $p$ -values, separately considering the exonic and intronic datasets, see tables J.1 and J.2. The findings are summarized in table 19.2.

Notably, about half of the motifs of Bechara et al. actually occur more frequently in shuffled controls than in the signal sequences. Furthermore, of those motifs that are more frequent in the signal sequences than in the controls, only few (<15% of the Bechara et al. motifs) are significantly enriched. Importantly, the Bechara et al. motifs that are significantly enriched in the Y. Wang et al. data match reasonably well to motifs reported by Discoverer and listed in table 19.1.

**RBM10 consensus motifs of Inoue et al. (2014) unsupported by PAR-CLIP data** Inoue et al. (2014) derived RBM10 consensus motifs based on the sequences of 5' splice sites of two

<sup>5</sup>Note that 95 words are more than 9% of all 1024 5mers.

exons affected by RBM10 knock-down. These motifs are essentially identical to the motif reported to be bound by the RanBP2-type zinc finger domain of RBM10 (Loughlin et al., 2009; Nguyen et al., 2011).

We counted the number of sequences in the PAR-CLIP data of Y. Wang et al. (2013) that have at least one occurrence of these words, see tables J.3 and J.4. The resulting number are extremely low ( $< 1\%$  of sequences), in spite of our respecting all word occurrences and not just those that overlap 5' splice sites. We found the two Inoue et al. motifs not to be enriched in either the exonic or intronic PAR-CLIP sequences compared to shuffled sequences.





## Chapter 20

# Mouse embryonic stem cell transcription factors

In this chapter, we apply DMD with Discover to ChIP-Seq data of two studies of mouse embryonic stem cell (ESC) transcription factors (TFs) from X. Chen et al. (2008) and Marson et al. (2008). The proteins studied include the four factors Oct4, Sox2, c-Myc, and Klf4, which can be used to reprogram mouse fibroblasts into an induced pluripotent stem cell (iPSC) state (K. Takahashi and Yamanaka, 2006).

### 20.1 Materials and methods

**Mouse ChIP-Seq data** The first study (X. Chen et al., 2008) performed ChIP-Seq experiments for 13 sequence-specific TFs in mouse ESCs E14. These include Oct4, Sox2, Nanog, Esrrb, and Zfx, involved in ESC self-renewal (Nichols et al., 1998; Niwa, Miyazaki, and A. G. Smith, 2000)(Avilion et al., 2003)(Chambers et al., 2003; Mitsui et al., 2003)(Feng et al., 2009; X. Zhang et al., 2008)(Galan-Caridad et al., 2007), Klf4, c-Myc, and n-Myc, which contribute to reprogramming of somatic cells to a pluripotent state (J. Jiang et al., 2008)(Cartwright et al., 2005), the cell cycle regulator E2f1 (Bieda et al., 2006), and Ctcf, which insulates transcriptional domains (T. H. Kim et al., 2007), as well as Tcfcp2l1, which is preferentially upregulated in ESCs (Ivanova et al., 2006). In addition, two factors downstream of signalling pathways are included: BMP1-induced Smad1 (Ying et al., 2003), and LIF-induced Stat3 (Niwa, Burdon, et al., 1998).

The second study (Marson et al., 2008) produced additional ChIP-Seq data from mouse ESCs V6.5 for Oct4, Sox2, Nanog, and Tcf3, a repressor of key pluripotency gene expression (Cole et al., 2008; Tam et al., 2008; F. Yi, Pereira, and Merrill, 2008).

**Data preparation and analytical treatment** We retrieved the bound regions of the mouse ESC ChIP-Seq data of X. Chen et al. and Marson et al. Sequences of length 101 nt spanning 50 nt up- and down-stream of bound region midpoints were extracted from mm8.

In order to find binding site patterns with discriminative learning, the ChIP-Seq datasets were individually contrasted to dinucleotide frequency preserving shuffles of the sig-

nal sequences. Using Discover, DMD was run for motifs of lengths of 5–16 nt, using MICO as objective function. The three most discriminative IUPAC words of each length and 1-nt-shifted variants of each were considered as seeds for HMM optimization. The multiple motif discovery mode was used.

## 20.2 Discriminative motifs in ChIP-Seq data

For each ChIP-Seq dataset one or more motifs were discovered, see table 20.1. In total, 44 motifs were discovered in the 17 datasets. Many motifs are discovered multiple times in different datasets, and in these cases the motifs are quantitatively highly consistent.

TOMTOM (Gupta et al., 2007) was used to identify factors known to bind to the discovered motifs. This identified 40 of the 44 discovered motifs<sup>1</sup>. One of the four unidentified motifs' reliability is further corroborated by additional evidence presented later in this chapter.




























For 12 of the 17 datasets the top-ranking discovered motif is known to be bound by the assayed factor. The other five cases comprise E2f1, Smad1, Tcf3, and both Nanog datasets.

Table 20.1: Discriminative motif analysis of mouse ChIP-Seq data. Protein: ChIP'd protein. N: number of signal sequences. Motifs: One or more motifs discovered in the sequences of the ChIP'd protein. Factor: TF (family) known to bind the discovered motif (TOMTOM  $q$ -value  $\leq 0.05$  (Gupta et al., 2007)), bold if one of the ChIP'd proteins. IC: information content. S and C: expected relative frequency of signal and control sequences with at least one motif occurrence. MICO: mutual information of condition and motif occurrence.  $\log$ - $p$ : MICO-based  $\log$ - $p$  value, corrected for motif length. Data sources: <sup>A</sup> X. Chen et al. (2008), <sup>B</sup> Marson et al. (2008).

Protein	N	Motifs	Factor	IC [bit]	S [%]	C [%]	MICO [bit]	$\log$ - $p$
c-Myc <sup>A</sup>	3422		<b>Myc</b>	12.5	40.9	8.4	747.8	-467.1
Ctcf <sup>A</sup>	39609		<b>Ctcf</b>	18.7	83.4	3.5	43892.7	-30350.6
			?	20.3	5.2	0.2	1753.4	-1154.5
E2f1 <sup>A</sup>	20699		Ets TF family	14.6	5.6	1.4	414.8	-210.9
			Yy1	14.9	3.8	0.6	397.5	-223.9
			Nrf1	17.1	3.1	0.4	360.9	-183.4
Esrrb <sup>A</sup>	21647		<b>Esrrb</b>	14.0	68.3	6.7	14108.6	-9735.0
			?	21.2	3.5	0.5	406.0	-219.8
Klf4 <sup>A</sup>	10875		<b>Klf/Sp1</b>	14.8	57.5	8.0	4767.3	-3254.4
Nanog <sup>A</sup>	10343		<b>Sox2</b>	13.4	29.8	6.8	1410.1	-916.5
			<b>Sox2-Oct4</b>	16.6	21.4	2.9	1341.9	-859.2
			Zic	14.8	13.1	2.2	685.0	-408.4
			<b>Nanog</b>	13.0	8.0	2.8	202.7	-103.5
Nanog <sup>B</sup>	16667		<b>Sox2-Oct4</b>	15.5	25.8	4.1	2430.6	-1619.1
			<b>Sox2</b>	13.3	24.1	6.6	1487.1	-979.9
			<b>Klf/Sp1</b>	14.0	10.1	2.2	699.1	-443.2
			Zic	15.8	6.6	1.2	519.8	-288.8

<sup>1</sup>Where a motif is known to be bound by more than one protein of a TF family, the family name is used to refer to the motif.

Table 20.1: continued from previous page.

Protein	N	Motifs	Factor	IC [bit]	S [%]	C [%]	MICO [bit]	log-p
n-Myc <sup>A</sup>	7182	 	Myc	12.1	33.8	7.6	1152.3	-747.7
			Klf/Sp1	16.0	10.5	3.1	235.4	-116.3
Oct4 <sup>A</sup>	3761	 	Sox2-Oct4	18.1	42.3	2.6	1422.7	-910.3
			Klf/Sp1	16.1	7.3	1.6	109.8	-28.8
Oct4 <sup>B</sup>	17 225	 	Sox2-Oct4	17.4	46.8	3.8	6898.7	-4706.8
			Klf/Sp1	15.5	8.4	1.7	635.2	-388.9
Smad1 <sup>A</sup>	1126	   	Sox2	11.2	31.0	9.6	119.5	-50.7
			Oct4	12.2	17.5	4.5	74.3	-14.1
			Klf/Sp1	14.2	13.2	2.5	69.7	-5.8
			Esrrb	13.3	10.1	1.7	56.3	-1.5
Sox2 <sup>A</sup>	4526	   	Sox2	13.5	64.0	10.3	2177.7	-1434.0
			Oct4	13.0	11.3	2.7	194.2	-102.6
			Klf/Sp1	14.1	4.7	1.2	74.1	-13.9
			Zic	16.3	6.2	0.9	149.2	-41.3
Sox2 <sup>B</sup>	15 036	  	Sox2-Oct4	17.5	38.9	3.3	4720.2	-3196.6
			Sox2	13.2	30.2	6.7	2132.0	-1432.1
			Klf/Sp1	14.1	7.7	1.4	554.1	-342.6
Stat3 <sup>A</sup>	2546	 	Stat3	14.1	40.0	3.7	800.8	-498.9
			?	18.6	6.8	0.8	98.6	-11.0
Tcf3 <sup>B</sup>	6257	   	Sox2-Oct4	17.2	38.0	3.7	1829.5	-1192.4
			Tcf3	14.5	22.5	4.3	702.2	-430.5
			Zic	15.2	5.4	0.8	175.2	-64.4
			Klf/Sp1	14.1	5.5	1.1	150.5	-47.2
Tcfcp2l1 <sup>A</sup>	26 910	 	Tcfcp2l1	14.6	75.0	10.9	17 829.4	-12 284.1
			Esrrb	15.5	5.2	0.8	713.0	-457.9
Zfx <sup>A</sup>	10 338	 	Zfx	11.5	44.4	17.5	1287.5	-841.4
			?	15.0	6.8	2.4	170.3	-66.0

The occurrence numbers given in table 20.1 are based on posterior decoding. This means that overlapping motifs compete for binding sites, and the contributions of overlapping motifs for binding to a given site is averaged probabilistically, incorporating both the fit of the motifs to the sequence of the site and the motifs' prior occurrence probability.

**Sox2-Oct4 heterodimer motif** Sox2 and Oct4 are known to bind DNA as heterodimer (Kuroda et al., 2005; Ng et al., 2012; Nishimoto et al., 1999; Tokuzawa et al., 2003; Tomioka et al., 2002). The corresponding Sox2-Oct4 heterodimer binding site pattern is discovered in six of the datasets, and the data in which it is discovered yield highly consistent motifs. Moreover, in five of these six datasets, it is found to be the most discriminative motif. In particular, it is independently discovered to be the most discriminative motif in both Oct4 datasets and one Sox2 dataset.

The Sox2-Oct4 heterodimer pattern is also the most discriminative motif in data of the functionally related factors Nanog and Tcf3. The enrichment for the Sox2-Oct4 heterodimer pattern is in concordance with the overlap of regions bound by Sox2 and Oct4 with those bound by Nanog (X. Chen et al., 2008) or those bound by Tcf3 (Marson et al.,

2008), as well as with the functional overlap of Sox2 and Oct4 with Nanog (Boyer et al., 2005; Loh et al., 2006) and with Tcf3 (Cole et al., 2008; Tam et al., 2008; F. Yi, Pereira, and Merrill, 2008).

**Sox2 and Oct4 monomer motifs** The monomer binding motifs of Sox2 and Oct4 are discovered in five and two datasets, respectively. The Sox2 monomer motif is the most discriminative motif in three dataset of X. Chen et al. These include the Sox2 dataset, and those of Nanog and Smad1. The Oct4 monomer motif is the second most discriminative motif for Smad1 data and in one Sox2 dataset.

**Co-discovery of Sox2 and Oct4 monomer and heterodimer motifs** Interestingly, in some datasets both the Sox2-Oct4 heterodimer and the Sox2 monomer motifs are discovered. This is the case for both Nanog datasets, as well as one Sox2 dataset. In the other Sox2 dataset, as well as in the Smad1 dataset, both Sox2 and Oct4 monomer motifs are discovered. The two Oct4 datasets, on the other hand, exhibit only the Sox2-Oct4 heterodimer pattern.

Overall, in each dataset where a Sox2 monomer motif is discovered, also the Oct4 monomer motif or the Sox2-Oct4 heterodimer motif is discovered. Conversely, in each dataset for which the Oct4 monomer motif is discovered, also the Sox2 monomer motif is discovered.

**Motif of pluripotency regulator Zic3 frequently co-discovered with Sox2 and Oct4 motifs** Zic3 is required for maintenance of pluripotency in ESCs (L. S. Lim, Loh, et al., 2007), and is a direct activator of the Nanog promoter in ESCs (L. S. Lim, Hong, et al., 2010). In addition, Zic3 has been shown to enhance the generation of mouse iPSCs (Declercq et al., 2013). We co-discovered the Zic family motif in four of the eight samples in which either the Sox2-Oct4 heterodimer motif or the Sox2 and Oct4 monomer motifs were found. In particular, this includes both Nanog datasets.

**Motif of pluripotency factor Klf4 co-discovered with Sox2 and Oct4 motifs** The Klf/Sp1 motif<sup>2</sup> is the only motif discovered in the ChIP-Seq dataset of the pluripotency factor Klf4. It is co-discovered in eight other datasets: both Oct4 datasets, both Sox2 datasets, the n-Myc, Smad1 and Tcf3 datasets, and one Nanog dataset. Importantly, in seven of the eight datasets in which the Klf/Sp1 motif is co-discovered, also the Sox2-Oct4 heterodimer motif or both Sox2 and Oct4 monomer motifs are discovered.

**c-Myc and n-Myc** The proto-oncogene c-Myc was one of the factors used in the first reported reprogramming of mouse somatic cells to iPSCs (K. Takahashi and Yamanaka, 2006). In human stem cells it was found to induce apoptosis and differentiation into extraembryonic tissue types (Sumi et al., 2007) and it is dispensable for induction of pluripotency in human somatic cells (Yu et al., 2007).

---

<sup>2</sup>Because of their similarity we group together the motifs of the family of Krüppel-like factors (Klf) with that of the evolutionarily related protein Sp1 (A. R. Black, J. D. Black, and Azizkhan-Clifford, 2001; Kaczynski, Cook, and Urrutia, 2003).

Table 20.2: Longer motif variants of Tcfcp2l1. The first row corresponds to the motif given in table 20.1, the motifs in the other rows were seeded as indicated. Seeds are given as IUPAC regexes, see table 3.1. Other columns as in table 20.1.

Seed	Motif	IC [bit]	S [%]	C [%]	MICO [bit]	log-p
DDCYDRHYNNDCYDN		14.6	75.0	10.9	17 829.4	-12 284.1
NNCCGGTTNNAACCGGNN		14.8	76.2	10.3	18 903.4	-13 018.6
NNCCGGTTNNAACCGGNNN		15.0	76.7	9.9	19 506.4	-13 426.6
NNNNCCGGTTNNAACCGGNNNN		15.1	76.5	9.7	19 543.0	-13 441.9

The ChIP-Seq data of c-Myc and n-Myc are both enriched for the E-box motif bound by the Myc TF family. In addition, in the n-Myc dataset the Klf/Sp1 motif is co-discovered.

That Myc proteins and Sp1 interact is well known: Myc and Sp1 cooperate in activating transcription of the human telomerase reverse transcriptase gene hTERT (Kyo et al., 2000), the catalytic subunit of human telomerase important for cellular immortalization and carcinogenesis; c-Myc interacts via its Zinc-finger domain with Sp1 (Gartel et al., 2001); promoters bound by both c-Myc and Sp1 exhibit higher degrees of conservation between human and rodents, stronger correlation with TFIID-bound promoters, and preference for permissive chromatin state (F. Parisi, Wirapati, and Naef, 2007); and interaction of Myc and Sp1 is also important for the regulation of clock controlled genes (Bozek et al., 2010) and various cancers (Gopisetty et al., 2013; H.-B. Wang et al., 2013).

**Cognate motifs are discovered for Ctfc, Esrrb, Stat3, Tcfcp2l1 and Zfx** In the ChIP-Seq data of Ctfc, Esrrb, Stat3, Tcfcp2l1 and Zfx the most discriminative motifs versus shuffled sequences are the known, cognate motifs.

The Ctfc motif shows the highest amount of evidence of all discovered motifs, followed by the Tcfcp2l1 motif. Two reasons are responsible for the large amount of evidence for these motifs: size of datasets and motif enrichment. Comprising respectively 39 609 and 26 901 ChIP-Seq sequences, the Ctfc and Tcfcp2l1 datasets are the largest of all considered datasets. Simultaneously, the respective motif frequency in the signal sequences, as well as the enrichment over the control sequences are the largest across all considered ChIP-Seq datasets: 83.4% vs. 3.5% for Ctfc and 75% vs. 10.9% for Tcfcp2l1 (signal frequency vs. control frequency).

In each of these five ChIP-Seq datasets secondary motifs are co-discovered. For the Tcfcp2l1 the Esrrb motif is co-discovered. Affinity purification and subsequent mass spectrometry of Esrrb interacting proteins revealed Tcfcp2l1 to be the third-highest scoring TF (van den Berg et al., 2010), and conversely Esrrb was the sixth-highest scoring TF among the interacting proteins of Tcfcp2l1 (van den Berg et al., 2010).

The four co-discovered motifs in the Ctfc, Esrrb, Stat3, and Zfx data are not identified by database queries using TOMTOM, and will be further discussed below.

**A long Tcfcp2l1 motif composed of identical, reverse-complementary halves** Seeds of length 5–16 nt were used to discover the motifs of table 20.1, and the Tcfcp2l1 motif listed there indicates that the true motif might be even longer. Table 20.2 compares the statistics

of the discovered motif from table 20.1 with those of longer HMM motifs seeded on a core sequence of CCGGTTNNAACCGG extended with Ns on both sides to lengths of 18, 20, and 22 nt. As is visible from this table, the longer motif variants yield increased MICO, but beyond 20 nt length further flanking positions contribute only little IC. Given these results, it appears that the preferred Tcfcp2l1 bind site consists of two adjacent, nearly identical motif-halves. The halves consist of reverse-complementary 10mers of which the central 8 positions are responsible for discrimination. Previous analyses of this dataset have only reported shorter motifs (Bailey, 2011; X. Chen et al., 2008).

**Cognate Nanog and Tcf3 motifs are not the most discriminative motifs** In the Tcf3 ChIP-Seq data, the Tcf motif is less discriminative than the Sox2-Oct4 heterodimer motif (Sox2-Oct4 motif: 38% vs. 3.7%, Tcf motif: 22.5% vs. 4.3%, signal frequency vs. control frequency). As noted above, the discovery of the Sox2 and Oct4 motifs in the Tcf3 data is consistent with their functional overlap (Cole et al., 2008; Tam et al., 2008; F. Yi, Pereira, and Merrill, 2008).

In the Nanog ChIP-Seq data of X. Chen et al., the Nanog motif is only the fourth most discriminative motif, outranked by the Sox2 monomer motif, the Sox2-Oct4 heterodimer motif, and the Zic motif. Notably, the discovered Nanog motif's occurrence frequency is rather low in the signal sequences at only 8% (and 2.8% in the control sequences). In the Nanog ChIP-Seq data of Marson et al., the Nanog motif is not discovered, while those of the Sox2-Oct4 heterodimer, the Sox2 monomer, Klf/Sp1, and Zic are discovered. As also already noted above, the discovery of the Sox2, Oct4, Klf/Sp1, and Zic motifs in the Nanog data is consistent with their shared contribution to maintaining pluripotency (Boyer et al., 2005; Loh et al., 2006), and for the cases of Sox2 and Oct4 with their overlapping set of ChIP-Seq bound promoters (X. Chen et al., 2008).

The datasets of Nanog and Tcf3 will be subjected to further analysis in section 20.4.

**Purported E2f consensus motif not supported by E2f1 motif discovery results** The motifs discovered in the ChIP-Seq data of E2f1 are identified by TOMTOM database queries as those of the Ets family of TFs, of Yy1, and of Nrf1. All of these motifs have two positions with purines, either followed by or embedded in a stretch of pyrimidines. As such, they are not unlike the purported E2f consensus motif TTTSSCGC<sup>3</sup> derived from *in-vitro* experiments (Chittenden, Livingston, and Kaelin, 1991). However, our results do not include a motif that is identified by database searches as the E2f consensus motif, and occurrence statistics of the consensus motif<sup>4</sup> are not discriminative between the X. Chen et al. ChIP-Seq sequences and shuffles.

Already earlier ChIP-Chip and ChIP-Seq experiments for E2f binding specificity did not support the validity of the E2f consensus motif (Rabinovich et al., 2008; X. Xu et al., 2007), in that few of the identified targets possess occurrences of the consensus motif. In fact, Bailey and Machanick (2012) analyzed the X. Chen et al. E2f1 dataset consider here and arrived at the same conclusions.

<sup>3</sup>Note that S stands for C or G, see table 3.1.

<sup>4</sup>Purported consensus motif in X. Chen et al. E2f1 ChIP-Seq sequences: 206 of 20 699 = 0.995%, shuffled control sequences: 189 of 20 699 = 0.913%; MICO: 0.53 bit, log *p*-value  $\geq$  0.

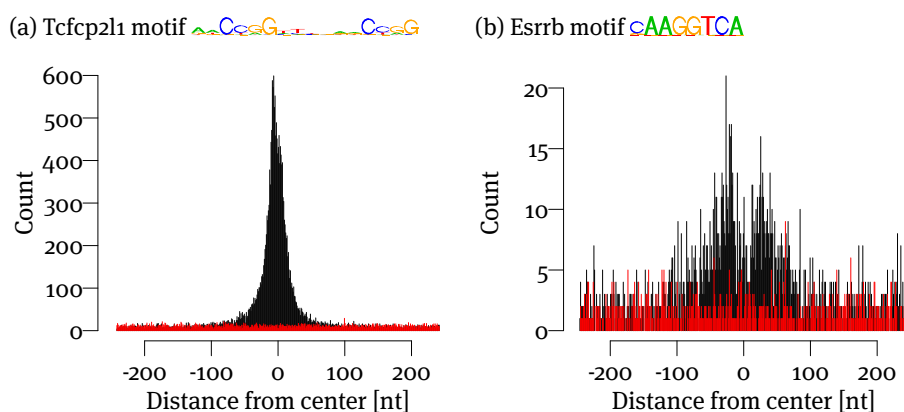


Figure 20.1: Position-wise spatial distribution of motif occurrences in up to 250 nt distance from midpoints of Tcfcp2l1 ChIP-Seq regions. (a) Tcfcp2l1 motif discovered in Tcfcp2l1 sequences. (b) Esrrb motif co-discovered in Tcfcp2l1 sequences. Black: occurrences in the signal sequences, red: occurrences in the control sequences. Note that windows of length 101 nt were used for motif discovery and are the basis of the statistics listed in table 20.1. See table K.1 for the positional distributions of all discovered ChIP-Seq motifs. Data source: X. Chen et al. (2008).

**Smad1** Extrinsic signaling pathways are important for maintenance of pluripotency in ESCs. In ESCs the cytokine leukemia inhibitory factor, LIF, acts to suppress differentiation via STAT3 (Niwa, Burdon, et al., 1998). Another external signal, bone morphogenetic factor, BMP, feeds into nuclear regulation of ESC transcriptional programs via the Smad proteins (Attisano and Wrana, 2002; von Bubnoff and Cho, 2001). The mechanisms of LIF- and BMP-signalling dependent regulation are complex and their effects context- and inter-dependent (Ying et al., 2003). Nanog has been shown to physically interact with Smad1 and interfere with recruitment of co-activators (Suzuki et al., 2006).

In the Smad1 ChIP-Seq data both Sox2 and Oct4 monomer motifs are discovered by Discover. Also, the motifs of Klf/Sp1 and Esrrb are discovered, albeit at lower enrichment. Our *de-novo* motif discovery results do not, however, include the Smad family motifs, and thus yield no evidence for direct binding of Smad1 in the mouse ESC ChIP-Seq sequences of X. Chen et al. This is in accordance with the requirement for co-activator interaction of Smad1 for transcriptional regulation, as well as with previous findings by Bailey and Machanick (2012), who, using database motifs, found central enrichment for the Sox2 and Oct4 monomer motifs in the Smad1 sequences of X. Chen et al., but no central enrichment for the Smad family motifs.

## 20.3 Spatial distribution of motif occurrences

Next we will consider spatial information of the discovered motifs' occurrences within the ChIP-Seq sequences. Table K.1 gives the spatial occurrence distribution around sequence midpoints for all motifs listed in table 20.1, an illustrative case is shown in figure 20.1, and we will make several observations from these.



**Cognate motifs exhibit higher central enrichment than co-factor motifs** As illustrated for the Tcfcp2l1 motif in the Tcfcp2l1 dataset in figure 20.1a, the spatial distribution around sequence midpoints of occurrences of cognate motifs is peaking close to zero. Conversely, the spatial occurrences distributions of co-discovered co-factor motifs often lack central enrichment or are even depleted around sequence midpoints, as illustrated for the co-discovered Esrrb motif in the Tcfcp2l1 sequences in figure 20.1b.

**Spatial occurrence distribution corroborates relevance of unidentified motif in Ctfc data** Among the four unidentified motifs discovered in the ChIP-Seq data only the one found in the Ctfc data exhibits a distribution of motif occurrences around sequence midpoint that is strongly indicative of it being a relevant Ctfc co-factor motif. The unidentified motifs co-discovered in the Esrrb and Stat3 data appears to be slightly enriched around the sequence midpoints but both also have a uniform contribution mixed into their spatial distribution, and due to the low number of occurrences the central enrichment is not very strong. Finally, the unidentified motif in the Zfx data appears rather uniformly distributed and even somewhat depleted around sequence midpoints.

Concluding, only the unidentified motif in the Ctfc data is strongly corroborated by the spatial distribution of occurrences around sequence midpoints.

## 20.4 Contrasting Nanog and Tcf3 against other ChIP-Seq data

For Nanog and Tcf3 the most discriminative motifs against shuffled sequences are those of Sox2 and Oct4. Extending earlier analyses of Bailey (2011), we contrasted the Nanog and Tcf3 sequences to other factors' ChIP-Seq sequences highly enriched for the Sox2 and Oct4 motifs. The most discriminative motifs were identified based on the three most discriminative seeds of lengths 5–16 nt and 1-nt shifted variants. These analyses yield in 15/16 cases the cognate motifs of Nanog and Tcf3, see table 20.3. Only Nanog (Marson et al., 2008) versus Oct4 (X. Chen et al., 2008) yields another motif: the Sox2 monomer motif is more discriminative across this contrast than the cognate motif.

In general, the cross-contrasting inter-dataset analyses for the Nanog datasets, especially that of Marson et al., show more variability in the discovered motif than those for the Tcf3 dataset. The reason for this likely is the low frequency of the Nanog motif in the Nanog datasets<sup>5</sup>.

## 20.5 Comparing results of DREME, FIRE, and Discover






The DREME publication (Bailey, 2011) analyzed 13 of the 17 datasets studied here and reported more motifs than discovered with our method. Exemplarily, we investigated the differences of Discover and DREME analyses of the Oct4 data of X. Chen et al., also including FIRE into this comparison. We generated two additional sets of shuffled sequences, and applied the methods on the three contrasts. Table 20.4 lists the results.

<sup>5</sup>The Nanog motif is found in about 25% in the X. Chen et al. sequences, and about 15% in those of Marson et al., while the Tcf3 motif is found in about 30% of Tcf3 sequences, see table 20.3.








Table 20.3: Inter-dataset comparison reveals motifs discriminating Nanog and Tcf3 data from other ChIP-Seq data. (a),(b) comparing Nanog ChIP-Seq sequences against those of Oct4, Sox2, and Tcf3. (c): comparing Tcf3 ChIP-Seq sequences against those of Nanog, Oct4, and Sox2. Data sources: <sup>A</sup> X. Chen et al. (2008), <sup>B</sup> Marson et al. (2008).

(a) Nanog<sup>A</sup>, N=10 343

vs. Protein	N	Motifs	Factor	IC [bit]	S [%]	C [%]	MICO [bit]	log-p
Oct4 <sup>A</sup>	3761		Nanog	10.1	30.8	10.7	478.7	-255.3
Oct4 <sup>B</sup>	17 225		Nanog	10.8	23.4	8.5	827.4	-507.3
Sox2 <sup>A</sup>	4526		Nanog	10.1	23.9	11.0	256.7	-131.1
Sox2 <sup>B</sup>	15 036		Nanog	10.5	22.6	9.7	568.5	-342.6
Tcf3 <sup>B</sup>	6257		Nanog	10.5	19.9	8.6	290.0	-159.2

(b) Nanog<sup>B</sup>, N=16 667

vs. Protein	N	Motifs	Factor	IC [bit]	S [%]	C [%]	MICO [bit]	log-p
Oct4 <sup>A</sup>	3761		Sox2	8.3	53.0	31.9	402.9	-227.7
Oct4 <sup>B</sup>	17 225		Nanog	11.0	14.0	6.5	374.7	-208.0
Sox2 <sup>A</sup>	4526		Nanog	10.9	10.3	5.1	92.4	-21.7
Sox2 <sup>B</sup>	15 036		Nanog	11.3	11.1	5.7	217.5	-103.8
Tcf3 <sup>B</sup>	6257		Nanog	8.5	25.4	19.8	57.8	-12.5

(c) Tcf3<sup>B</sup>, N=6257







vs. Protein	N	Motifs	Factor	IC [bit]	S [%]	C [%]	MICO [bit]	log-p
Nanog <sup>A</sup>	10 343		Tcf3	14.7	26.0	6.4	883.9	-556.4
Nanog <sup>B</sup>	16 667		Tcf3	14.9	24.6	6.6	924.4	-584.6
Oct4 <sup>A</sup>	3761		Tcf3	12.0	38.4	11.2	681.7	-416.2
Oct4 <sup>B</sup>	17 225		Tcf3	13.8	30.1	8.8	1078.2	-686.2
Sox2 <sup>A</sup>	4526		Tcf3	13.8	29.7	7.5	636.5	-389.8
Sox2 <sup>B</sup>	15 036		Tcf3	14.7	24.6	5.6	1042.9	-666.8

Table 20.4: Comparison of discriminative motif discovery with (a) Discover, (b) DREME, and (c) FIRE on Oct4 data of X. Chen et al. (2008) against three sets of shuffled sequences. Motifs are presented in decreasing order of significance as reported by the method. Factor: factor binding the motif; factor identification based on TOMTOM searches (with  $q$ -value  $< 0.05$ ), with manual judgement in ambiguous cases, and in cases (denoted “?”) where TOMTOM did not identify matches. DREME and FIRE discover IUPAC regex motifs, shown in these tables. Note, that for DREME and FIRE PWMs could be built from the statistics of words matching the discovered regexes. The final rows list wall clock and CPU time (hours:minutes:seconds).

(a) Discover

Rank	Shuffles 1		Shuffles 2		Shuffles 3	
	Motif	Factor	Motif	Factor	Motif	Factor
1		Sox2-Oct4		Sox2-Oct4		Sox2-Oct4
2		Klf/Sp1		Klf/Sp1		Klf/Sp1
Wall	01:54:12		01:54:41		01:57:28	
CPU	11:38:04		11:32:09		11:30:48	

(b) DREME

Rank	Shuffles 1		Shuffles 2		Shuffles 3	
	Motif	Factor	Motif	Factor	Motif	Factor
1		Oct4		Oct4		Oct4
2		Sox2-Oct4		Sox2-Oct4		Oct4
3		Oct4		Oct4		Sox2
4		Klf/Sp1		Sox2		Klf/Sp1
5		Sox2		Klf/Sp1		Ets TF family
6		Ets TF family		Ets TF family?		Sox2-Oct4
7		Oct4?		Oct4?		Oct4?
8		Esrrb		Esrrb		Esrrb
9		Oct4?		?		Oct4?
10		Klf/Sp1		?		?
11		?		Oct4?		Oct4?
12		Sox2?		Klf/Sp1		Sp/Egr
13		Oct4?		Sox2		Sox8
14		?		?		?
15		Oct4?		Oct4?		Myc
16		Myc		?		Zic
17		?		Oct4?		Oct4
18		?		?		?
19		Yy1		Myc		
20				?		
21				Oct4?		
Wall	00:16:36		00:16:50		00:15:43	
CPU	00:16:36		00:16:50		00:15:43	

Table 20.4: continued from previous page. E: sample in which the FIRE motif is enriched; +: signal sequences, -: control sequences.

(c) FIRE

Rank	Shuffles 1			Shuffles 2			Shuffles 3		
	Motif	Factor	E	Motif	Factor	E	Motif	Factor	E
1	ATGAT	Oct4?	+	ATGAA	Oct4?	+	ATGAT	Oct4?	+
2	CCCC	Klf/Sp1	+	TTG	?	-	TTTGT	Sox2?	+
3	AT	?	-	CCC	Klf/Sp1	+	AT	?	-
4	TTTGT	Sox2?	+	ATCA	Oct4?	+	CCC	Klf/Sp1	+
5	AAA	?	-	TCA	?	-	TTCC	Ets TF family	+
6	TGCA	?	-	TTCC	Ets TF family	+	TCA	?	-
7	AAAG	?	-	AGCA	Nr4A2	+	TTG	?	-
8	TCC	Ets TF family?	+	TCCGAA	Oct4?	+	ATCCA	Oct4?	+
9	TCCGA	Oct4?	+	TTATT	?	+	GATAA	?	+
10	GTTG	?	-	CCCT	?	-	TGCA	?	-
11	ATATGC	Sox2-Oct4?	+	TGG	?	-	AAGTCA	Esrrb	+
12	AAAGA	?	+	CCAA	?	-	GAAG	?	+
13	CAAGGT	Esrrb	+	TTGTT	Sox2?	+	CACAAG	?	+
14				ATTT	?	-			
15				TGAA	Oct4?	+			
16				TTGAAA	Oct4?	+			
Wall	00:13:02			00:14:54			00:14:03		
CPU	00:13:02			00:14:54			00:14:03		

Discover consistently reports the full-length Sox2-Oct4 heterodimer motif and the Klf/Sp1 motif (table 20.4a). DREME finds 18–21 IUPAC RE motifs (table 20.4b) and FIRE 13–16, of which 8–10 are enriched in the signal sequences (table 20.4c). Both DREME and FIRE are designed for the discovery of short motifs, and respectively 8–10 and 3–6 of their motifs are variants of partially overlapping segments of the Sox2-Oct4 heterodimer pattern. DREME also consistently finds the Esrrb and Myc motifs, while FIRE finds the Esrrb motif in 2 of 3 analyses. Other motifs found by DREME and FIRE are not identifiable as known motifs, or are not reproduced when running repeatedly against different sets of shuffled sequences.

Note that the results in tables 20.4a to 20.4c are listed in order of significance, as reported by the methods. Inspection of motifs discovered by DREME and FIRE reveals that the more dubious motifs are those with lesser significance. Thus, when a higher significance threshold would be chosen than the standard one used by the methods, then the DREME and FIRE results would include fewer dubious motifs, such as unidentified motifs or motifs that are not reproduced when different shuffles are used as controls. Consequently, lower false-positive rates would be expected for DREME and FIRE. However, at the same time both methods would also report fewer true-positive co-factor motifs, while the redundancy issues would not be resolved.



**Part V**

**Discussion**



## Chapter 21

# Supervised motif discovery performance experiments

This chapter discusses the supervised MD experiments based on synthetic data. Section 21.1 summarizes the influence of the parameters varied in the synthetic data, and relates the observed effects to theoretical expectations. Advantages and disadvantages of generative and discriminative learning approaches are discussed in sections 21.2 and 21.3. Section 21.4 comments on the robustness of hybrid learning. Finally, section 21.5 compares the MD performance of Discoverer and that of published methods.

### 21.1 Influence of parameters varied in synthetic data

**Sequence number** How much data is necessary to saturate MD performance was studied by varying the number of sequences available for learning. For all methods MD performance substantially increased when going from 100 to 1000 sequences. With 10 000 sequences MD performance further approached the limit of motif recognizability.

**Sequence context size and motif implantation frequency** Varying sequence context size and implantation frequency allowed to study the sensitivity of MD performance with respect to signal preponderance, i.e. the product of the implantation frequency and the reciprocal of sequence context size. MD performance was found to react approximately linearly to logarithmic changes of preponderance. This observation is in line with theoretical expectations according to motif occurrence inference schemes (see section 3.1.3.1) in which the negative log-odds of the (position-wise) occurrence frequency acts as cutoff on a PSSM score.

**Motif information content** Control over IC allows to determine the signal/noise ratio of binding site predictions. Among the parameters varied, IC of the true motif was found to be the most important determinant of motif recognizability, and thus of the limit of MD performance. MD performance exhibits a non-linear response to variation of IC, with a sigmoidal contribution due to the response of sSn.

To understand the sigmoidal response of  $sSn$ , note the remark in equation (3.17) about the interpretation of the IC of a motif as the expected score of a PSSM generated by the PSFM. It implies that reductions in IC lead to a larger overlap of the distribution of PSSM scores generated by the PSFM, and the distribution of PSSM scores generated by background. As the number of words typically generated by the background, i.e. its information entropy, is generally much larger<sup>1</sup> than that of the PSFM, the increasing distributional overlap with diminishing IC causes the observed sigmoidal response in  $sSn$ : when IC is high, the distributions of PSSM scores of signal and background are well separated, leading to a dominance of words generated by signal among the words that score above threshold, and, correspondingly, an  $sSn$  that is close to 1; conversely, when IC is low, due to the high number of words generated by background, any threshold on the PSSM score that captures meaningful proportions of the signal words, must also capture many background words. In combination with the fact that only few positions in the synthetic sequences correspond to implanted motifs<sup>2</sup>, this leads to a dominance of background positions among the positions that score above threshold, which corresponds to an  $sSn$  close to 0. In between these extremes is a phase of intermediate IC, in which the proportion of words with PSSM above threshold shifts from signal dominated to background dominated. In summary, this behaviour is tantamount to the properties of the sigmoidal function.

**Sensitivity and positive predictive value** Predictions of most methods<sup>3</sup> generally yield higher sPPV than  $sSn$ . This is likely due to the relative scarcity of binding sites in the data, i.e. the fact the number of binding site positions is much less than the number of non-binding site positions. When signal is sparse, overall performance is typically maximized by tuning prediction so as to avoid type II errors (equivalent to high sPPV), at the price of increased type I errors (equivalent to low  $sSn$ ), see section 3.1.3.1 for the definitions of these error types. Due to the relative scarcity of true site positions, only few type I errors are possible. But if predictions were prone to type II errors, many type II errors would happen due to the high number of non-binding site positions.

Similarly, DMD results yield higher sPPV than recognizability, while  $sSn$  of DMD is lower than that of recognizability. This means that MD tends to miss true motifs, but rarely errs when reporting motifs, which shows that MD results are conservative, in the sense of only predicting motifs when sufficient evidence is available.

**Averaging over remaining variates** It should be noted that the observations regarding the sensitivity of MD performance with respect to one parameter are based on averaging over the values of other parameters, and the additional gains from increased sequence numbers are larger in more difficult learning problems. Thus, e.g. saturation of MD performance with increasing number of sequences occurs earlier for frequent, high-IC motifs, and later for infrequent, low-IC motifs.

<sup>1</sup>They are equal only for motifs of maximal information entropy, i.e. motifs that by definition are indistinguishable from a uniform background

<sup>2</sup>When  $p$  is the motif implantation probability and  $n$  the sequence length then motifs are implanted with a position-wise frequency of only  $\frac{p}{n} \ll 1$ . In the synthetic data  $0.01 \leq p \leq 1$  and  $20 \leq n \leq 1000$ , so  $10^{-5} \leq \frac{p}{n} \leq 0.05$ .

<sup>3</sup>With the exception of CMF and DECOD.



**Performance measurements also depend on invariants** The observations made in the synthetic data experiments also hinge on the choices for constant parameters of data generation. For example, a length of 8 nt was chosen to model typical RBP binding sites. However, as exemplified in the ChIP-Seq data analyzed in chapter 20, many DNA-binding factors recognize longer motifs.

**PSSM as motif models** Another point of discussion regarding the synthetic data experiments concerns the choice of PSSM as simulated motif model to generate motif occurrences. While being much used in the field of sequence analysis (Benos, Bulyk, and Stormo, 2002), PSSMs disregard the possibility of dependent emission probabilities at different positions, and thus need not be good approximations to many, let alone most binding site patterns.

## 21.2 Generative, signal-only learning

**Opportunistic cases allow generative learning to excel** The Baum-Welch algorithm employed for generative learning performs maximum likelihood (ML) estimation of HMM parameters. ML estimation has the property of statistical consistency<sup>4</sup>. It is this property that enables generative, signal-only learning in opportunistic cases<sup>5</sup> to approximate the motif recognizability limit (figure 17.2(a)). Indeed, it became apparent that already sample sizes of 1000 sequences are (within the ranges of the other parameters) frequently sufficient for generative, signal-only learning's MD performance to come very close to this limit. Yet, the model correctness assumption on which consistency of generative, signal-only learning rests may frequently be violated due to various factors.

**Un-modeled effects deteriorate MD performance of generative learning** Relative to motif recognizability the MD performance of generative, signal-only learning degrades when the background is of a more complex nature than modeled (figure 17.2(b)), or when confounding motifs are present (figure 17.2(c)). When real data contains effects that are not part of the model, generative, signal-only learning does not immediately become useless. Instead, it will yield models that incorporate characteristics of the unmodeled effects, accommodating them according to the structure of the model. If these effects are of small magnitude, the models will tend to be useful to describe the signal. When they get stronger, as is the case when the relative frequency or IC of confounding motifs increases, then generative, signal-only learning may fail to identify the true signal and yield spurious relationships. In such cases, the discovered motif may represent the confounding motif, or one that is frequent in the background sequence.

**Improving generative learning with discriminative seeding or filtering** We considered several ways how deficits of generative learning can be ameliorated by using contrasting information in seeding, learning, or filtering of models. Using discriminative measures

<sup>4</sup>With increasing sample size estimates converge on their true values.

<sup>5</sup>I.e. when the true model is used.

to seed models which are then optimized via generative learning may improve MD performance, as was observed with the decoy experiments (BW-MICO vs. BW in figure 17.1). Yet, In other cases discriminative seeding has little consequence and generative learning invariably identifies the (non-discriminative) likelihood maximum by drifting away into non-discriminative regions of parameter space, as in the 3'UTR experiments (BW-MICO vs. BW in figure 17.1). When generative models are not discriminative versus suitable control data, the true parameters are likely different from those of the model. By not predicting models in such cases, MD performance is substantially improved (BW and BW-MICO in figure H.7). Applying discriminative seeding and discriminative filtering to generative models ameliorated the deficiencies that generative learning exhibits in its performance on the decoy experiments as compared to recognizability (BW-MICO in figure H.7). However, in the 3'UTR experiments the MD performance deficiency of generative learning is not equally ameliorated by discriminative seeding and filtering (BW-MICO in figure H.7). Also, this could not be fixed by increasing the order of the used background model (results not shown).

There exist further possibilities—not considered here—to approach the particular difficulties faced by generative learning in the synthetic data. E.g. the presence of confounding motifs in the decoy experiments could be tackled by fixing the misspecification of the model, by making use of multi motif models. After identifying parameters of a multi motif model one would then ask which of the motifs, if any, correspond to signal. This could be decided by considering discriminative measures.

### 21.3 Discriminative learning

#### **When generative learning performs optimally discriminative learning is nearly as good**

The basic set of synthetic sequence experiments demonstrates the optimality of generative learning when true model is within the parametric family. When these conditions are fulfilled, generative learning based on the ML principle achieves a MD performance close to the limit of motif recognizability. MD performance of discriminative learning is only marginally inferior in this set of experiments. The largest reductions of MD performance with respect to motif recognizability is seen for both learning approaches when data is limited.

**Leveraging control data ameliorates model mis-specifications** The supervised experiments utilizing real human 3'UTR sequences as background exhibit MD performance deficits of generative, signal-only learning that are not seen for discriminative learning. As the same set of parameters are used in this set of experiments as in the basic set of experiments, any difference in MD performance is attributable to the increased complexity of the sequence background into which motifs are implanted. The complex sequence composition of human 3'UTR sequences is sufficient to confound generative learning in many instances. This results in models representing (possibly real) motifs present in human 3'UTRs whose enrichment in the data is however purely due to sampling effects. Such a decrease in MD performance is not observed with discriminative learning.

A particular model mis-specification was realized in the set of decoy motif experiments, namely single motif models for data containing two motifs. Using single motif models, MD performance was reduced for generative, signal-only learning but not for discriminative learning.

**Pre-requisites for discriminative learning** Availability control data is a necessary condition in order for discriminative learning to be applicable, and, as discussed in section 7.1, there are many ways in which useful control data can be procured. Yet, as the statistical differences between signal and control data are the information source for discriminative learning, it is clear that control data should be chosen diligently. Careless choice of control data may beget misguided conclusions regarding the sequence specificity of the analyzed proteins. While we did not systematically study the consequences of such effects empirically, there are theoretical considerations that indicate problems best avoided in choosing control data:

**Matching length distributions** When control sequences are much shorter than the signal sequences, discriminative learning might yield fairly unspecific motifs, because longer sequences have a higher chance of containing “random” occurrences of unspecific motifs. Conversely, when control sequences are much longer than the signal sequences, discriminative learning might face difficulties in accepting the true motifs, if—due to the sequence length bias—too many control sequence would exhibit occurrences. Thus, control data should exhibit comparable length distributions to the signal data.

Note though, that in studying the PUF proteins not all contrasts had matching signal and control length distributions, yet MD was still successful.

**Matching sequence composition** Many sets of biological sequences have characteristic basic sequence compositions, such as GC-bias in core promoters, or AT-bias in 3’UTRs. As DMD is designed to pick up sequence features with differing frequencies across the studied contrast, it is important not to introduce such differences inadvertently. Thus, one would be ill-advised to contrast a set of promoter sequences with synthetic control sequences generated by a uniform,  $o^{\text{th}}$ -order Markov chain, as then discriminative learning would likely produce low-IC, GC-rich motifs.

For this reason, in the absence of suitable biologically motivated control sequences, shuffling the signal sequences guarantees that such basic sequence composition effects are avoided. However, if analysis based on shuffled sequences does not reveal credible motifs, the question remains whether the signal sequences might not potentially be selected precisely because of their basic sequence composition.

**Conclusions** Defining and measuring MD performance in terms of a supervised classification problem may explain why discriminative learning, being suited to such problems, is performing so well in these experiments. However, this is not an unfair comparison, as the only additional information provided to discriminative learning is the set of control samples. In particular, the locations of the implanted sites or which of the signal

sequences actually contained occurrences were not provided for learning, and this information was only used during evaluation to measure MD performance.

Summarizing these experiments, generative learning was found to be optimal when the true model is used and in these cases discriminative learning is nearly as good. But in situations involving slight model mis-specification, which likely is the general case, discriminative learning discovers motifs more robustly than generative learning.

## 21.4 Robustness of hybrid learning

As mentioned in section 12.3, termination of Discover’s hybrid learning scheme is not guaranteed in the general case. This is because, on the one hand, it is not clear that parameter choices exist whose generative and discriminative components are simultaneously (locally) optimal for their respective objective functions. On the other hand, numerical issues may prevent hybrid learning from reaching acceptable solutions.

However, in practice such problems were not observed, and there are multiple reasons why this anecdotal evidence may be indicative of a more general phenomenon. First, gradient optimization of discriminative objectives and Baum-Welch learning involve the choice of some threshold for the iteration-wise parameter change upon reaching of which learning is terminated. This naturally censors any theoretically possible non-terminating optimization as long as optimization progress reaches a sufficiently small scale. Another theoretical consideration is the compatibility of generative and discriminative learning hinted at in the introduction, in cases in which they are mutually ancillary, i.e. when generative parameters represent variables of the model that play a peripheral but necessary role for discrimination, and in which the discriminative parameters are not among the parameters for which likelihood is most sensitive.

**Hybrid learning gracefully terminates for hard MD problems, predicting no motifs** For hard MD problems<sup>6</sup>, we observed that during hybrid learning when the discriminative learning of emissions strives towards higher-IC motifs, re-estimation of the generative learning of context parameters<sup>7</sup> frequently tends to reduce the motif priors. This generates a positive feedback loop, as less frequently occurring motifs can only be discriminative if their IC is high enough, and the higher the IC, the lower the occurrence counts are. The result of these dynamics is that the motif degenerates to full IC and vanishing motif prior. When practically no motif occurrences are predicted, the discriminative gradient vanishes, line searching for the next local optimum fails, and no increase in the discriminative objective function happens in that iteration. As hybrid learning finishes once both the relative objective function change of the discriminative learning step and the absolute parameter change of the re-estimation step fall below respective thresholds, this eventually terminates hybrid learning, and results in models that represent motifs for which no occurrences are predicted.

---

<sup>6</sup>Low-IC, low implantation frequency, large sequence context, few sequences.

<sup>7</sup>Remember that the context parameters include the motif priors, see section 12.1.

## 21.5 Comparison to published motif discovery methods

Our HMM-based method Discoverer achieved the highest MD performance of all considered methods in the synthetic data experiments. The second-best method was our IUPAC regex based seeding method Plasma. The best-performing published DMD tools were MoAn and DREME. The latter is an RE-based MD method and performed consistently better than CMF, DECOD, and DME2, which are all based on PWMs. This shows that RE-based sequence specificity models are not necessarily inferior to probabilistic ones when different objective functions and optimization procedures are used. Conversely, while FIRE uses the same objective function as Discoverer, its MD performance is much lower, demonstrating that aside from the objective function, also other properties of MD tools are important.

It was interesting to observe that the MD performance of DREME did not significantly benefit from the update that allowed to perform single-stranded motif analysis. While MoAn exhibited the highest MD performance of all published methods, it was also the slowest of all considered methods<sup>8</sup>.

Of the discriminative objective functions implemented in Discoverer, DFREQ exhibits the lowest MD performance when applied to the synthetic data experiments. Similarly, the MD performance of CMF and DECOD, which also use this objective function, was found to be worse than that of many other DMD methods.

It appears probable that the inferior performance of DFREQ is due to that, unlike other measures considered here, it does not embody a notion of significance of association. To illustrate this, consider a motif that has a relative occurrence frequency of  $o$  in the control sequences, and of  $x < \frac{1}{2}$  in the signal sequences. It should probably be judged more significantly associated with the contrast than if the numbers were  $\frac{1}{2}$  and  $\frac{1}{2} + x$  respectively. Yet, DFREQ measures these cases as equally strongly associated with the contrast. This leads us to discourage the use of DFREQ as discriminative objective function.

The two published signal-only MD tools included in the comparison, BioProspector (X. Liu, Brutlag, and J. S. Liu, 2001) and MDscan (X. S. Liu, Brutlag, and J. S. Liu, 2002), showed MD performance comparable to or below that of the worst-performing DMD methods considered.

**Runtime** Plasma and Discoverer, the IUPAC regex and HMM based MD methods presented in this thesis, respectively achieve the best MD performance for methods of their kind. Simultaneously, they are also the fastest methods in their respective categories of MD tools. DREME, the second best IUPAC regex based method, was the third fastest IUPAC regex based method, and faster than Discoverer. MoAn, after Discoverer the second-best method based on probabilistic motif modeling, was the slowest of all considered methods.

Various reasons are responsible for the higher speed of Plasma and Discoverer compared to other methods. They comprise clever choice and design of algorithms, efficient coding techniques in a compiled language, and usage of parallelism.

<sup>8</sup>Of course, the three excluded methods—DEME, DIPS, and Dispom—were slower still, see table H.1.

**Parallel computation** Surprisingly, in spite of the natural parallelizability of the MD statistics collection problem<sup>9</sup>, only few MD methods make use of multi-threading offered by most current computers. Aside from Plasma and Discover, the only MD tool considered here that makes use of parallel computation is Dispom. Given how easy parallelization can be leveraged in most programming languages by using libraries such as OpenMP, it would likely not take much work to parallelize the other methods.

---

<sup>9</sup>The problem is “embarrassingly parallel”, as the independence assumption of the individual sequences implies *inter alia* that occurrence statistics for the sequences can be collected in parallel.

## Chapter 22

# Sequence motifs in biological data

This chapter discusses the experiences made by applying discriminative learning methods to real biological data. Section 22.1 discusses MD results for data of the PUF RBP family, section 22.2 those of RBM10, and section 22.3 turns to those of the mouse ESC TF ChIP-Seq experiments.

### 22.1 RIP-Chip and PAR-CLIP data of PUF family RBPs

**Reproducing previous findings** Analysis of the PUF RBP family data demonstrated that the Discover framework is capable of reproducing previous findings regarding the sequence specificity of this well-studied RBP family.

**Showcasing multiple kinds of contrasts** The application to the PUF RBP family showcased the usage of multiple kinds of contrasts, including comparison of bound genes versus unbound ones, of bound genes versus the genomic complement, of multiple groups of genes ranked by binding evidence, as well as the comparison of signal to shuffled sequences.

**Sequence length distributions differences do not impede DMD** The PUF RIP-Chip signal and control data are not perfectly length distribution matched, demonstrating that DMD is feasible also under relaxed constraints regarding equality of signal and control length distributions.

**Benefits of MICO-based  $p$ -values** Using MICO as objective function and the corresponding length corrected  $p$ -values, the MICO MD framework automatically yielded the individual family members' motif length preferences.

**Benefits of PAR-CLIP data** Discover analysis of the PUM2 PAR-CLIP data revealed the relevance of weak-affinity variants that do not conform to the classic PRE UGUAAUA. In particular, the first four nucleotides of the motif are each found to exhibit close to 2 bit IC per position, but the second half exhibits lower position-wise IC. While HAUA is a relevant

IUPAC consensus for the second half, it is perhaps better described as excluding C on its first position, followed by three to four purines.

It is noteworthy that this insight was enabled specifically by two facts: the fine spatial resolution of PAR-CLIP data, and inclusion of lower-ranking PAR-CLIP cluster sequences. As the dilution analysis showed, with larger sequence context, DMD yields high-IC positions also on the second motif half. And, as previous analyses of the same data showed, considering only top-ranking sequences similarly leads to the conclusion of high-IC positions in the second half of the motif.

## 22.2 Alternative splicing regulator RBM10

We performed DMD for the alternative splicing regulator RBM10, jointly analyzing PAR-CLIP repeat experiments, separately for exonic and intronic sequences. The analysis revealed motifs previously implicated in splicing regulation. We also analyzed enrichment of previously reported RBM10 motifs in the same data.

**Earlier RBM10 motif analyses mostly uncorroborated** Earlier analyses revealed that the RanBP2 zinc finger domain of RBM10 binds *in vitro* to single-stranded RNA with the motif AGGUA (Loughlin et al., 2009; Nguyen et al., 2011), a pattern very similar to the conserved metazoan 5' splice site sequence (Ast, 2004; M. Q. Zhang, 1998). However, as these analyses were done *in vitro* and with GST-fusion proteins containing only the zinc finger domain rather than with the full RBM10 proteins including the other RNA-binding domains, the relevance of these insights for *in vivo* RNA-binding specificity of RBM10 is dubious. Yet, Inoue et al. (2014) found that two exons whose splicing is affected by RBM10 carry this motif at their 5' splice sites, and from this concluded this motif to be a RBM10 consensus motif. Another recent analysis by Bechara et al. (2013) performed CLIP-Seq for RBM10 and reported 94 words of length 5—nearly a tenth of all 5mers—to be enriched in the RBM10 CLIP-Seq sequences. The Bechara et al. motifs do not include the purported motif related to the 5' splice site consensus pattern.

We investigated whether these motifs are also enriched in the RBM10 PAR-CLIP data of Y. Wang et al. (2013). We found the motif bound by the zinc finger domain not to be enriched. Similarly, most (~85%) of the words reported by Bechara et al. were either more frequent in shuffled control sequences, or not significantly enriched in the PAR-CLIP sequences of Y. Wang et al. The few Bechara et al. RBM10 motifs that were enriched in the PAR-CLIP sequences are consistent with the motifs that our own analysis revealed.

**Discover reports splicing-relevant motifs in PAR-CLIP data** The most differential motif we found with Discover in the exonic clusters of the Y. Wang et al. (2013) PAR-CLIP data is a purine-rich exonic splicing enhancer (ESE) signal reported to be bound by SFRS1 (Ramchatesingh et al., 1995; Sanford et al., 2009; Tacke and Manley, 1995; X. Wang, Juan, et al., 2011) and by eIF4AIII (Saulière et al., 2012). The most differential intronic motif resembles the signal of the polypyrimidine tract, a signal in the vicinity of the splicing branch point sequence that is bound by U2AF65 and promotes assembly of the spliceosome.



**Correlation with PAR-CLIP conversion events** While both the purine- and the pyrimidine-rich motifs are overall much more frequent in their respective signal sequences than in shuffled control sequences, they are not identically distributed across the ranks of PAR-CLIP conversion ordered sequences. The purine-rich ESE motif found in the exonic sequences is positively correlated with PAR-CLIP conversions, while the pyrimidine-rich one from the intronic sequences is negatively correlated. In other words, the prior is found most frequently in the PAR-CLIP clusters with most conversions, while the latter is more frequent in clusters with fewer conversions.

A mixture of reasons determines the number of conversion events observed in any given cluster. On the one hand, due to stoichiometric reasons, transcripts with a high cellular abundance may more frequently be bound by a given RBP than less highly expressed ones, and thus show elevated levels of PAR-CLIP conversion events. On the other hand, sequences with higher-affinity binding sites would tend to be more tightly and longer bound by a RBP. Thus, in PAR-CLIP datasets the clusters with most conversion events tend to be a mixture of highly expressed transcripts and of high-affinity targets.

Note that sequence composition may also affect PAR-CLIP cross-linking efficiency and thus the number of PAR-CLIP nucleotide conversion events observed in a cluster. E.g. when PAR-CLIP is performed with 4SU—as is the case with the two libraries analyzed here—then sequences that lack U would be at a disadvantage. However, this effect is unlikely to be of consequence here as analysis of an (unpublished) PAR-CLIP library for RBM10 generated with 6SG instead of 4SU yielded the same observations (results not shown).

In summary, the positive correlation with PAR-CLIP conversion events suggests that the purine-rich motif may be tightly bound by RBM10. Conversely, the negative correlation with PAR-CLIP conversions of the pyrimidine-rich motif could be explained by several factors. These include the pyrimidine-rich motif being bound with lesser affinity, perhaps by a different domain of RBM10; or the pyrimidine-rich motif could be indirectly bound by RBM10 due to involvement of RNA secondary structures or interacting proteins. Yet, sequence composition effects could also account for the apparent lower efficiency of induction of PAR-CLIP cross-linking in the vicinity of the pyrimidine-rich motif.

**Interpretation** RBM5, a splicing factor related and highly similar to RBM10, is known to compete for binding to the polypyrimidine tract with U2AF65 (Jin et al., 2012). RBM5 and RBM10 antagonistically regulate proliferative capacity of cancer cells (Bechara et al., 2013) and have different effects on splicing: RBM10 effects mainly exon exclusion (Bechara et al., 2013; Y. Wang et al., 2013), while the main effect of RBM5 appears to be exon-inclusion (Bechara et al., 2013).

The polypyrimidine tract binding protein<sup>1</sup>, PTB, is a multi-functional protein with roles in diverse biological processes including splicing, polyadenylation, mRNA stability, and translation initiation (Sawicka et al., 2008). It binds to intronic pyrimidine-rich sequences and inhibits formation of the U2 snRNP/pre-mRNA complex and thereby splicing (García-Blanco, Jamison, and Sharp, 1989; R. Singh, Valcárcel, and Green, 1995). PTB binds to pyrimidine-rich sequences in a variety of structural contexts (Clerte and Hall,

<sup>1</sup>PTB is also known as heterogeneous nuclear RNP (ribonucleoprotein) I, hnRNP I.

2009), including—intriguingly—the double stranded region of a hairpin secondary structure motif whose one arm consists of pyrimidine-rich sequence, while the other consists of purine-rich sequence (Mitchell et al., 2005). It is possible that similar secondary structure might also be of importance to the regulation exerted by RBM10, which could either favor or disfavor the formation of such hairpins and influence splicing through this mechanism.

Alternatively, it is conceivable that interaction of RBM10 with PTB or other co-factors explains why the pyrimidine-rich motif found is negatively correlated with PAR-CLIP conversions. According to this idea, the sequences most tightly bound by RBM10 would be those containing the purine-rich motif correlated positively with PAR-CLIP conversions, and the sequences containing the pyrimidine-rich motif would be negatively correlated with PAR-CLIP conversions because their association to RBM10 is indirect via interaction with PTB or other factors that are responsible for binding the pyrimidine-rich motif.

**Conclusions** Due to the presence of multiple RNA-binding domains, it would not be surprising for RBM10 to present complex RNA-binding specificities. For this reason, our negative results regarding enrichment in the PAR-CLIP data of Y. Wang et al. for most previously published RBM10 motifs are not necessarily in conflict with earlier analyses (Bechara et al., 2013; Inoue et al., 2014).

Despite the fact that both the purine-rich and the pyrimidine-rich motifs found by Discover in the Y. Wang et al. data are included in the motifs reported by Bechara et al., it seems justified to stress that our findings are first in underlining their central importance for RBM10 binding, given that Bechara et al. report a total of 94 words, only few of which are consistent with our motifs. Bechara et al. do not draw particular attention to the motifs highlighted here, and instead follow up on other motifs that lack—as we showed here—statistical evidence for enrichment in the Y. Wang et al. PAR-CLIP data.

Our results are lent credence by RBM10's role in alternative splicing regulation and the well-established involvement in RNA splicing regulation of the primary motifs we found, the ESE motif in the exonic RBM10 clusters and the pyrimidine-rich motif in the intronic ones. In summary, based on the motif analysis findings, two mechanisms might be responsible for the reported exon-skipping mediated by RBM10: (I) competition of RBM10 with splicing enhancers for the ESE motif, and (II) competition of RBM10 with U2AF65 for binding to the polypyrimidine tract—either through RNA secondary structure or via co-factors.

## 22.3 CHIP-Seq data of mouse ESC TFs

**Coping with the entirety of large datasets** Using contrasting information obviated the need to apply repeat masking or other kinds of filtering to preprocess the data, and MD was applied directly to the significant ChIP-Seq bound regions. Discover leveraged the full size of these datasets that numbered up to 39 609 signal sequences<sup>2</sup>.

---

<sup>2</sup>And equal numbers of shuffled sequences.

**Concordance with earlier analyses** By discriminative learning using MICO as objective function, cognate sequence motifs were successfully *de-novo* discovered for most of the analyzed ChIP-Seq data. In two cases, discriminative statistics provided evidence—in accordance with previous reports—for the absence of purported motifs.

**Cognate motifs are sometimes not the most discriminative ones** In some ChIP-Seq datasets, exemplified by Nanog and Tcf3, the cognate motif is less discriminative than those of co-factors. For such cases, discriminative learning with shuffled sequences as contrast may fail to discover the cognate motifs when only considering the top-ranking motif, or it may report the cognate motifs at a sub-maximal rank when discovering multiple motifs. However, frequently the cognate motifs can then be easily discovered by cross-contrasting these datasets to other datasets that are enriched for the same distracting motifs, as shown in section 20.4.

**Discover results are stringent and robust** The Discover analysis results of ChIP-Seq data appear to be stringent and robust. This is indicated by (I) the strong similarity of multiply discovered motifs, (II) the high proportion of previously described motifs recovered, (III) the high proportion of known co-factor motifs among the co-discovered motifs, and (IV) the consistent results when applied to multiple sets of shuffled sequences.

**Interpreting spatial distribution of motif occurrences** The ChIP-Seq wet-lab procedure and subsequent analysis by peak-calling software are designed to yield genome-wide, localized, quantitative evidence for the binding of the assayed factor. Hence, binding sites of the assayed factor are expected to occur close to the identified ChIP-Seq peak signals. While the exact distance from the peak signal of a ChIP-Seq region to the nearest binding site is a function of many factors<sup>3</sup>, statistics of the distances between sequence midpoints to motif occurrences should distinguish between the motifs of the assayed factor and of co-factors.

Building on this principle, Bailey and Machanick (2012) present a method that uses a database of previously discovered motifs, and calculates for each motif the central enrichment around sequence midpoints of its occurrences in the sequences. Spatial information can also be used for the purposes of MD<sup>4</sup>, but is not (currently) used Discover. Thus, it was available here to validate the discovered motifs as it constitutes evidence orthogonal to the sequence information.

**Motif presence versus *de-novo* discovery** It is important to note the distinction between *de-novo* motif discovery revealing a motif in a given dataset and confirming the presence of a given, known motif in a set of sequences. Clearly, discovering a motif needs more evidence than merely confirming motif presence. Consequently, an MD method can not be expected to co-discover all motifs of co-factors for which association to the assayed factor

<sup>3</sup>Influential factors include stringency of the wet-lab processing, as well as adequacy of the models used by the software to identify enriched regions.

<sup>4</sup>And is indeed used for this purpose for example by the DMD method Dispom (which we excluded from the synthetic data analysis due to its slowness), as well as several other non-discriminative methods.

is known, while it is also to be expected that more motifs' presence can be confirmed than will be discovered by an MD method in a given dataset.

**Discover yields non-redundant, full-length motifs with high true-positive rate** Comparing Discover, DREME, and FIRE DMD results on Oct4 data in section 20.5 showed that DREME and FIRE yield more motifs than Discover, and that these motifs may include presumed-true co-factor motifs that are not identified by Discover. However, the motifs yielded by the regex-based methods are short, redundant and contain motifs that are either not identifiable as known motifs, or that are not reproduced in multiple runs with different sets of shuffled controls. Thus, while potentially missing some true motifs, Discover consistently and robustly identifies a non-redundant set of full-length motifs with a higher true positive rate than competing methods.

## Chapter 23

# Outlook

This chapter gives an overview over possible avenues for further research. We comment on alternative, global learning approaches in section 23.1. Instead of set-based MD approaches, section 23.2 discusses possible usage of mutual information in the context of rank-based MD problems. Section 23.3 mentions theoretical possibilities to reduce the number of learning iterations by using the curvature of probabilistic models. In addition to sequence data alone, MD methods might leverage further information sources, as explained in section 23.4. Finally, section 23.5 mentions settings outside of MD in nucleic acids to which the methods discussed in this thesis could be applied.

### 23.1 Global optimization

Because gradient learning is an inherently local search method, it is desirable to consider alternatives that globally explore parameter space. To this end, Monte-Carlo Markov chain (MCMC) sampling has previously been used for DMD in MoAn (Valen et al., 2009). Aside from gradient optimization, we thus implemented routines in Discover to allow the user to explore MCMC sampling-based optimization of HMM parameters. For this, Discover uses parallel tempering, also known as replica-exchange, to increase the efficiency of the sampling (Earl and Deem, 2005). We found (results not shown) that parallel tempering optimization of HMM parameter is feasible but expectedly much less efficient than gradient optimization, and also computationally more intensive than MCMC optimization of PSSM parameters, as offered in MoAn. Due to the high efficiency we recommend to first try gradient learning, and turn to MCMC sampling when in doubt whether the maximum found by gradient learning is global. Short runs of MCMC sampling can also be used to find starting points for gradient optimization. Furthermore, while the current version of Discover does not support it, gradient optimization could be combined with MCMC sampling in the form of Langevin MCMC (Stramer and Tweedie, 1999).

## 23.2 Rank-based learning - Average rank information

Frequently data from experiments do not present themselves simply in the form of data points with equally strong evidence either supporting or rejecting a certain hypothesis. Rather, for a given sequence to be bound by a certain factor, it is typical that we may have means to associate a quantitative value that measures how strong our believe is for the sequence to be bound or not. While it is in most cases true, that these measures, being based on real data, are far from perfect, they do frequently exhibit correlation with the underlying hypotheses.

In this thesis, we based MD methods on data that is thresholded in different manners, yielding sets of sequences where the sequences within each set are treated interchangeably. For example, we consider sets of positive or negative sequences, or sets of sequences for which evidence favors being bound and sequences for which the evidence is larger for them not to be bound.

Such evidence comes in different kinds of forms. For example, the RIP-Chip RBP data of Morris, Mukherjee, and Keene (2008) presents itself in the form of LOD scores; PAR-CLIP data may be quantified based on the amount of T-to-C conversions observed in the bound regions (Corcoran et al., 2011). ChIP-Seq data is analyzed using tools that employ various probabilistic models that measure enrichment in a signal pull-down over a mock pull-down (e.g. Y. Zhang et al., 2008). This heterogeneity makes it difficult to apply general methods that do not respect specific properties of the particular kind of evidence used.

However, rather than thresholding evidence and treating data in groups resulting from that as interchangeable, MD methods may also try to leverage correlation or association between the quantified evidence and the hypotheses of interest. Such methods include cERMIT (Georgiev et al., 2010) and DRIM (Leibovich, Paz, et al., 2013; Leibovich and Yakhini, 2012).

One particular way of approaching rank-based learning in the context of information theoretic measures is to consider the *average rank information* (RI). Given a ranked set of  $n$  sequences  $(\mathbf{X}_i)_{i=1,\dots,n}$ , we may consider all  $n - 1$  contrasts  $C_i$  that result from splitting the set of sequences in two sets at rank  $i$  for  $1 < i < n$ . Then, the average rank information of a motif  $M$  is defined as the expected value of MICO over all such binary contrasts,

$$\text{RI}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) = \mathbb{E}_i [\text{MICO}(C_i, M)] = \sum_i \frac{1}{n-1} \text{MICO}(C_i, M) = \sum_i \frac{n}{n-1} \mathbb{I}(C_i, M). \quad (23.1)$$

Average rank information has already been implemented in Discoverer, and preliminary evaluation has shown it to be useful for MD (results not shown).

## 23.3 Faster learning

As with all gradient-based methods it may make sense to consider higher order methods to accelerate learning. One such avenue might be to consider the natural gradient, like suggested by Amari and Douglas (1998). With the natural gradient, the parameter updates

of iteration  $k$  from (6.9) take the form

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + n_k \frac{\nabla_{\boldsymbol{\theta}} \mathbb{P}(D|M(\boldsymbol{\theta}))}{\nabla_{\boldsymbol{\theta}}^2 \mathbb{P}(D|M(\boldsymbol{\theta}))}, \quad (23.2)$$

where  $\nabla_{\boldsymbol{\theta}}^2 f(\boldsymbol{\theta})$  is the Hessian matrix of  $f$ . For the step size  $n_k$  the same options exist as explained previously in section 6.2. This method to find maxima of a function is essentially an application of the Newton-Raphson method to the derivative of the function. A requirement for the applicability of this approach is availability of expressions for the Hessian of the objective function. As the computation of the Hessian is computationally more expensive than that of the gradient, the advantage of such approaches would likely hinge on the factors hidden by the Landau notation for runtime complexity.

Alternatively, in the absence of (manageable) expressions for the Hessian, one may resort to methods like the Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963), or quasi-Newton methods like the BFGS method (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970).

## 23.4 Additional information sources

Another avenue of advancement might be to follow the path of the FIRE approach (Elemento, Slonim, and Tavazoie, 2007). FIRE pioneered the usage of mutual information as a general association measure, used as objective function for MD. Importantly, Elemento, Slonim, and Tavazoie use mutual information of the motif occurrence not only in a binary signal-control distinction but also in a multiple class setting. Equally importantly they leverage additional information like strand bias and position bias or covariation with expression to find likely regulatory motifs. Clearly, the same kind of gradient-based approach presented in this thesis is generalizable to make use of such additional information.

Possible sources of additional information include chromatin signatures (Won et al., 2008), sequence conservation or more general phylogenetic information (Siddharthan, Siggia, and Nimwegen, 2005), PAR-CLIP conversion events (Corcoran et al., 2011; Hafner et al., 2010; Majoros et al., 2013), spatial information (Bailey and Machanick, 2012), or RNA-secondary structures (Hiller et al., 2006).

Today, Discover, using e.g. MICO as objective function, can operate on contrasts of many conditions, and can also utilize multiple contrasts, and therefore already supports such kinds of analyses to a certain extent. For this, the user has to split the supplied sequences according to the desired annotation, e.g. discretize some annotation of interest, like expression of RIP-Chip LOD, and group the sequences accordingly. We demonstrated this kind of analysis for the FBF-1 data of Kershner and Kimble (2010) in chapter 18 where we split the data into 15 groups by LOD.

However, Discover does not currently utilize annotation along individual sequences. This could be achieved by considering a product emission space, composed of the nucleic acids emitted at a position as well as discrete or continuous annotated signals. It might also be interesting to investigate whether such additional information sources could au-

tomatically be mined, as FIRE does, to e.g. automatically determine strand bias.

## 23.5 Other applications

Aside from protein binding sites in nucleic acids, discriminative motif learning has already found application in other fields. One example is sequence features in amino acid sequences. Some publications that present DMD methods for amino acid sequences include Lieber, Elemento, and Tavazoie (2010), T.-H. Lin, Murphy, and Bar-Joseph (2011), and Vens, Rosso, and Danchin (2011).

To first approximation, modifying Discover to work on amino acid sequences would merely amount to using a larger alphabet. However, it seems likely that some learning parameters, like pseudo-counts and convergence criteria, might also have to be adapted.



## Chapter 24

# Conclusions

A novel MD method was presented that integrates different generative and discriminative objective functions. The engineering aspects of the software enable analysis of large and complex data setups. Synthetic data allowed controlled performance experiments of our and published MD methods. Our method, Discover, utilizing any of several adequate objective functions, yields the highest MD performance in these experiments. MICO appears the most versatile among the best-performing DMD objective functions, as it is applicable in both discrete and smooth parameter learning settings for seeding and HMM optimization, can leverage contrasts of more than two conditions, as well as allowing joint analysis of repeat experiments. In addition, information theory offers suitable generalizations in the form of cMI, that Discover uses to learn multiple motif models.

Mutual information was introduced to quantify the capacity of noisy signal transmission channels in communications theory (Shannon, 1948). The application to MD in the form of MICO suggests to conceive of regulatory regions in nucleic acids as information transmitting channels. In essence, enhancer and promoter regions mediate inherited control information to specialized receptors: sequence specific DNA-binding TFs. Similarly, stretches of RNA molecules transmit control information to RBPs. Regulatory, nucleic acid binding proteins sample their respective channels, and, upon discovering their cognate signals in the nucleic acid patterns, these proteins bind and thereby initiate the execution of their regulatory purpose.

Analysis of RBP and TF data proved Discover's applicability and utility for real biological data. Motif discovery for PUF family RBPs accorded with earlier reports, and thus validated the presented MD framework. Dilution analyses explained differences in motif discovery results between RIP-Chip and PAR-CLIP data, implying that the finer spatial resolution of PAR-CLIP translates into refined motif models. For the alternative-splicing regulator RBM10 we discovered *de-novo* known splicing-relevant motifs whose importance was previously not adequately appreciated. The analysis of ChIP-Seq datasets of mouse ESC TFs proved Discover's multiple motif mode to robustly discover non-redundant sets of full-length motifs with a high true-positive rate.

In conclusion, the presented MD method Discover is expected to be a highly useful tool to decipher sequence binding specificities from the expanding amounts data generated for nucleic acid binding proteins.



# **Appendices**



# Appendix A

## Proof of correctness of the scaling procedure

This appendix chapter presents a proof for the correctness of the scaling procedure presented in section 4.5.

### A.1 Correctness for the forward matrix recursion

We want to show that

$$\alpha_t(i) = \tilde{\alpha}_t(i) \prod_{k=0}^t s_k. \quad (\text{A.1})$$

The proof is by induction on the time of observation  $t$ .

For the initial step of the proof we observe that the initialization is identical between the scaled and the unscaled variants. The identity thus trivially holds for  $t = 0$ .

Assume now that it holds for some  $t$  and all  $1 \leq j \leq N$  that

$$\alpha_t(j) = \tilde{\alpha}_t(j) \prod_{k=0}^t s_k. \quad (\text{A.2})$$

From the definition of  $\alpha_{t+1}(j)$  we have for  $1 \leq j \leq N$

$$\begin{aligned} \alpha_{t+1}(j) &= b_{t+1}(j) \sum_{i=1}^N \alpha_t(i) a_{ij} = b_{t+1}(j) \sum_{i=1}^N \tilde{\alpha}_t(i) a_{ij} \prod_{k=0}^t s_k = \hat{\alpha}_{t+1}(j) \prod_{k=0}^t s_k \\ &= \frac{\hat{\alpha}_{t+1}(j)}{s_{t+1}} \prod_{k=0}^{t+1} s_k = \tilde{\alpha}_{t+1}(j) \prod_{k=0}^{t+1} s_k. \end{aligned} \quad (\text{A.3})$$

This shows that if the relation holds for any time  $t$  then it also holds for time  $t + 1$ .

Together with the initial step this concludes this proof by induction.  $\square$

## A.2 Correctness for the backward matrix recursion

We want to show that

$$\beta_t(i) = \tilde{\beta}_t(i) \prod_{k=t}^{T+1} s_k. \quad (\text{A.4})$$

The proof is by induction on the time of observation  $t$ , starting with the last position.

Again, the basis of the induction is trivially given, because initialization of the algorithms to compute scaled and unscaled matrices are identical.

Assume now that it holds for some  $t + 1$  and all  $1 \leq i \leq N$  that

$$\beta_{t+1}(i) = \tilde{\beta}_{t+1}(i) \prod_{k=t+1}^{T+1} s_k. \quad (\text{A.5})$$

Then we have for  $1 \leq i \leq N$

$$\begin{aligned} \beta_t(i) &= \sum_{j=1}^N a_{ij} b_j(X_{t+1}) \beta_{t+1}(j) = \sum_{j=1}^N a_{ij} b_j(X_{t+1}) \tilde{\beta}_{t+1}(i) \prod_{k=t+1}^{T+1} s_k = \hat{\beta}_t(i) \prod_{k=t+1}^{T+1} s_k \\ &= \frac{\hat{\beta}_t(i)}{s_t} \prod_{k=t}^{T+1} s_k = \tilde{\beta}_t(i) \prod_{k=t}^{T+1} s_k. \end{aligned} \quad (\text{A.6})$$

This shows that if the relation holds for any time  $t + 1$  then it also holds for time  $t$ .

Together with the initial step this concludes this proof by induction.  $\square$

## Appendix B

# Runtime of HMM inference algorithms

This chapter explains how the runtime of the HMM inference algorithms presented in chapter 4 can be reduced. The textbook runtime for the Viterbi algorithm, and for the algorithms to compute the forward and backward variables are all  $\mathcal{O}(TN^2)$ , where  $T$  is the length of the data and  $N$  is the number of states in the HMM. This appendix shows how the computations of these algorithms can be arranged so as to yield a runtime of  $\mathcal{O}(TE)$  where  $E$  is the number of transitions (edges) in the HMM whose corresponding transition probability is non-zero. For general HMMs with a connected topology  $N - 1 \leq E \leq N^2$ .

The upper limit is reached with equality for HMMs that have a complete graph as topology, the lower limit is reached with equality if the HMM is a chain of states of which each has one incoming and one outgoing edge, except for the start and end state, which have one outgoing, and one incoming, respectively.

The runtime reduction is illustrated by contrasting a HMM with topology of a complete graph (figure B.1a), and one which is close to the lower limit of edges (figure B.1b).

If we want to compute the Viterbi algorithm, or the forward or backward matrix calculations on these two HMMs we need to visit each state  $j$  at each point in time  $t$  yielding a runtime of  $\mathcal{O}(TN)$ . Then, for each state  $j$  and time  $t$  we need to inspect all possible states  $i$  at time  $t - 1$  for the Viterbi algorithm or for the forward matrix calculation. Similarly, for the backward matrix calculation for each state  $i$  and time  $t$  we need to inspect all possible states  $j$  at time  $t + 1$ . The trellis structure corresponding to this is shown in figure B.2.

The important point is that only those states  $i$  need to be considered as predecessors, or successors, whose transitions probabilities to or from the state  $j$ ,  $a_{ij}$  are non-zero, and thus the runtime can be reduced when the HMM topology is not a complete graph. When there are  $E$  such transitions with non-zero probability in the HMM, for each time  $t$  only exactly  $E$  predecessor states at time  $t - 1$  or time  $t + 1$  behind the  $E$  transitions need to be considered, yielding a runtime of  $\mathcal{O}(TE)$ . The resulting trellis for the less connected HMM is shown in figure B.3.

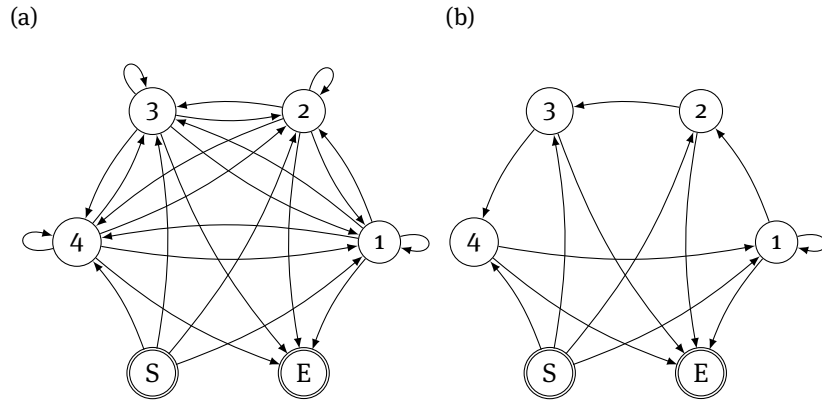


Figure B.1: The transition graphs of two HMMs. (a) An HMM with a full transition topology, with four states and 16 transitions, in which each state may transition to each other state. (b) An HMM with a four states but only five possible transitions.

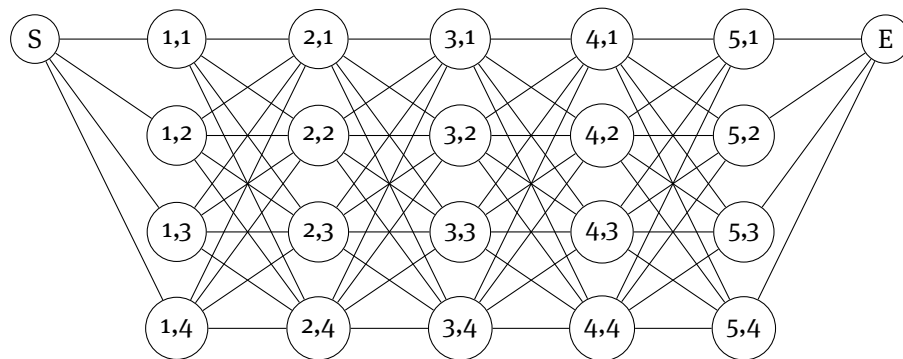


Figure B.2: The trellis structure of the Viterbi, forward and backward algorithms for a sequence of length 5 on the HMM shown in figure B.1a.

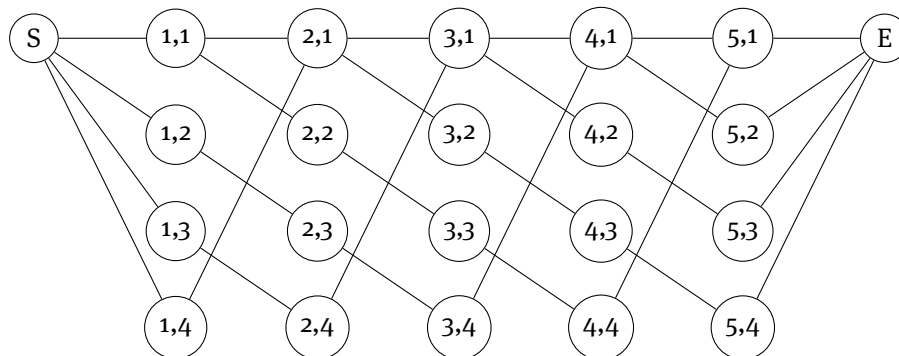


Figure B.3: The trellis structure of the Viterbi, forward and backward algorithms for a sequence of length 5 on the HMM shown in figure B.1b.



# Appendix C

## Information theory

This chapter summarizes the fundamental concepts of information theory. For more comprehensive discussion of these topics the reader is referred to textbooks (Cover and Thomas, 2006; Kullback, 1959; MacKay, 2003).

### C.1 Communication systems

The subject of information theory, as defined by Shannon (1948), are communication systems. As depicted in figure C.1, communication systems are conceived of as a process involving as entities an information source, a transmitter, a channel, a receiver, and a destination, as well as a noise source.

The source is modeled as a probabilistic process that generates messages or sequences of messages from a set of possible messages. In turn, the messages generated by the source are encoded by a transmitter into a form suitable for transmission over the channel. On the receiving side of the channel the encoding of the transmitter is inverted by a receiver to yield the received message. More modern terms for transmitter and receiver are encoder and decoder, respectively.

As indicated, noise may perturb the transmitted signal, thus causing reception of modified signals and consequently differences may exist between the source and received message. Note that noise may affect any of the processes from generation of the message, over encoding, transmission, reception, and decoding. It is however sufficient for a general discussion to assume the noise source to act on the channel.

### C.2 Fundamental quantities of information theory

In order to be able reason about fidelity of information transmission it is necessary to be able to quantify it. Information entropy and the related quantities that we define below have certain beneficial properties making them uniquely suitable to this task, see (Shannon, 1948).

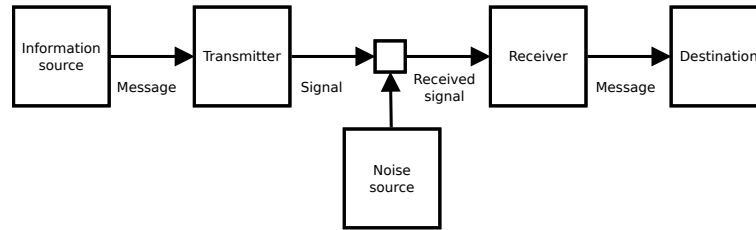


Figure C.1: Schematic diagram of a general communication system; reproduction of figure 1 of Shannon (1948).

### C.2.1 Information entropy

Information entropy  $\mathbb{H}(X)$  quantifies the uncertainty about the realization of a random variable  $X$ ,

$$\mathbb{H}(X) = - \sum_{x \in X} \mathbb{P}(x) \log_2 \mathbb{P}(x). \quad (\text{C.1})$$

Similarly, by considering the joint state space of two variables  $X$  and  $Y$ , the joint information entropy of  $X$  and  $Y$  is defined as

$$\mathbb{H}(X, Y) = - \sum_{\substack{x \in X \\ y \in Y}} \mathbb{P}(x, y) \log_2 \mathbb{P}(x, y). \quad (\text{C.2})$$

By referring to the negative logarithmic probability  $-\log_2 \mathbb{P}(x)$  as the information of  $\mathbb{P}(x)$  we may note that the information entropy of  $X$  corresponds to the expected information over  $X$ .

**Units of information** Note that the choice of the logarithmic base determines the units in which information entropy, and the other quantities defined below, are measured. The choice of logarithms to the base 2 corresponds to units of bit, an abbreviation for binary digits, a term proposed by J. W. Tukey, as Shannon noted. Other choices like the natural logarithm yield units of nats, which is short for natural digits. If logarithmic base 10 is used the units are referred to as bans<sup>1</sup>, or Hartley<sup>2</sup>.

### C.2.2 Conditional information entropy

Conditional information entropy  $\mathbb{H}(X|Y)$  is a measure for how much uncertainty about  $X$  is remaining when  $Y$  is known,

$$\mathbb{H}(X|Y) = - \sum_{\substack{x \in X \\ y \in Y}} \mathbb{P}(x, y) \log_2 \mathbb{P}(x|y). \quad (\text{C.3})$$

<sup>1</sup>The term ban refers to the cryptanalytic procedure Banburismus developed by Alan Turing and I. J. Good at Bletchley Park, close to Banbury.

<sup>2</sup>Ralph Hartley had been working in the field of information theory already prior to Shannon's seminal paper (Hartley, 1928).

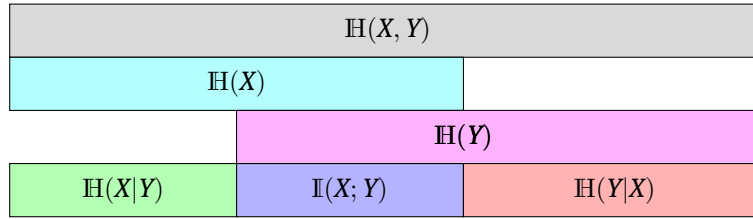


Figure C.2: Relationship of information theoretic quantities.

When  $X$  is the source message and  $Y$  the received message, Shannon used the term *equivocation* for  $\mathbb{H}(X|Y)$  to label the average ambiguity of received messages.

### C.2.3 Mutual information

Mutual information  $\mathbb{I}(X; Y)$  is a measure of the reduction of uncertainty about one variable when the state of the other variable is known,

$$\mathbb{I}(X; Y) = \sum_{\substack{x \in X \\ y \in Y}} P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}. \quad (\text{C.4})$$

Shannon defined mutual information to capture the actual rate of transmission through the channel. In turn, considering the maximal possible mutual information through a channel allowed Shannon to define the channel's capacity.

The symmetry  $\mathbb{I}(X; Y) = \mathbb{I}(Y; X)$  of mutual information is trivially given by the definition.

### C.2.4 Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence of two probability distributions  $P$  and  $Q$  (Kullback and Leibler, 1951), also known as relative entropy, is defined as

$$D_{\text{KL}}(P||Q) = \sum_{x \in X} P(x) \log_2 \frac{P(x)}{Q(x)}. \quad (\text{C.5})$$

Due to Gibbs inequality (Gibbs, 1902), KL divergences are never negative,

$$D_{\text{KL}}(P||Q) \geq 0. \quad (\text{C.6})$$

Gibbs inequality can be proven using an inequality due to Jensen (1906), see (Dembo, Cover, and Thomas, 1991). Note that mutual information is the KL divergence of the joint and independence models, where  $P(X, Y) = P(X, Y)$  and  $Q(X, Y) = P(X)P(Y)$ ,

$$\mathbb{I}(X; Y) = D_{\text{KL}}(P(X, Y)||P(X)P(Y)). \quad (\text{C.7})$$

### C.2.5 Relationships between information theoretic quantities

The information theoretic quantities defined in the preceding subsections are interrelated in various ways. Among them are the following relations:

$$\mathbb{H}(X|Y) = \mathbb{H}(X, Y) - \mathbb{H}(Y), \quad (\text{C.8})$$

and symmetrically,

$$\mathbb{H}(Y|X) = \mathbb{H}(X, Y) - \mathbb{H}(X). \quad (\text{C.9})$$

Mutual information may be related to these quantities as follows:

$$\mathbb{I}(X; Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) \quad (\text{C.10})$$

$$= \mathbb{H}(X) - \mathbb{H}(X|Y) \quad (\text{C.11})$$

$$= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X, Y). \quad (\text{C.12})$$

These relations are displayed graphically in figure C.2.

### C.2.6 Conditional mutual information

We will also make use of conditional mutual information (cMI) (Cover and Thomas, 2006). The cMI of a variable  $X$  and a variable  $Y$  given a variable  $Z$ ,  $\mathbb{I}(X; Y|Z)$ , is defined as,

$$\begin{aligned} \mathbb{I}(X; Y|Z) &= \sum_{\substack{x \in X \\ y \in Y \\ z \in Z}} \mathbb{P}(x, y, z) \log_2 \frac{\mathbb{P}(x, y|z)}{\mathbb{P}(x|z)\mathbb{P}(y|z)} \\ &= \sum_{\substack{x \in X \\ y \in Y \\ z \in Z}} \mathbb{P}(x, y, z) \log_2 \frac{\mathbb{P}(x, y, z)\mathbb{P}(z)}{\mathbb{P}(x, z)\mathbb{P}(y, z)}. \end{aligned} \quad (\text{C.13})$$

### C.2.7 Chain rules

There are chain rules for the calculation of joint information entropy and mutual information. The chain rule for joint information entropy generalizes equations (C.8) and (C.9),

$$\mathbb{H}(X_1, X_2, \dots, X_n) = \sum_{i=1}^n \mathbb{H}(X_i | X_{i-1}, \dots, X_1). \quad (\text{C.14})$$

The corresponding chain rule for joint mutual information is

$$\mathbb{I}(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n \mathbb{I}(X_i; Y | X_{i-1}, \dots, X_1). \quad (\text{C.15})$$

For proofs of equations (C.14) and (C.15) see section 2.5 of Cover and Thomas (2006).

## Appendix D

# Likelihood ratio, mutual information and $\chi^2$ statistic

If  $X_1, \dots, X_k$  are  $k$  independent, standard normally distributed random variables, then the sum of their squares,

$$Z = \sum_{i=1}^k X_i^2, \quad (\text{D.1})$$

is distributed according to the  $\chi^2$  distribution with  $k$  degrees of freedom,

$$Z \sim \chi^2(k) \quad (\text{D.2})$$

The probability density function of the  $\chi^2$  distribution with  $k$  degrees of freedom is given by

$$f(x, k) = \frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, \quad (\text{D.3})$$

where  $\Gamma(x)$  is the Gamma function. Figure D.1 displays the probability density function for different degrees of freedom.

### D.1 The likelihood ratio statistic for goodness of fit

When considering the likelihood ratio statistic for goodness of fit we first need to define the null and alternative hypotheses. The null hypothesis is that a categorical random variable is distributed according to a given distribution  $\mathbf{p}_0$ ,

$$H_0 : \mathbf{p} = \mathbf{p}_0. \quad (\text{D.4})$$

This is to be contrasted with the alternative hypothesis that  $\mathbf{p}$  assumes a different value,

$$H_A : \mathbf{p} \neq \mathbf{p}_0. \quad (\text{D.5})$$

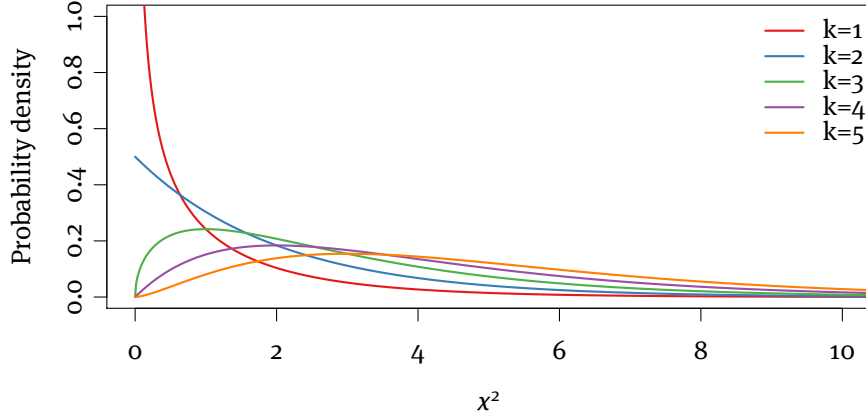


Figure D.1: Probability density function of the  $\chi^2$  distribution for different degrees of freedom.

When observing a multinomial sample of counts  $\mathbf{f}$ , the likelihood given the null hypothesis is

$$\mathbb{L}(\mathbf{p}_0) = \mathbb{P}(\mathbf{f}|\mathbf{p}_0) = \frac{n!}{\prod_{j=1}^k f_j!} \prod_{j=1}^k p_j^{f_j}. \quad (\text{D.6})$$

For the alternative hypothesis we evoke the maximum likelihood principle and define the likelihood to be

$$\mathbb{L}(\hat{\mathbf{p}}) = \mathbb{P}(\mathbf{f}|\hat{\mathbf{p}}) = \frac{n!}{\prod_{j=1}^k f_j!} \prod_{j=1}^k \left(\frac{f_j}{n}\right)^{f_j}. \quad (\text{D.7})$$

Following Lindgren (1993 and 1998), the likelihood ratio statistic  $\Lambda$  is given by

$$\Lambda = \frac{\mathbb{L}(\mathbf{p}_0)}{\mathbb{L}(\hat{\mathbf{p}})} = \prod_{j=1}^k \frac{(p_j)^{f_j}}{\left(\frac{f_j}{n}\right)^{f_j}}. \quad (\text{D.8})$$

Consequently, the log likelihood ratio statistic is

$$\log \Lambda = \sum_{j=1}^k f_j \log \frac{p_j}{\frac{f_j}{n}} = \sum_{j=1}^k f_j \log \frac{np_j}{f_j} = \sum_{j=1}^k f_j \log \frac{\tilde{f}_j}{f_j}. \quad (\text{D.9})$$

Here,  $\tilde{\mathbf{f}}$  is the vector of expected counts under the null hypothesis. Clearly,  $\Lambda \leq 1$  and  $\log \Lambda \leq 0$ , as the likelihood of the restricted null hypothesis must always be less or equal to that of the unrestricted alternative hypothesis.

When the null hypothesis is true and for large sample sizes  $n$ , the log likelihood ratio statistic  $-2 \log \Lambda$  is distributed like  $\chi^2$  with a suitable number of degrees of freedom. Named after Samuel S. Wilks, this fact is known as Wilks' theorem (Dudley, Spring 2003; Wilks, 1938, 1962).

## D.2 G-test

When applied to test independence in a contingency tables, the likelihood ratio statistic is also known as G-test (Sokal and Rohlf, 1969). The value of the G-test is given by

$$G = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}, \quad (\text{D.10})$$

and the summation is over all non-empty cells of the contingency table,  $O_{ij}$  are the observed counts of the cell in row  $i$  and column  $j$ , and  $E_{ij}$  are the corresponding expected counts, and are calculated according to

$$E_{ij} = \frac{1}{n} \left( \sum_k O_{ik} \right) \left( \sum_k O_{kj} \right). \quad (\text{D.11})$$

The relation of the G-test and the likelihood ratio test is

$$G = -2 \log \Lambda. \quad (\text{D.12})$$

Together with Wilks's theorem this shows that the G-test statistic approximates the  $\chi^2$  distribution with  $(n_{\text{row}} - 1) \times (n_{\text{columns}} - 1)$  degrees of freedom.

## D.3 Likelihood ratio and mutual information

We consider the likelihood ratio statistic test for a null hypothesis of independence of two variables  $X$  and  $Y$ ,

$$H_0 : X \perp Y \Leftrightarrow \mathbb{P}(x, y) = \mathbb{P}(x)\mathbb{P}(y), \quad \text{for all } x, y. \quad (\text{D.13})$$

Then, the log likelihood ratio statistic is

$$\log \Lambda = \sum_{j=1}^J f_j \log \frac{p_j}{\frac{f_j}{n}}. \quad (\text{D.14})$$

From this the relation of the likelihood ratio statistic and mutual information is apparent,

$$-\frac{1}{n} \log \Lambda = \sum_{j=1}^J \frac{f_j}{n} \log \frac{f_j}{p_j} = D_{\text{KL}}(\mathbf{p} \parallel \mathbf{p}_0) = \mathbb{I}_{\mathbf{p}}(X; Y). \quad (\text{D.15})$$





## Appendix E

# Limits of Matthews correlation coefficient

This appendix demonstrates that the limiting value of the Matthew's correlation coefficient is zero, when two of the statistics that it is based on approach zero.

First, let us recapitulate the definition of the  $mCC$ :

$$mCC = \frac{nTP \cdot nTN - nFP \cdot nFN}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}}. \quad (E.1)$$

For the sake of brevity of notation, in this appendix we will refer to the number of true and false positives,  $nTP$  and  $nFP$ , as  $a$  and  $b$ , and those of true and false negatives,  $nTN$  and  $nFN$ , as  $d$  and  $c$ , respectively. Thus, we have

$$mCC = \frac{a \cdot d - b \cdot c}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}. \quad (E.2)$$

A zero factor appears in the denominator when at least one of  $a$  or  $d$ , and at least one of  $b$  or  $c$  become zero. Without loss of generality let us assume that  $a$  and  $b$  approach zero. In fact, let us first assume that  $a$  is zero, and  $b$  approaches zero. We then have

$$mCC = \frac{-b \cdot c}{\sqrt{b \cdot c(b + d)(c + d)}}. \quad (E.3)$$

Now, by the rule of L'Hôpital we have

$$\lim_{b \rightarrow 0} mCC = \lim_{b \rightarrow 0} \frac{-b \cdot c}{\sqrt{b \cdot c(b + d)(c + d)}} \quad (E.4)$$

$$= \lim_{b \rightarrow 0} \frac{-c}{\frac{1}{2} \sqrt{c} \sqrt{c + d} \frac{1}{\sqrt{b^2 + b \cdot d}} (2b + d)} \quad (E.5)$$

$$= \lim_{b \rightarrow 0} \frac{-2c \sqrt{b^2 + b \cdot d}}{\sqrt{c} \sqrt{c + d} (2b + d)} \quad (E.6)$$

$$= \frac{0}{\sqrt{c} \sqrt{c + d} \cdot d} = 0. \quad (E.7)$$

Let us now study the case when we take  $b$  to be zero, and letting  $a$  approach zero. In this case we have

$$mCC = \frac{a \cdot d}{\sqrt{a \cdot d(a+c)(c+d)}}. \quad (\text{E.8})$$

Also in this case the rule of L'Hôpital gives

$$\lim_{a \rightarrow 0} mCC = \lim_{a \rightarrow 0} \frac{a \cdot d}{\sqrt{a \cdot d(a+c)(c+d)}} \quad (\text{E.9})$$

$$= \lim_{a \rightarrow 0} \frac{d}{\frac{1}{2}\sqrt{d}\sqrt{c+d} \frac{1}{\sqrt{a^2+a \cdot c}} (2a+c)} \quad (\text{E.10})$$

$$= \lim_{a \rightarrow 0} \frac{2d\sqrt{a^2+a \cdot c}}{\sqrt{d}\sqrt{c+d} (2a+c)} \quad (\text{E.11})$$

$$= \frac{0}{\sqrt{d}\sqrt{c+d} \cdot c} = 0. \quad (\text{E.12})$$

## Appendix F

# Gradient calculus

In this chapter we derive an expression for the gradient of MICO, starting from equation (8.12), which we repeat here:

$$\begin{aligned} \mathbb{I}(C; M) = \log_2 N + \frac{1}{N} & \left( \sum_{i=1}^k m_i(\boldsymbol{\theta}) \log_2 \frac{m_i(\boldsymbol{\theta})}{n_i \sum_{j=1}^k m_j(\boldsymbol{\theta})} \right. \\ & \left. + \sum_{i=1}^k (n_i - m_i(\boldsymbol{\theta})) \log_2 \frac{n_i - m_i(\boldsymbol{\theta})}{n_i \sum_{j=1}^k (n_j - m_j(\boldsymbol{\theta}))} \right), \end{aligned} \quad (\text{F.1})$$

where  $N = \sum_{j=1}^k n_j$ ,  $n_i$  is the number of sequences in condition  $i$ , and  $m_i(\boldsymbol{\theta})$  is the number of sequences in condition  $i$  with at least one motif occurrence.

For the purpose of this chapter, we will suppress the notational expression of the dependence on the parameters  $\boldsymbol{\theta}$  of the number of sequences in condition  $i$  with at least one motif occurrence,  $m_i(\boldsymbol{\theta})$ , and simply use  $m_i$ . Later, when we determine the gradient of expressions based on this, we will remember that  $m_i$  depends on the parameters, but not  $n_i$ .

Also, as the sums are generally running from 1 to  $k$ , where  $k$  is the number of conditions in the contrast, we will in this chapter avoid writing out the limits of summation, and only indicate the summation variable.

$$\begin{aligned} \mathbb{I}(C; M) = \log_2 N + \frac{1}{N} & \left( \sum_i m_i \log_2 \frac{m_i}{n_i \sum_j m_j} \right. \\ & \left. + \sum_i (n_i - m_i) \log_2 \frac{n_i - m_i}{n_i \sum_j (n_j - m_j)} \right) \end{aligned} \quad (\text{F.2})$$

$$\begin{aligned} = \log_2 N + \frac{1}{N} & \left( \sum_i m_i \log_2 \frac{m_i}{\sum_j m_j} - \sum_i m_i \log_2 n_i \right. \\ & \left. + \sum_i (n_i - m_i) \log_2 \frac{n_i - m_i}{\sum_j (n_j - m_j)} - \sum_i (n_i - m_i) \log_2 n_i \right) \end{aligned} \quad (\text{F.3})$$

$$\begin{aligned}
 &= \log_2 N + \frac{1}{N} \left( \sum_i m_i \log_2 \frac{m_i}{\sum_j m_j} \right. \\
 &\quad \left. + \sum_i (n_i - m_i) \log_2 \frac{n_i - m_i}{\sum_j (n_j - m_j)} \right. \\
 &\quad \left. - \sum_i n_i \log_2 n_i \right). \tag{F.4}
 \end{aligned}$$

We now consider the gradient of equation (F.4), remembering that the number of sequences in condition  $i$ ,  $n_i$ , are not dependent on the parameters, and thus  $\nabla n_i = 0$ .

$$\begin{aligned}
 \nabla \mathbb{I}(C; M) &= \frac{1}{N} \left( \sum_i (\nabla m_i) \log_2 \frac{m_i}{\sum_j m_j} + \sum_i m_i \nabla \log_2 \frac{m_i}{\sum_j m_j} \right. \\
 &\quad \left. + \sum_i (\nabla (n_i - m_i)) \log_2 \frac{n_i - m_i}{\sum_j (n_j - m_j)} + \sum_i (n_i - m_i) \nabla \log_2 \frac{n_i - m_i}{\sum_j (n_j - m_j)} \right) \tag{F.5}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{N} \left( \sum_i (\nabla m_i) \log_2 \frac{m_i}{\sum_j m_j} + \sum_i m_i \nabla \log_2 \frac{m_i}{\sum_j m_j} \right. \\
 &\quad \left. - \sum_i (\nabla m_i) \log_2 \frac{n_i - m_i}{\sum_j (n_j - m_j)} + \sum_i (n_i - m_i) \nabla \log_2 \frac{n_i - m_i}{\sum_j (n_j - m_j)} \right) \tag{F.6}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{N} \left( \sum_i (\nabla m_i) \log_2 \frac{m_i}{n_i - m_i} - \sum_i (\nabla m_i) \log_2 \frac{\sum_j m_j}{\sum_j (n_j - m_j)} \right. \\
 &\quad \left. + \sum_i \nabla m_i - \sum_i \nabla m_i + \sum_i \nabla (n_i - m_i) - \sum_i \nabla (n_i - m_i) \right) \tag{F.7}
 \end{aligned}$$

$$= \frac{1}{N} \left( \sum_i (\nabla m_i) \log_2 \frac{m_i}{n_i - m_i} - \left( \sum_i \nabla m_i \right) \log_2 \frac{\sum_i m_i}{\sum_i (n_i - m_i)} \right). \tag{F.8}$$

## Appendix G

# Running parameters

The synthetic data were analyzed with a fixed motif length of eight nucleotides. Below, SIGNAL and CONTROL denote the respective paths of signal and control FASTA files. OUTPUT is the path for the output file(s).

### Discover

Discover (vo.3-306-g6b8b582) was run with the following command line:

```
discover --score SCORE signal:SIGNAL control:CONTROL -o OUTPUT -m signal:8
```

SCORE stands for the objective function, and is one of bw, dfreq, dlogl, mcc, mi, or mmie. The switch `-m signal:8` lets Discover search for a motif of length 8 named signal. The signal sequence file SIGNAL is annotated with signal, to indicate that the motif named signal is supposed to be enriched in it. The label control for the control sequence file CONTROL indicates that this file is not to be used for learning the generative parameters in hybrid learning.

### CMF

CMF (patched version, see appendix [L.1](#)) was run with the following command line:

```
cmf -f OUT_DIR -o SEED_OUTPUT -i1 SIGNAL -i2 CONTROL -w 8 > OUTPUT
```

The option `-w 8` signifies that only motifs of length eight nucleotides are sought. OUT\_DIR specifies a directory where output files will be created and SEED\_OUTPUT gives a path to which seed statistics are written.

CMF does not support single stranded MD analysis.

### DECOD

DECOD (v1.01) was run with the following command line:

```
java -Xms2G -Xmx2G -jar DECOD-20110613.jar -nogui -pos SIGNAL -neg CONTROL \  
-w 8 -nmotif 1 -o OUTPUT -strand forward
```

The options `-Xms2G` and `-Xmx2G` set the initial and maximal size of the memory allocation pool to 2 GiB. The option `-w 8` signifies that only motifs of length eight nucleotides are sought.

## DEME

DEME (v1.0) was run with the following command line:

```
deme -p SIGNAL -n CONTROL -w 8
```

The option `-w 8` signifies that only motifs of length eight nucleotides are sought. By default, DEME searches only the forward strand for motifs.

## DIPS

DIPS (v1.1) was run with the following command line:

```
dmotif -positive SIGNAL -negative CONTROL -len 8
```

The option `-len 8` specifies that only motifs of length eight nucleotides are sought. DIPS can only perform double-stranded MD analysis.

## Dispom

Dispom (unversioned) was run with the following command line:

```
java -Xms1G -Xmx4G -jar Dispom.jar home=OUT_DIR fg=SIGNAL bg=CONTROL \  
  bothStrands=false init=best-random=100 p-val=1E-4 position=UNIFORM \  
  length=8 > OUTPUT
```

The option `length=8` specifies that only motifs of length eight nucleotides are sought. `OUT_DIR` specifies a directory in which output files will be generated. Single stranded motif discovery mode is selected by the switch `bothStrands=false`. The switches `-Xms1G` and `-Xmx4G` instruct the Java virtual machine to use 1 GiB of memory initially and 4 GiB maximally, respectively. `position=UNIFORM` uses a flat spatial motif occurrence distribution prior.

## DME

DME (v2) was run with the following command line:

```
dme SIGNAL -b CONTROL -w 8 > OUTPUT
```

The option `-w 8` signifies that only motifs of length eight nucleotides are sought.

DME does not support single stranded MD analysis.

## DREME

DREME (v4.9.1) was run with the following command line:

```
dreme -norc -m 1 -oc OUT_DIR -p SIGNAL -n CONTROL -k 8 > OUTPUT
```

The option `-k 8` signifies that only motifs of length eight nucleotides are sought. `OUT_DIR` specifies a directory where output files will be created.

---

## FIRE

FIRE (v1.1a) was run with the following command line:

```
cat SIGNAL CONTROL | upseq.rb > JOINT
prep-fire.rb SIGNAL CONTROL EXPR
perl fire.pl -k 8 --dodna=0 --expfiles=EXPR --exptype=discrete \
  --fastafire_rna=JOINT --seqlen_rna=LENGTH --nodups=1 > OUTPUT
```

First, both signal and control sequences are transformed to upper case<sup>1</sup> and concatenated into the file JOINT. Then, a script prep-fire.rb is run that creates an annotation file EXPR that is used to inform FIRE which sequences belong to which class. Finally, FIRE is run in RNA mode: the switch `dodna=0` instructs not to perform a double-stranded analysis. The option `-k 8` signifies that only motifs of length eight nucleotides are sought. FIRE also needs to be informed of the length of the sequences via the switch `seqlen_rna`. The switch `nodups` tells FIRE not to try to remove duplicate sequences.

## MDscan

MDscan (unversioned) was run with the following command line:

```
moan -i SIGNAL -w 8 -t 100 > OUTPUT
```

The option `-w 8` signifies that only motifs of length eight nucleotides are sought. The switch `-t 100` instructs MDscan to consider the top 100 sequences to look for candidate motifs (default=5).

MDscan does not support single stranded MD analysis.

## MoAn

MoAn (v1.01, patched to fix a problem with the random number generation routine, see appendix L.1) was run with the following command line:

```
moan SIGNAL CONTROL -R 8,8 -i ITERATIONS > OUTPUT
```

The option `-R 8,8` signifies that only motifs of length eight nucleotides are sought. MoAn by default searches the forward strand only and performs  $3 \times 10^7$  iterations. By using the switch `-i ITERATIONS` MoAn was run with  $3 \times 10^6$  iterations, a tenth of the default value.

---

<sup>1</sup>FIRE ignores lower case regions of sequences.





## Appendix H

# Synthetic data experiments

This chapter gives supplementary results for the synthetic data experiments presented in chapter 17.

Table H.1 compares the runtime of several MD methods on one particular dataset. Compared are Discover, MoAn-3M, MoAn, DEME, DIPS, and Dispom. As is visible, with MoAn, by reducing the number of iterations to a tenth of the default number (MoAn-3M), the runtime is similarly decreased about tenfold. The methods DEME, DIPS, and Dispom all run more than 1000 times as long as Discover. As Discover needs a total of 16.9 h on the synthetic datasets (figure H.3), based on this estimate, these three methods would be expected to run more than 16 900 h on the synthetic dataset, or more than 1.92 years. For this reason they were excluded from the analysis presented in chapter 17.

In addition to the nCC, as shown in figure 17.1, figure H.1 presents site-level supervised metrics for MD performance, including sSn, sPPV, sAP, and sF<sub>1</sub>. The MD performance in terms of nCC of additional methods is shown in figure H.2, and numerically in table H.2. Figure H.3 gives runtimes of the considered methods. Figures H.4 to H.6 display sAP, sSn, and sPPV summarized in the same manner as nCC in figure 17.2. Finally, Figure H.7 illustrates the effect of discriminative significance filtering. Particularly, the non-discriminative methods profit from discriminative filtering.

Table H.1: Runtime of several motif discovery method on one pair of signal and control sequence sets. We considered one particular motif discovery experiment from the basic dataset with 10 000 signal and 10 000 control sequences each of length 1000 with a motif implantation probability of 1 % at an information content of 1.4 bit. Times are given as hours:minutes:seconds. The last two columns give average CPU utilization in percent of a single CPU, and wall clock time relative to that of Discrover. Experiments were run on an Intel® Xeon® E5645 CPU running at 2.40GHz with 12 CPU cores. Note: DEME did not finish after more than 74 days, 20 hours. Note: DIPS experienced contention for the used CPU, otherwise wall clock time would be around the same as the CPU time.

Method	Wall clock	CPU time	CPU [%]	Relative wall clock
Discrover	00:01:59	00:08:12	413	1.00
DEME	> 74 days	> 74 days	100	> 54330.59
DIPS	628:27:21	605:10:46	96	19008.43
Dispom	40:09:41	88:09:16	219	1214.73
MoAn-3M	00:45:31	00:45:26	100	22.95
MoAn	08:20:57	08:20:04	100	252.53

Table H.2: Motif discovery performance. nCC: nucleotide-level Matthews correlation coefficient, %: nCC relative to recognizability. NA: not available. Numbers are plotted in figure H.2.

	Basic		3'UTR		Decoy	
	nCC	%	nCC	%	nCC	%
BioProspector	0.43	0.63	0.43	0.63	0.21	0.36
CMF	0.40	0.59	0.45	0.66	0.27	0.45
DECOD	0.42	0.62	0.34	0.51	0.42	0.72
DME	0.48	0.71	0.41	0.60	0.44	0.75
DREME DNA	0.56	0.83	0.59	0.86	0.52	0.88
DREME RNA	0.56	0.83	0.59	0.86	0.53	0.90
DREME RNA*	0.56	0.83	0.59	0.87	0.53	0.90
FIRE	0.39	0.58	0.40	0.59	0.37	0.62
MDscan	0.15	0.23	0.19	0.28	0.09	0.15
MoAn-3M	0.60	0.89	0.62	0.91	0.50	0.85
MoAn	0.63	0.92	0.64	0.94	NA	NA
Recognizability	0.68	1.00	0.68	1.00	0.59	1.00
Discrover - BW (MICO)	0.67	0.98	0.51	0.76	0.55	0.93
Discrover - BW	0.67	0.98	0.51	0.74	0.49	0.83
Discrover - DFREQ	0.61	0.89	0.59	0.87	0.30	0.51
Discrover - DLOGL	0.66	0.97	0.66	0.97	0.58	0.97
Discrover - MCC	0.66	0.97	0.66	0.97	0.56	0.95
Discrover - MICO-DREME	0.65	0.96	0.66	0.97	0.57	0.96
Discrover - MICO	0.66	0.97	0.66	0.97	0.57	0.96
Discrover - MMIE	0.66	0.98	0.67	0.98	0.57	0.96
Plasma	0.63	0.93	0.61	0.90	0.56	0.95

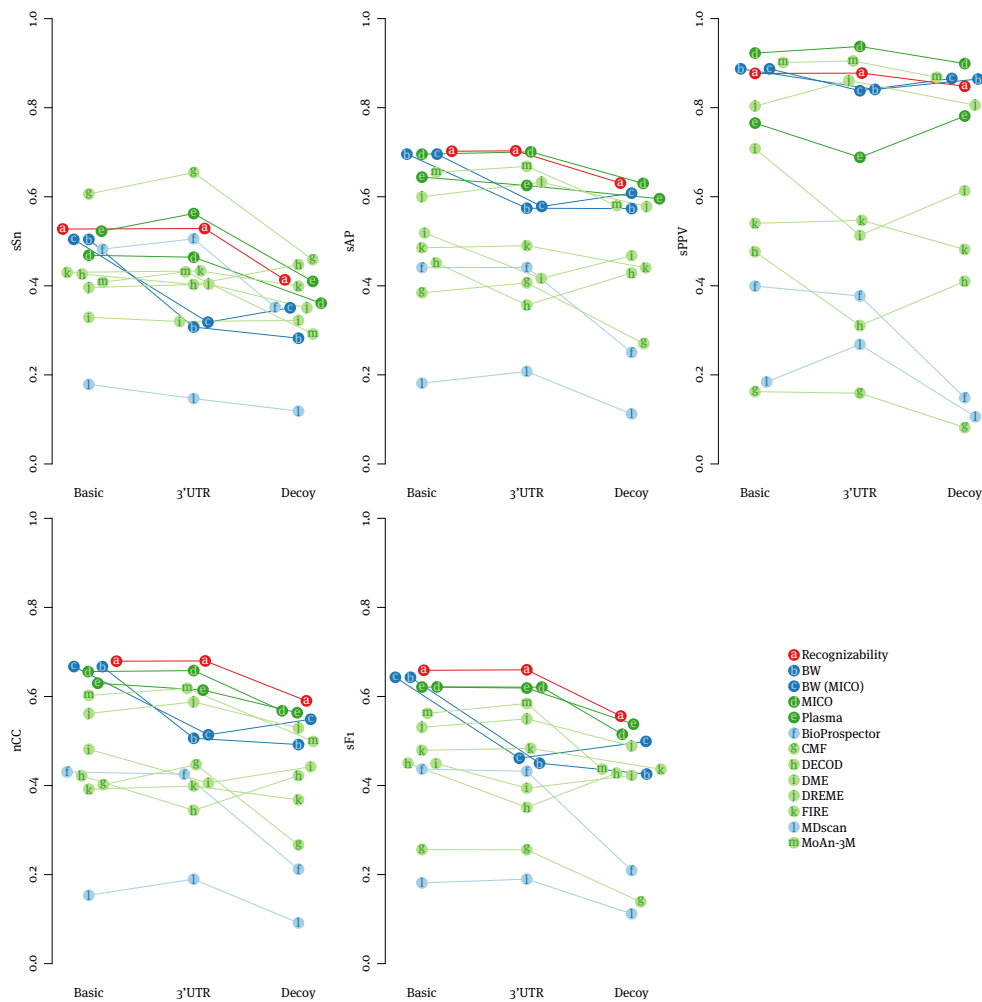


Figure H.1: Summarized motif finding performance of various methods on three synthetic datasets measured by the nucleotide-level Matthews correlation coefficient (nCC), average site performance (sAP), site sensitivity (sSn), and site positive predictive value (sPPV), as well as the  $sF_1$ -score. See (16.1)–(16.5) for definition of the metrics. Recognizability (red) serves as reference. Blue denotes signal-only motif learning methods, while green denotes discriminative motif discovery methods. Dark letters and light background denote published motif finding methods, light letters and dark background denote motif finding with objective functions implemented in Discover. BW: Baum-Welch training of HMMs seeded with the most frequent IUPAC regex motifs of degeneracy maximally 2, BW (MICO): Baum-Welch training of HMMs seeded with IUPAC regex motifs maximizing MICO. Plasma: IUPAC regex motif optimization with MICO as objective function.

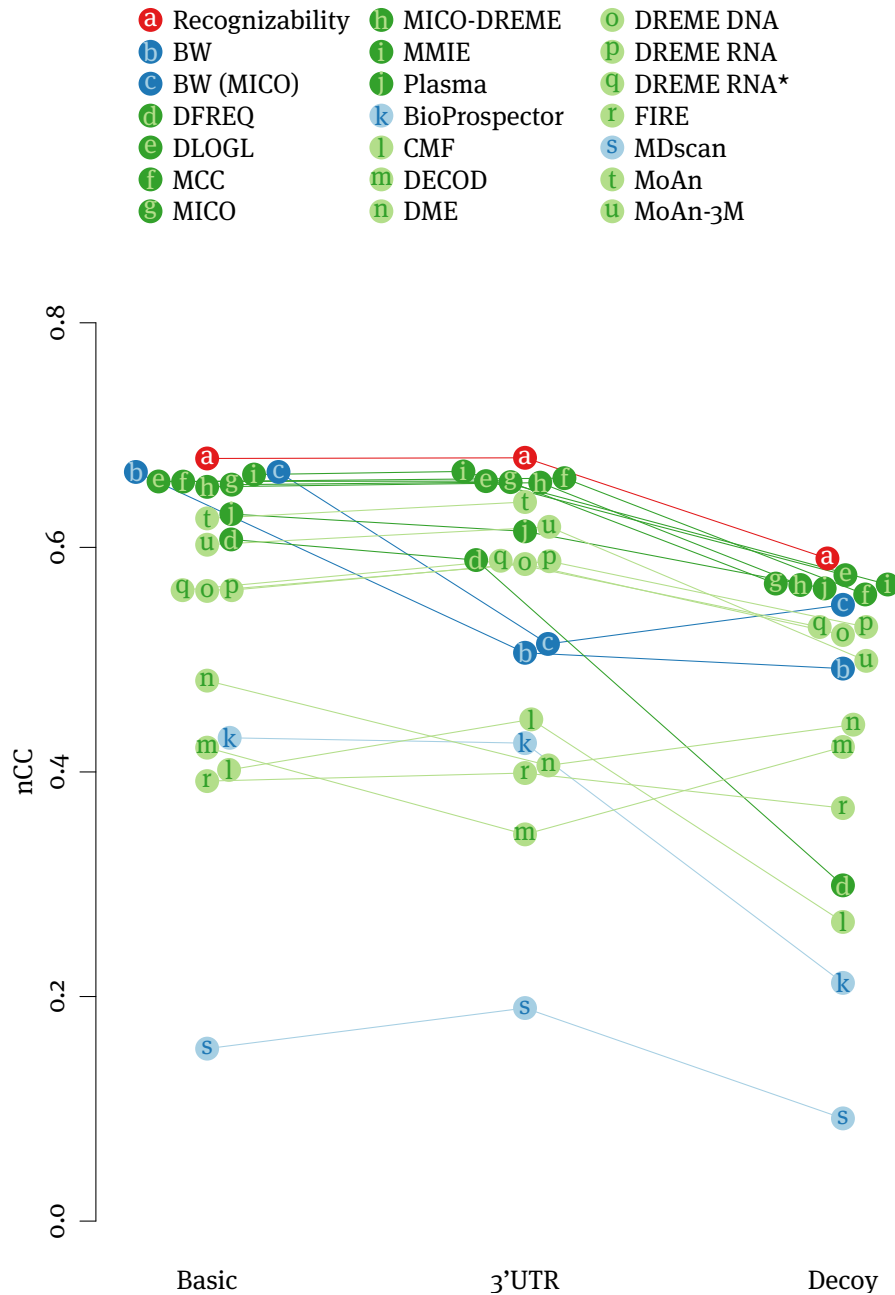


Figure H.2: Summarized motif finding performance of methods on three synthetic datasets measured by the nucleotide-level Matthews correlation coefficient (nCC). See figure 17.1 for description. Numbers are tabulated in table H.2. MoAn-3M: MoAn with  $3 \times 10^6$  iterations. MoAn: MoAn with  $3 \times 10^7$  iterations; note that it was infeasible to evaluate the decoy dataset in this case. MICO-DREME: DREME provides seeds, on which HMMs are seeded and further optimized for MICO by Discover. DREME DNA: DREME in double-stranded motif analysis mode, suitable for DNA-binding protein analysis; providing one seed. DREME RNA: DREME in single-stranded motif analysis mode, suitable for RNA-binding protein analysis; providing one seed. DREME RNA\*: DREME in single-stranded motif analysis mode, discovering motifs as long as the  $E$ -value threshold is met, of which subsequently the highest scoring one is used for evaluation.

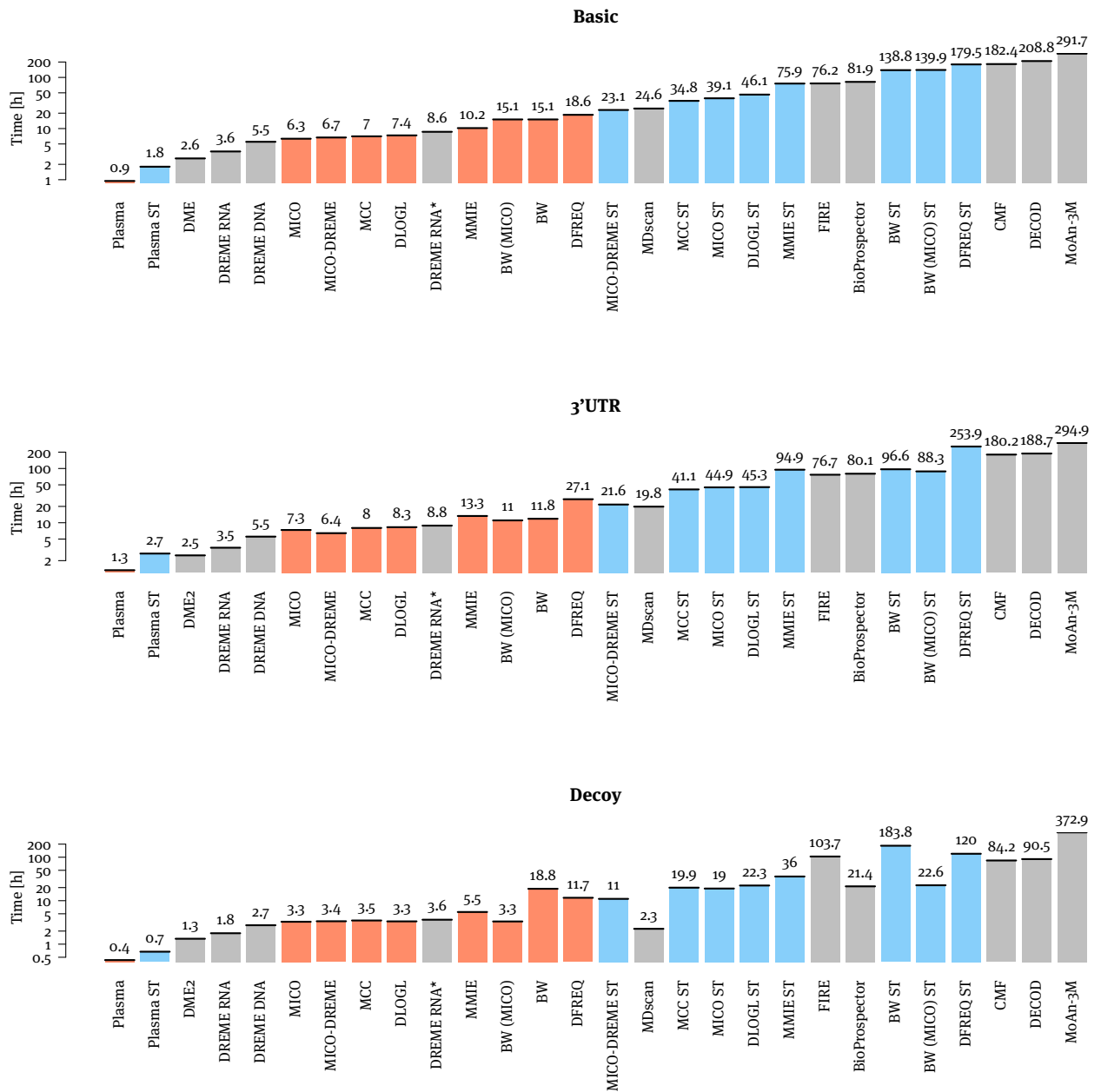


Figure H.3: Runtime of motif discovery methods on the three synthetic datasets using an Intel® Xeon® E5645 CPU running at 2.40GHz with 12 CPU cores. Grey bars denote published methods. As Discover can utilize multiple threads, we include two time measurements: orange bars denote multi-threaded runtime (wall clock time), blue bars denote single-threaded runtime (CPU time). Note that the runtime of MoAn with the default number of iterations, which we performed for the basic and 3'UTR experiments is not included, as it was run on a different compute, and thus the runtimes are not comparable.

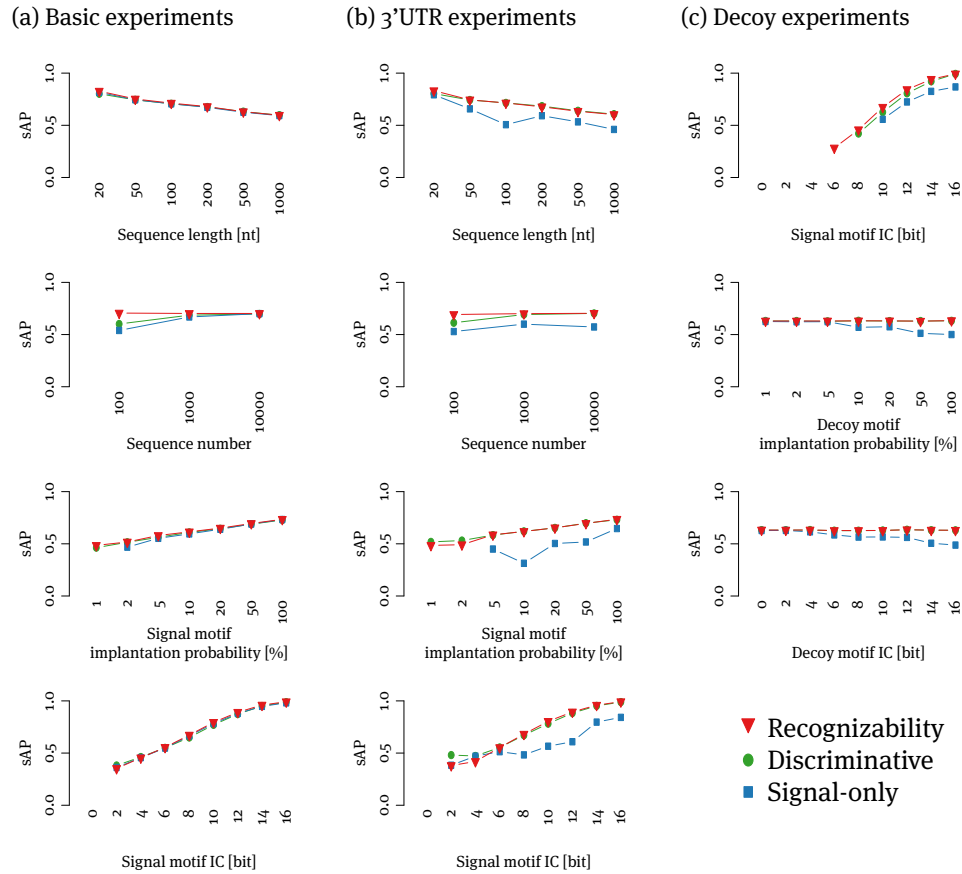


Figure H.4: Motif recognizability and discovery performance measured by average site performance (sAP) on synthetic data in the (a) basic, (b) 3'UTR, and (c) decoy experiments. Note that sAP is not defined when sPPV is not defined (see figure H.6). See legend of figure 17.2 for further explanations.

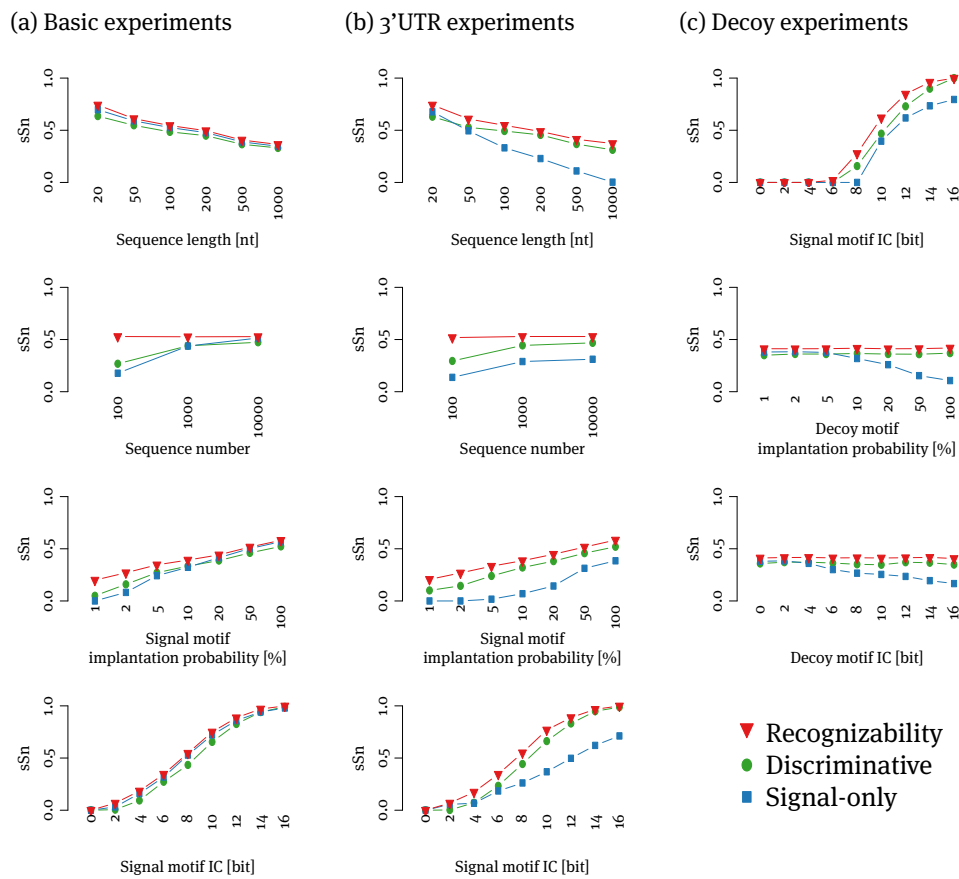


Figure H.5: Motif recognizability and discovery performance measured by site-level sensitivity (sSn) on synthetic data in the (a) basic, (b) 3'UTR, and (c) decoy experiments. See legend of figure 17.2 for further explanations.

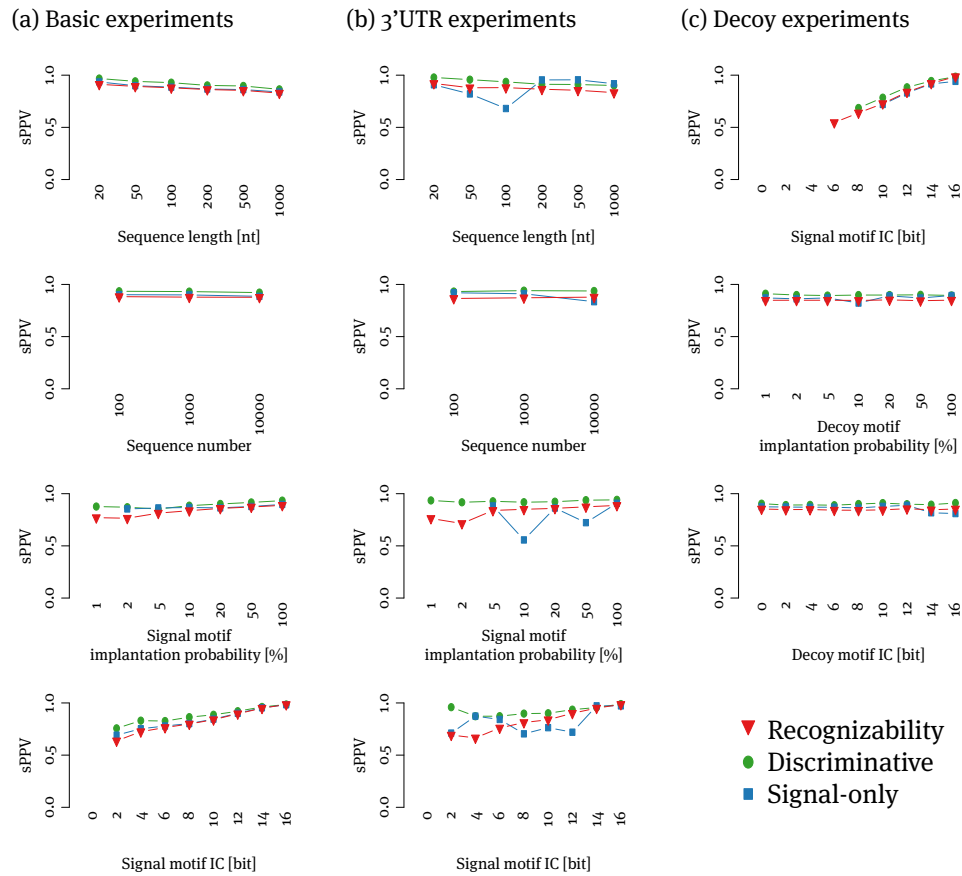


Figure H.6: Motif recognizability and discovery performance measured by site-level positive predictive value (sPPV) on synthetic data in the (a) basic, (b) 3'UTR, and (c) decoy experiments. See legend of figure 17.2 for further explanations. Note that sPPV is not defined when no motif occurrences are predicted, e.g. for low signal motif IC values in (c).



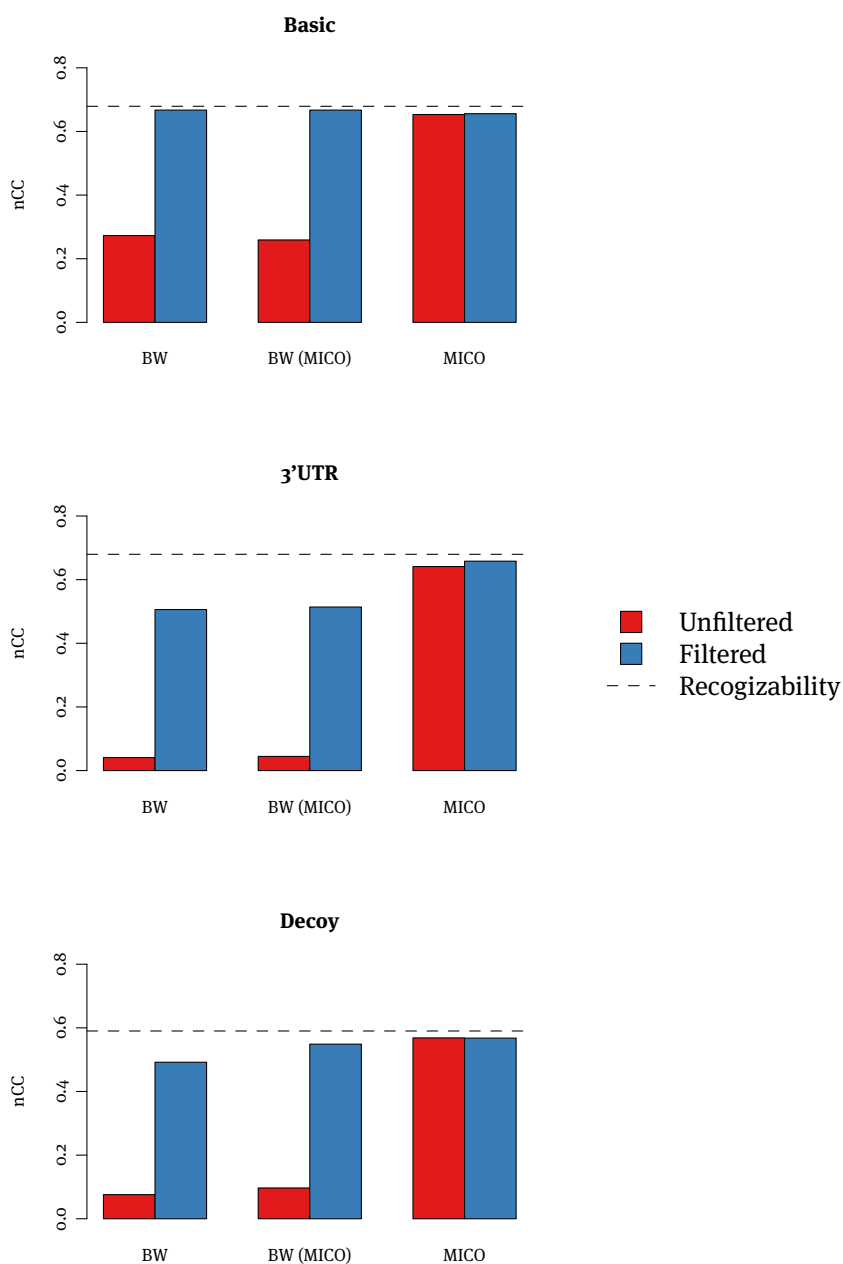


Figure H.7: Effect of significance filtering on motif discovery performance measured by the nucleotide-level Matthews correlation coefficient (nCC). Blue and red bars give motif discovery performance respectively with or without discriminative filtering for significance of association after learning. For reference, the nCC of recognizability is indicated by the dashed line. Significance filtering is done by evaluating MICO on the signal and control example sequences, computing the associated  $p$ -value, correcting for multiple testing, and discarding motifs failing the significance threshold. BW: signal-only learning of HMM parameters with the Baum-Welch algorithm, using as seeds the 8mers of degeneracy at most 2 that are most frequent in the signal data. BW (MICO): signal-only learning of HMM parameters with the Baum-Welch algorithm, using discriminative seeds that maximize MICO for IUPAC regexes on the signal and control data. MICO: Discriminative learning of HMM parameters by MICO, with discriminative seed determined by optimizing MICO.



# Appendix I

## PUF RBP family data

This chapter presents supplementary results for the motif analyses in chapter 18 of the PUF RBP family datasets.

Figure I.1 shows the length distributions of sequences in the different datasets. Frequently, but not for all datasets, the signal data have longer sequences than the controls. In particular, this is visible for the worm, fly, and human RIP-Chip datasets, and is probably explained by an expression bias of genes that have long 3'UTRs. This length inequality is not as pronounced for the yeast data, and is absent from the length-matched PAR-CLIP data.

Figure I.2 gives the relative numbers of sequences in each of these datasets that have at least one occurrence of the IUPAC regex motif UGUAHAUA. This shows that the members of this RBP family fall into two classes. The datasets of Puf3, Pumilio, PUM1, and PUM2 all show enrichment for this motif, while the other datasets are not enriched for this 8mer.

Figures I.3 to I.6 give scatter plots of sequences that have at least one occurrence of a given words for the PUM1 and PUM2 datasets of Galgano et al. (2008), Hafner et al. (2010), and Morris, Mukherjee, and Keene (2008).

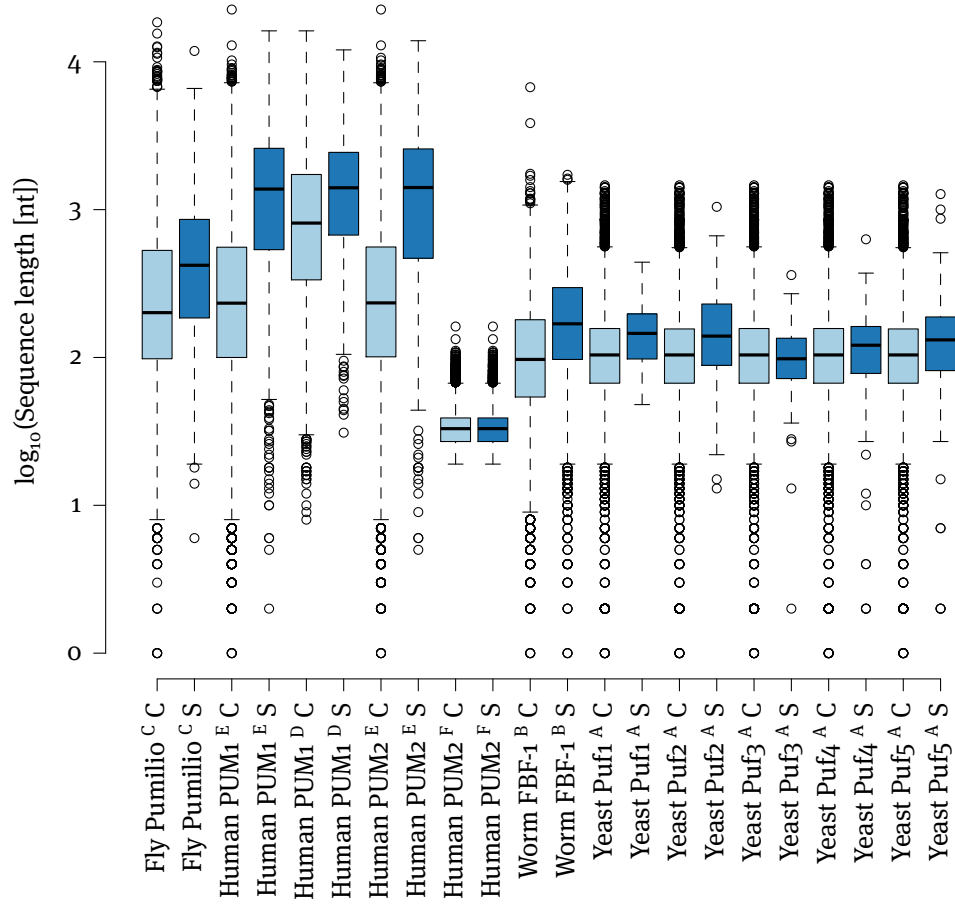


Figure I.1: Boxplot of sequence lengths for the PUF RBP family datasets and their controls. S: Signal sequences (dark blue). C: Control sequences (light blue). Data sources: <sup>A</sup> Gerber, Herschlag, and Brown (2004), <sup>B</sup> Kershner and Kimble (2010), <sup>C</sup> Gerber, Luschnig, et al. (2006), <sup>D</sup> Morris, Mukherjee, and Keene (2008), <sup>E</sup> Galgano et al. (2008), <sup>F</sup> Hafner et al. (2010).

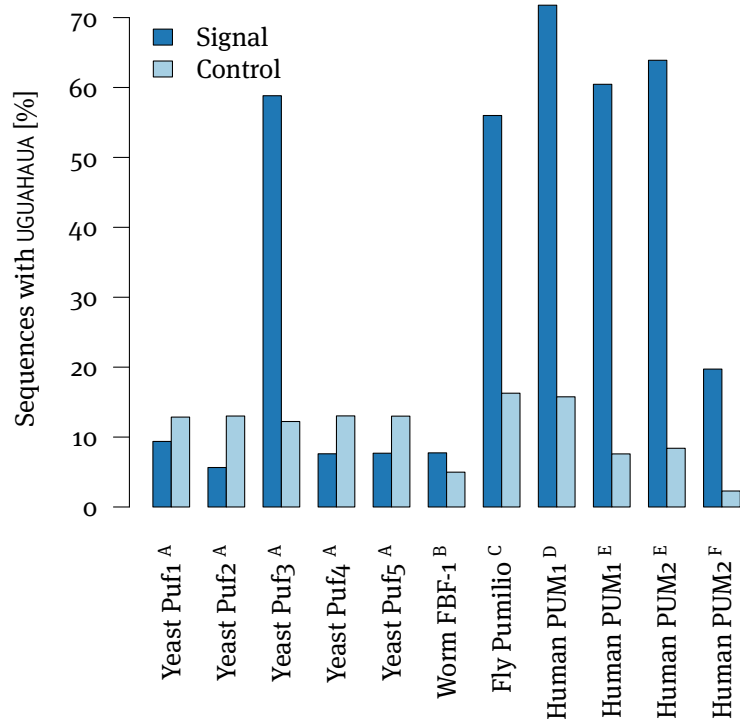


Figure I.2: Number of sequences with at least one occurrence of the IUPAC motif UGUAHAUA in the PUF RBP family data. Data sources: <sup>A</sup> Gerber, Herschlag, and Brown (2004), <sup>B</sup> Kershner and Kimble (2010), <sup>C</sup> Gerber, Luschnig, et al. (2006), <sup>D</sup> Morris, Mukherjee, and Keene (2008), <sup>E</sup> Galgano et al. (2008), <sup>F</sup> Hafner et al. (2010).













## Appendix J

# Alternative splicing regulator

## RBM10

Table J.1: RBM10 motifs of Bechara et al. (2013) in exonic RBM10 PAR-CLIP sequences of Y. Wang et al. (2013). Bechara et al. (2013) performed CLIP-Seq to identify RNA-binding sites of RBM10 and defined 9 groups of 5mers as RBM10 target motifs. This table gives the number of exonic sequences in the two PAR-CLIP datasets of Y. Wang et al. (2013) that have occurrences of these 5mers. S and C: relative frequency of signal and control sequences with at least one motif occurrence. MICO: mutual information of condition and motif occurrence.  $\log-p$ : MICO-based  $\log-p$  value, corrected for motif length. The bars visualize the  $\log-p$  values; black and red bars respectively correspond to enrichment in the signal or control sequences.

Group	Motif	PAR-CLIP dataset 1				PAR-CLIP dataset 2			
		S [%]	C [%]	MICO [bit]	$\log-p$	S [%]	C [%]	MICO [bit]	$\log-p$
1	AACUC	3.5	3.7	0.3	0.0	3.7	4.0	2.0	0.0
1	AAGUC	3.5	3.8	0.5	0.0	3.8	3.9	0.2	0.0
1	UACUC	1.9	1.7	0.3	0.0	2.2	1.6	17.9	-7.3
1	AACUG	5.6	5.9	0.4	0.0	5.4	5.9	3.1	0.0
1	UACUG	3.0	2.5	2.2	0.0	3.1	2.5	9.5	-1.2
1	GACUU	5.3	4.4	5.2	0.0	5.0	4.3	8.9	-0.8
1	GACUC	3.4	3.7	0.5	0.0	3.7	3.9	1.2	0.0
1	GACUG	4.8	6.5	14.8	-5.1	4.7	6.3	39.9	-23.0 ■
1	UUCUC	3.4	2.9	1.5	0.0	4.0	3.2	15.5	-5.6
2	ACUCU	3.3	3.4	0.1	0.0	3.4	3.8	5.2	0.0
2	UCUGA	5.2	6.6	9.6	-1.3	5.1	6.6	37.4	-21.2 ■
2	UCUGG	6.4	4.5	17.6	-7.1	5.6	4.4	24.3	-11.9
2	CCUGA	5.6	5.5	0.0	0.0	5.4	6.4	16.4	-6.3
2	ACUCC	2.8	2.7	0.0	0.0	3.2	3.0	1.3	0.0
2	GCUUG	2.8	4.7	27.5	-14.2 ■	2.5	4.1	72.8	-46.0 ■
2	ACUGA	5.4	8.0	29.0	-15.3 ■	5.3	7.3	56.4	-34.6 ■
2	ACUUC	4.4	3.3	8.7	-0.6	5.1	3.2	76.2	-48.4 ■
2	UCUUG	3.0	4.4	13.8	-4.4	3.1	4.4	37.6	-21.4

Table J.1: continued from previous page.

Group	Motif	PAR-CLIP dataset 1				PAR-CLIP dataset 2			
		S [%]	C [%]	MICO [bit]	log-p	S [%]	C [%]	MICO [bit]	log-p
2	ACUGG	6.3	5.2	6.1	0.0	5.5	5.3	0.6	0.0
2	ACUCA	3.4	4.6	10.1	-1.7	3.9	4.5	8.0	-0.1
2	UCUUC	4.3	3.1	12.0	-3.1	4.6	3.3	39.7	-22.8 ■
2	ACUCG	1.2	1.5	1.9	0.0	1.1	1.3	4.8	0.0
2	ACUGU	2.9	2.9	0.0	0.0	3.3	3.1	1.2	0.0
2	UCUUA	2.4	2.3	0.0	0.0	2.3	2.6	2.1	0.0
2	ACUUG	3.9	4.6	3.8	0.0	3.9	4.7	10.5	-1.9
2	UCUGU	2.8	2.4	2.1	0.0	2.7	3.0	2.7	0.0
2	UGUGA	4.6	6.3	16.6	-6.4	4.9	6.3	34.4	-19.1 ■
3	CUCUG	4.7	4.3	0.8	0.0	4.7	5.0	1.7	0.0
3	UUCUG	5.3	4.1	9.1	-0.9	5.1	4.2	15.3	-5.4
3	GUGUU	2.0	1.8	0.7	0.0	2.3	1.9	6.1	0.0
3	GUCUU	2.2	2.3	0.1	0.0	2.3	2.4	0.8	0.0
3	CUCUC	2.6	2.8	0.5	0.0	3.0	3.4	5.6	0.0
3	CUGUG	4.7	4.6	0.1	0.0	5.1	4.5	6.1	0.0
3	CUUUG	4.7	3.9	3.5	0.0	4.9	4.4	3.9	0.0
3	CUCUU	3.7	3.1	2.5	0.0	3.5	3.6	0.8	0.0
3	CUGUC	1.9	2.4	3.1	0.0	2.5	2.6	0.4	0.0
3	CUGUU	2.8	2.7	0.1	0.0	2.9	3.0	0.1	0.0
3	CUUUC	3.1	3.3	0.3	0.0	3.6	3.2	4.7	0.0
3	GUUUG	2.9	3.3	1.5	0.0	3.1	3.1	0.0	0.0
3	GUCUG	2.6	3.1	2.4	0.0	2.8	3.4	9.0	-0.8
4	CUGAA	10.0	8.7	5.4	0.0	9.3	8.4	7.3	0.0
4	UUGUG	3.4	4.2	4.4	0.0	3.3	4.0	11.6	-2.8
4	UUGAC	2.8	4.5	21.9	-10.2	3.0	4.5	53.2	-32.3 ■
4	CUGAG	4.8	8.4	54.3	-33.1 ■	4.9	7.9	120.0	-79.1 ■
4	UUGUC	1.8	2.4	5.0	0.0	2.0	2.5	11.2	-2.5
4	CUGAC	3.0	4.2	11.5	-2.7	3.3	4.7	45.5	-27.0 ■
4	UUGGA	11.1	7.4	45.7	-27.0 ■	9.2	7.0	56.1	-34.4 ■
4	UUGGG	3.8	4.5	2.9	0.0	3.5	4.2	10.6	-2.0
4	UUGAA	8.3	10.9	21.9	-10.2	8.1	9.4	18.2	-7.6
4	CUGGA	12.5	7.1	89.5	-57.8 ■	11.1	6.9	185.1	-124.4 ■
4	UUGUA	1.7	2.0	1.1	0.0	1.9	2.0	0.7	0.0
4	CUUGA	4.0	7.2	50.4	-30.4	4.0	6.5	100.1	-65.2 ■
4	GUGGA	11.6	6.1	102.4	-66.8 ■	8.8	5.2	163.2	-109.2 ■
5	GAACU	6.8	6.3	1.0	0.0	6.0	5.6	3.1	0.0
5	GAAGG	9.7	9.4	0.2	0.0	8.8	8.1	6.0	0.0
5	GUACU	1.5	1.3	0.8	0.0	1.6	1.3	4.8	0.0
5	GAAGA	32.6	15.5	440.0	-301.6 ■	27.5	14.4	868.2	-598.7 ■
5	CAACU	3.6	4.0	1.7	0.0	4.0	3.9	0.0	0.0
5	GAGCU	5.7	5.1	1.6	0.0	5.0	5.3	1.2	0.0
5	GGACU	5.2	4.7	1.7	0.0	4.8	4.4	4.0	0.0

TableJ.1: continued from previous page.

Group	Motif	PAR-CLIP dataset 1				PAR-CLIP dataset 2			
		S [%]	C [%]	MICO [bit]	log-p	S [%]	C [%]	MICO [bit]	log-p
5	GAAGU	8.3	5.9	23.3	-11.2	8.0	5.6	73.7	-46.7 ■
6	UGGAA	14.6	9.3	72.6	-45.9 ■	13.3	8.8	167.7	-112.3 ■
6	UGUUG	3.5	3.8	0.5	0.0	3.5	3.7	1.1	0.0
6	UGUGC	2.4	3.2	6.0	0.0	2.6	2.9	2.4	0.0
6	UGUAG	2.3	2.0	1.4	0.0	2.0	2.3	2.4	0.0
6	UGGAG	13.5	8.5	68.6	-43.1 ■	11.8	8.0	133.2	-88.3 ■
6	UGUAC	2.0	1.3	8.1	-0.2	1.9	1.3	18.4	-7.7
6	UGAAG	20.2	11.5	153.9	-102.7 ■	18.0	10.4	395.8	-270.8 ■
6	UGUCC	2.2	2.2	0.0	0.0	2.6	2.3	3.2	0.0
6	UGAAC	5.1	6.3	6.8	0.0	5.0	5.6	5.9	0.0
6	UGUUC	2.4	2.0	1.6	0.0	2.7	2.4	2.7	0.0
6	UGGAC	9.7	5.2	77.1	-49.1 ■	7.2	4.6	97.9	-63.6 ■
6	UGUGG	6.7	4.8	17.0	-6.7	6.1	4.5	45.8	-27.1 ■
6	AGAAC	9.2	7.5	10.3	-1.8	8.1	7.3	9.2	-1.0
7	CUUUU	3.7	3.7	0.0	0.0	4.0	3.8	0.6	0.0
7	GAUCU	4.3	4.6	0.6	0.0	3.8	4.7	15.5	-5.6
7	UGUCU	2.4	2.9	3.2	0.0	2.8	2.9	0.7	0.0
7	CCUUU	3.1	2.9	0.4	0.0	3.6	3.2	3.7	0.0
7	GGUCU	1.7	2.3	4.8	0.0	1.8	2.4	15.0	-5.3
7	CUUCU	4.0	3.3	4.5	0.0	4.2	3.7	7.2	0.0
7	UCUCU	3.1	3.1	0.0	0.0	3.3	3.6	1.8	0.0
7	CUUGU	1.9	2.6	5.0	0.0	2.1	2.8	15.5	-5.6
7	CUUUAU	2.7	1.7	11.3	-2.5	2.5	2.1	5.1	0.0
7	UCUUU	3.2	3.5	0.4	0.0	3.7	3.8	0.4	0.0
7	GCUCU	3.4	2.9	2.6	0.0	3.3	3.3	0.1	0.0
7	GCUUU	3.2	2.8	1.3	0.0	3.5	3.0	5.8	0.0
7	CCUCU	3.2	3.0	0.3	0.0	3.4	3.6	0.8	0.0
7	AGUCU	2.8	3.3	3.0	0.0	2.7	3.5	17.9	-7.3
7	UGACU	3.8	5.3	13.6	-4.2	4.2	5.1	15.2	-5.4
8	UUCCU	4.0	3.0	9.4	-1.1	4.4	3.4	24.6	-12.1
8	UGCUU	2.6	3.6	9.2	-1.0	3.1	3.5	5.4	0.0
8	UCCUU	3.3	2.6	4.0	0.0	3.5	3.2	2.4	0.0
8	UCCCU	2.3	2.6	0.8	0.0	2.9	3.0	0.4	0.0
8	UUCUU	3.3	3.5	0.6	0.0	4.1	3.8	1.5	0.0

Table J.2: RBM10 motifs of Bechara et al. (2013) in intronic RBM10 PAR-CLIP sequences of Y. Wang et al. (2013). Bechara et al. (2013) performed CLIP-Seq to identify RNA-binding sites of RBM10 and defined 9 groups of 5mers as RBM10 target motifs. This table gives the number of intronic sequences in the two PAR-CLIP datasets of Y. Wang et al. (2013) that have occurrences of these 5mers. S and C: relative frequency of signal and control sequences with at least one motif occurrence. MICO: mutual information of condition and motif occurrence.  $\log-p$ : MICO-based  $\log-p$  value, corrected for motif length. The bars visualize the  $\log-p$  values; black and red bars respectively correspond to enrichment in the signal or control sequences.

Group	Motif	PAR-CLIP dataset 1				PAR-CLIP dataset 2			
		S [%]	C [%]	MICO [bit]	$\log-p$	S [%]	C [%]	MICO [bit]	$\log-p$
1	AACUC	3.8	3.7	0.2	0.0	3.7	3.7	0.1	0.0
1	AAGUC	1.8	1.8	0.0	0.0	2.3	2.6	3.7	0.0
1	UACUC	3.3	3.4	0.1	0.0	3.6	2.8	17.1	-6.8
1	AACUG	3.0	3.7	3.0	0.0	3.3	4.3	20.7	-9.3
1	UACUG	2.5	3.0	2.3	0.0	2.9	3.0	0.0	0.0
1	GACUU	3.4	3.7	0.3	0.0	3.7	3.9	1.0	0.0
1	GACUC	3.6	4.5	3.8	0.0	3.4	4.1	12.8	-3.7
1	GACUG	4.6	6.3	11.2	-2.5	4.4	5.1	8.2	-0.3
1	UUCUC	10.2	7.7	17.1	-6.8	10.7	7.7	85.5	-55.0 ■
2	ACUCU	6.3	6.9	1.3	0.0	5.9	6.2	1.8	0.0
2	UCUGA	5.2	6.3	4.8	0.0	5.3	6.1	9.7	-1.3
2	UCUGG	10.8	10.8	0.0	0.0	8.7	9.0	0.4	0.0
2	CCUGA	6.4	7.4	3.1	0.0	6.2	6.2	0.0	0.0
2	ACUCC	5.3	5.7	0.9	0.0	4.6	5.1	4.6	0.0
2	GCUUG	4.8	5.8	4.2	0.0	3.7	5.2	43.3	-25.4
2	ACUGA	3.5	3.9	1.3	0.0	3.7	4.5	13.0	-3.8
2	ACUUC	4.6	4.5	0.0	0.0	4.9	4.7	1.0	0.0
2	UCUUG	5.2	7.2	15.1	-5.3	5.1	7.3	63.6	-39.6
2	ACUGG	6.8	7.5	1.7	0.0	5.5	6.4	11.0	-2.3
2	ACUCA	3.9	4.9	4.2	0.0	4.3	4.7	3.0	0.0
2	UCUUC	8.5	7.8	1.3	0.0	8.9	7.4	20.9	-9.5
2	ACUCG	1.0	1.0	0.1	0.0	0.9	1.0	0.1	0.0
2	ACUGU	5.0	4.4	1.6	0.0	5.1	4.5	6.6	0.0
2	UCUUA	3.6	3.9	0.8	0.0	4.4	4.7	1.8	0.0
2	ACUUG	3.5	4.6	7.0	0.0	3.7	4.8	21.5	-9.9
2	UCUGU	7.3	6.3	3.1	0.0	7.0	6.2	8.0	-0.1
2	UGUGA	4.4	4.4	0.0	0.0	5.1	4.4	9.0	-0.8
3	CUCUG	12.3	12.1	0.2	0.0	10.7	10.4	0.6	0.0
3	UUCUG	8.5	8.1	0.5	0.0	8.0	7.5	3.8	0.0
3	GUGUU	3.2	3.1	0.1	0.0	3.7	3.1	7.9	-0.0
3	GUCUU	3.6	4.6	6.0	0.0	4.0	4.8	11.8	-2.9
3	CUCUC	10.0	10.1	0.0	0.0	8.8	8.7	0.1	0.0
3	CUGUG	12.6	9.8	16.8	-6.5	11.1	8.2	74.0	-46.9 ■
3	CUUUG	5.4	5.9	1.2	0.0	5.5	6.5	14.7	-5.0

TableJ.2: continued from previous page.

Group	Motif	PAR-CLIP dataset 1				PAR-CLIP dataset 2			
		S [%]	C [%]	MICO [bit]	log-p	S [%]	C [%]	MICO [bit]	log-p
3	CUCUU	10.3	10.1	0.1	0.0	9.7	9.3	1.9	0.0
3	CUGUC	7.0	6.7	0.3	0.0	6.5	6.3	0.3	0.0
3	CUGUU	5.5	5.5	0.0	0.0	5.7	5.6	0.2	0.0
3	CUUUC	9.2	7.3	9.4	-1.2	9.6	7.2	58.6	-36.1 ■
3	GUUUG	2.8	2.9	0.1	0.0	3.4	3.6	0.7	0.0
3	GUCUG	6.1	6.7	1.3	0.0	5.1	6.0	11.6	-2.8
4	CUGAA	4.5	4.1	0.7	0.0	4.8	4.6	0.6	0.0
4	UUGUG	4.0	4.6	1.4	0.0	4.1	5.0	14.2	-4.7
4	UUGAC	1.7	3.4	27.5	-14.2 ■	2.5	3.8	43.9	-25.8 ■
4	CUGAG	8.1	8.5	0.4	0.0	7.4	7.5	0.2	0.0
4	UUGUC	3.3	4.1	3.8	0.0	3.3	4.5	29.8	-15.9
4	CUGAC	4.5	6.1	10.3	-1.8	4.5	5.1	6.1	0.0
4	UUGGA	4.9	4.7	0.2	0.0	4.6	5.0	2.3	0.0
4	UUGGG	7.7	7.4	0.2	0.0	6.3	6.5	0.7	0.0
4	UUGAA	2.4	3.2	5.0	0.0	3.6	4.6	19.0	-8.1
4	CUGGA	10.1	7.9	12.8	-3.7	8.3	6.8	25.1	-12.5
4	UUGUA	1.4	1.8	1.9	0.0	2.0	2.7	13.7	-4.3
4	CUUGA	3.1	4.9	18.3	-7.6	3.5	5.5	70.5	-44.5 ■
4	GUGGA	6.0	4.6	8.4	-0.4	5.4	4.4	16.9	-6.6
5	GAACU	3.1	2.9	0.2	0.0	2.8	3.2	5.7	0.0
5	GAAGG	4.7	3.8	3.6	0.0	4.7	4.2	4.2	0.0
5	GUACU	1.8	2.0	0.9	0.0	1.7	2.0	3.0	0.0
5	GAAGA	4.9	2.9	22.5	-10.7	5.7	3.7	68.9	-43.4 ■
5	CAACU	2.4	4.5	27.8	-14.4 ■	2.9	4.3	42.9	-25.1 ■
5	GAGCU	5.0	5.4	0.8	0.0	4.8	5.0	0.6	0.0
5	GGACU	5.0	5.1	0.1	0.0	4.2	4.9	7.6	0.0
5	GAAGU	2.8	2.1	4.2	0.0	3.1	2.6	5.8	0.0
6	UGGAA	5.6	3.6	20.5	-9.2	5.9	4.3	38.8	-22.2 ■
6	UGUUG	4.0	4.3	0.7	0.0	3.8	4.6	12.2	-3.2
6	UGUGC	6.3	5.9	0.8	0.0	5.6	5.5	0.2	0.0
6	UGUAG	2.1	1.9	0.4	0.0	2.1	2.3	0.7	0.0
6	UGGAG	8.9	7.5	5.5	0.0	8.2	6.9	17.7	-7.2
6	UGUAC	1.6	2.0	1.6	0.0	2.0	2.2	0.9	0.0
6	UGAAG	4.9	3.6	8.5	-0.5	5.2	4.4	12.2	-3.2
6	UGUCC	6.1	5.5	1.5	0.0	5.6	5.2	3.2	0.0
6	UGAAC	2.5	3.1	2.7	0.0	2.6	3.1	6.1	0.0
6	UGUUC	4.2	3.9	0.5	0.0	4.5	4.6	0.3	0.0
6	UGGAC	6.3	5.1	5.3	0.0	4.7	4.8	0.1	0.0
6	UGUGG	10.6	8.8	8.4	-0.5	8.7	7.2	24.4	-12.0
6	AGAAC	2.7	1.9	5.3	0.0	2.7	2.6	0.1	0.0
7	CUUUU	7.6	7.4	0.1	0.0	9.2	8.4	5.8	0.0
7	GAUCU	2.2	3.4	9.8	-1.5	2.5	3.6	33.7	-18.6 ■

Table J.2: continued from previous page.

Group	Motif	PAR-CLIP dataset 1				PAR-CLIP dataset 2			
		S [%]	C [%]	MICO [bit]	log-p	S [%]	C [%]	MICO [bit]	log-p
7	UGUCU	5.9	6.1	0.2	0.0	5.8	6.2	1.9	0.0
7	CCUUU	8.1	7.3	2.1	0.0	8.5	7.0	22.4	-10.6
7	GGUCU	5.3	6.3	3.9	0.0	4.6	5.6	15.6	-5.7
7	CUUCU	9.9	10.1	0.0	0.0	9.7	8.8	7.7	0.0
7	UCUCU	10.3	9.7	1.0	0.0	10.4	9.1	17.2	-6.8
7	CUUGU	3.8	5.4	12.3	-3.3	3.9	5.6	51.7	-31.3 ■
7	CUUUAU	2.8	3.1	0.7	0.0	3.3	3.7	3.7	0.0
7	UCUUU	8.4	7.8	1.1	0.0	9.6	9.0	3.7	0.0
7	GCUCU	7.9	8.1	0.0	0.0	6.9	7.1	0.7	0.0
7	GCUUU	5.0	4.4	1.9	0.0	5.1	4.6	4.3	0.0
7	CCUCU	12.2	11.5	0.8	0.0	10.0	10.0	0.0	0.0
7	AGUCU	2.8	3.9	7.0	0.0	3.1	4.5	43.6	-25.6 ■
7	UGACU	4.3	5.6	8.3	-0.4	4.6	5.2	5.9	0.0
8	UUCUU	12.4	7.7	51.1	-30.9 ■	11.6	7.8	130.9	-86.6 ■
8	UGCUU	5.1	5.8	2.0	0.0	5.6	5.8	1.3	0.0
8	UCCUU	10.1	8.4	7.1	0.0	9.6	7.9	27.8	-14.4
8	UCCCU	9.5	9.0	0.7	0.0	8.5	7.9	4.1	0.0
8	UUCUU	8.2	7.9	0.3	0.0	9.5	8.5	9.8	-1.5

Table J.3: RBM10 motifs of Inoue et al. (2014) in exonic RBM10 PAR-CLIP sequences of Y. Wang et al. (2013). Inoue et al. (2014) defined motifs based on the sequences of 5' splice sites of two exons affected by RBM10 knock-down. This table gives the number of exonic sequences in the two PAR-CLIP datasets of Y. Wang et al. (2013) that have occurrences of these motifs. The vertical bar in the motif indicates exon-intron boundary of the two example sequences (note: all occurrences in PAR-CLIP sequences are counted, whether across exon-intron boundaries or not). S and C: relative frequency of signal and control sequences with at least one motif occurrence. MICO: mutual information of condition and motif occurrence. log-p: MICO-based log-p value, corrected for motif length. The bars visualize the log-p values; black and red bars respectively correspond to enrichment in the signal or control sequences.

Motif	PAR-CLIP dataset 1				PAR-CLIP dataset 2			
	S [%]	C [%]	MICO [bit]	log-p	S [%]	C [%]	MICO [bit]	log-p
AG GUAA	0.6	0.8	2.3	0.0	0.5	0.7	4.0	0.0
GG GUAAG	0.1	0.1	0.6	0.0	0.1	0.1	3.5	0.0



Table J.4: RBM10 motifs of Inoue et al. (2014) in intronic RBM10 PAR-CLIP sequences of Y. Wang et al. (2013). Inoue et al. (2014) defined motifs based on the sequences of 5' splice sites of two exons affected by RBM10 knock-down. This table gives the number of intronic sequences in the two PAR-CLIP datasets of Y. Wang et al. (2013) that have occurrences of these motifs. The vertical bar in the motif indicates exon-intron boundary of the two example sequences (note: all occurrences in PAR-CLIP sequences are counted, whether across exon-intron boundaries or not). S and C: relative frequency of signal and control sequences with at least one motif occurrence. MICO: mutual information of condition and motif occurrence.  $\log-p$ : MICO-based  $\log-p$  value, corrected for motif length. The bars visualize the  $\log-p$  values; black and red bars respectively correspond to enrichment in the signal or control sequences.

Motif	PAR-CLIP dataset 1				PAR-CLIP dataset 2			
	S [%]	C [%]	MICO [bit]	$\log-p$	S [%]	C [%]	MICO [bit]	$\log-p$
AG GUAA	0.5	0.3	4.4	0.0	0.6	0.4	2.3	0.0
GG GUAAG	0.2	0.1	0.0	0.0	0.1	0.1	0.1	0.0



# Appendix K

## Mouse ESC ChIP-Seq data

This chapter presents supplementary results for the mouse ESC ChIP-Seq motif analysis of chapter 20. Table K.1 gives the positional distributions of occurrences for all motifs presented in table 20.1 in windows of 501 nt around the midpoints of ChIP-Seq regions.

Table K.1: Positional distribution of Viterbi-decoded occurrences of predicted motifs of table 20.1 in 501 nt windows in distance up to 250 nt from the peak of the ChIP-Seq regions. Black: occurrences in the signal sequences, red: occurrences in the control sequences. Note that windows of length 101 nt were used for motif discovery and are the basis of the statistics listed in table 20.1. Data sources: <sup>A</sup> X. Chen et al. (2008), <sup>B</sup> Marson et al. (2008).

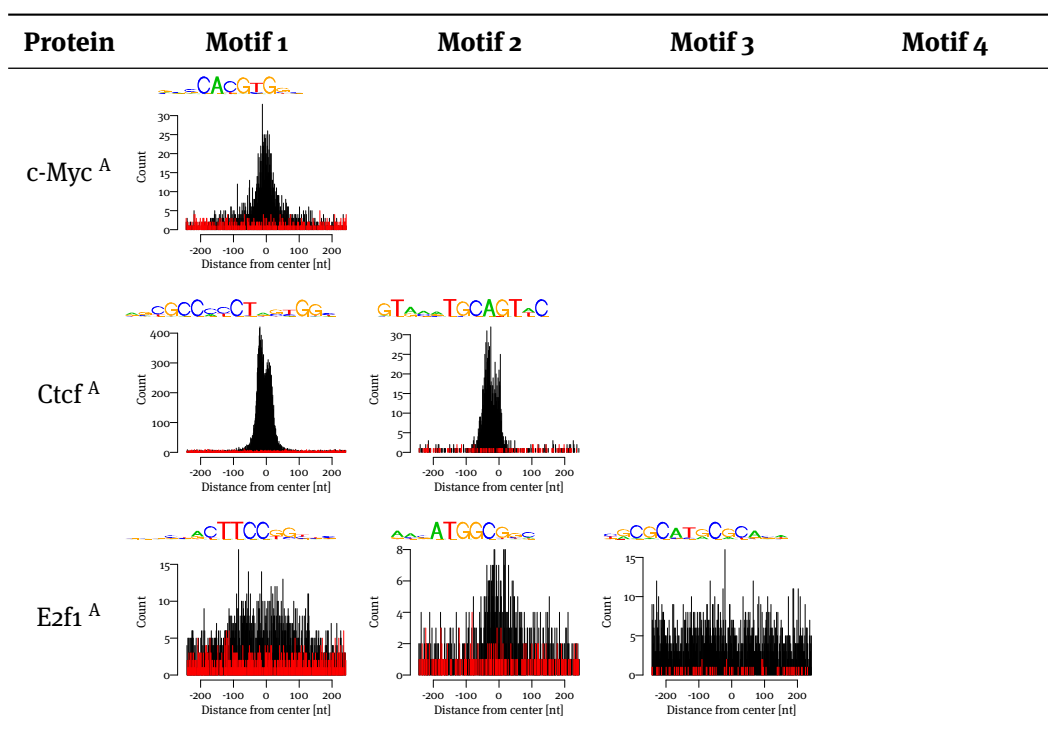


Table K.1: continued from previous page.

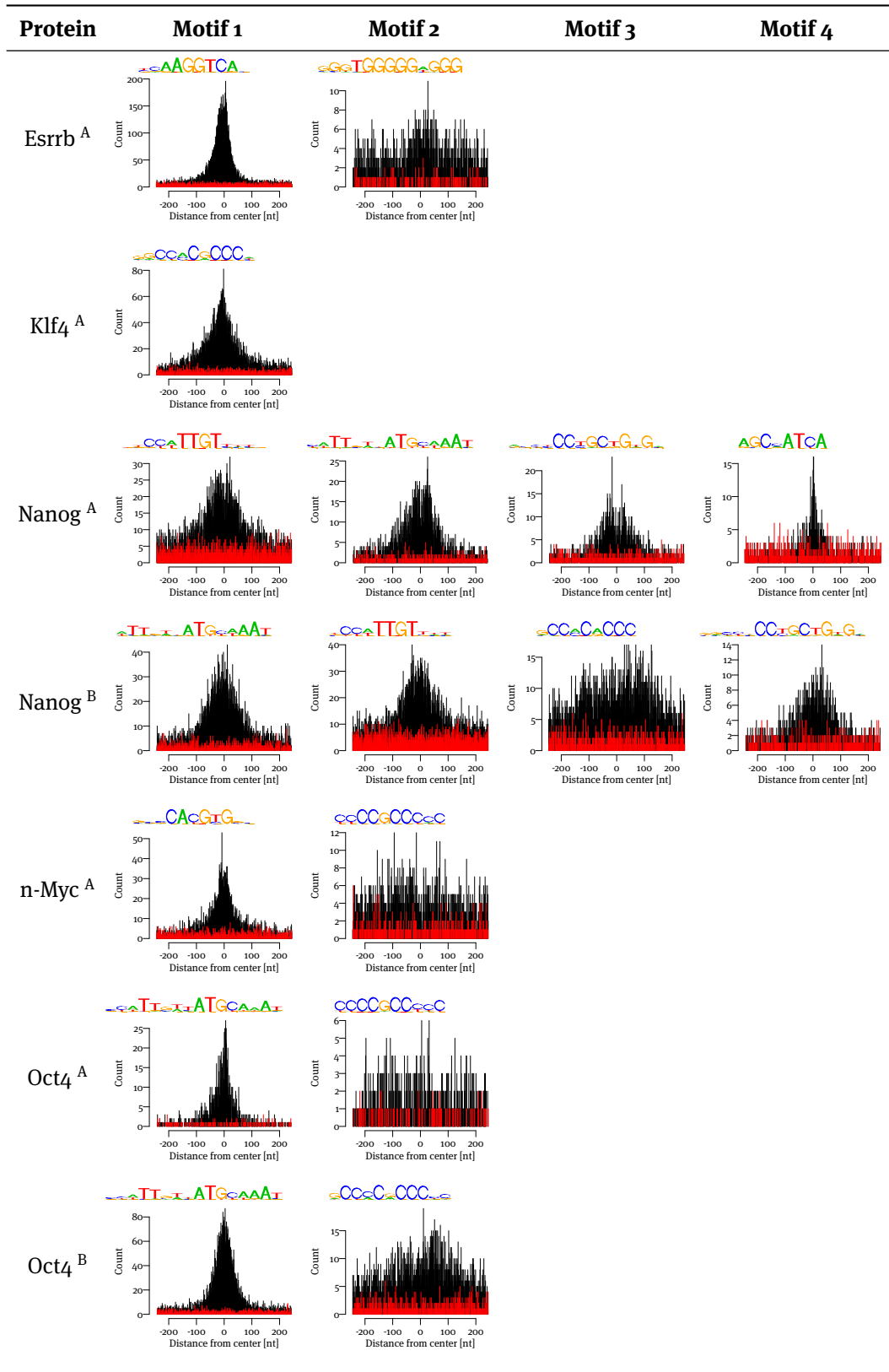
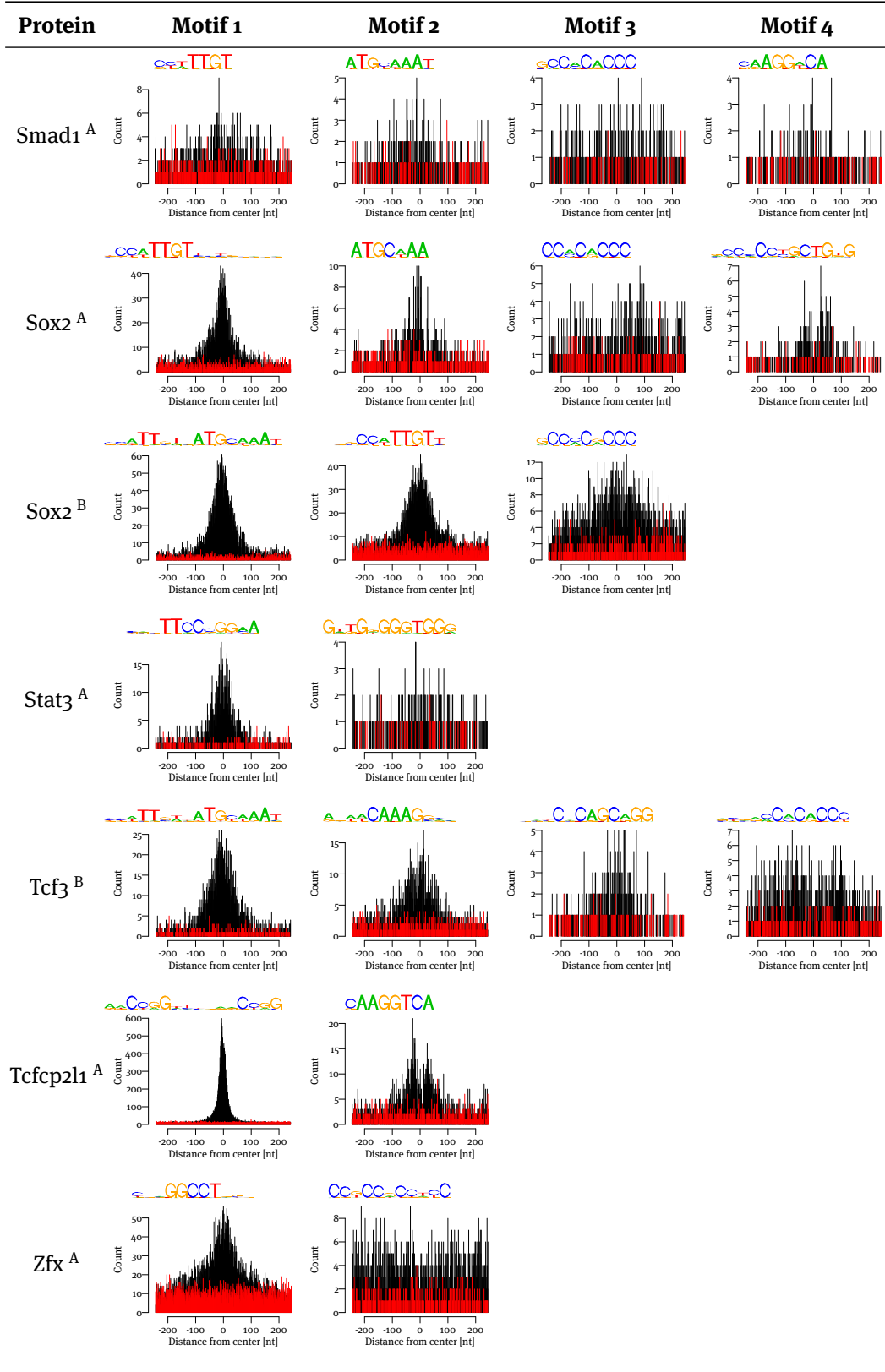


Table K.1: continued from previous page.





# Appendix L

## Research anecdotes

This chapter collects a number of anecdotes pertaining to topics presented in this dissertation. Appendix [L.1](#) addresses some bugs and errors in MD tools and publications, and appendix [L.2](#) recounts a misadventure during our first ChIP-Seq run for the investigation that lead to the publication von Eyss et al. ([2012](#)).

### L.1 Motif discovery anecdotes

**Mistake in equation of objective function of Dispom** A mistake in the manuscript of Dispom (Keilwagen et al., [2011](#)) was discovered regarding the equation describing the objective function. A note was sent to the authors, who were quick to acknowledge the fact, blaming the error on the editorial process. Subsequently, a formal correction was issued.

**Trivial bug in CMF** A trivial bug was found in the source code of CMF (Mason, Plath, and Zhou, [2010](#)) that prevented the method from reporting some occurrences of the identified motif. Specifically, CMF would not report motif occurrences that start on the first position or that end on the last position. The authors acknowledged the bug and promptly provided a patch to fix the problem. The MD performance analysis was based on the patched version.

**ALSE has bit-rotted** The method ALSE (Leung and Chin, [2006](#)) was found to not compile with modern versions of the GCC. Although the C++ source code could be patched such that the code compiled apparently successfully, the resulting program failed with segmentation violations (memory access errors) when run to search for motifs of length greater seven nucleotides. Thus the method was excluded from the analysis.

**Hard coded limits for sequence numbers in DIPS** The program DIPS (Sinha, [2006](#)) was modified to accommodate more sequences. In particular, the source code of DIPS hard-codes limits for the number of sequences with two constants. Both were increased to allow for tens of thousands of sequences to be processed. This did not affect the runtime issues, that ultimately lead us to exclude DIPS from the performance evaluation.

```
(a) #define RAND_INT(max) (int) (((float) max) * rand() / (RAND_MAX + 1.0))  
(b) #define RAND_INT(max) (int) (((double) max) * rand() / (RAND_MAX + 11))
```

Figure L.1: Routines for random number generation in MoAn. (a) Faulty random number generation routine in MoAn. (b) Proposed fix.

**Faulty random number generation in MoAn** The problem I found with MoAn was due to faulty integer random number generation. The routine used by MoAn is shown in figure L.1a. It is supposed to generate integers greater or equal to 0 and less than `max`. First, an integer  $x$  between 0 and `RAND_MAX` =  $2^{31} - 1 = 2147483647$  is generated using the system routine `rand()`. `max` is cast to `float` and multiplied with  $x$  which is implicitly cast to `float` for this. The product is subsequently divided by `RAND_INT+1`, a value strictly greater than  $x$ , thus supposedly yielding a value that is strictly less than `max`. Finally, the resulting floating point number is cast to `int` by truncating the fractional part.

However, this routine generates values that are equal to `max` with a probability of  $\frac{64}{2147483647}$  or about  $2.98 \times 10^{-8}$ . The reason for this is that variables of type `float` lack resolution to represent values that are close to, but less than 1. In particular, all division results that fall above the largest representable number less than 1 are rounded to one. The consequence is that instead of yielding numbers between 0 and  $n - 1$ , the routine sometimes returns  $n$ . As such numbers are used by MoAn to index arrays, segmentation violations may occur, and MoAn aborts. Although such events are individually relatively rare, MoAn generates a lot (millions) of random numbers per run, and thus the problems may occur with a frequency in the percent range. I fixed this problem by modifying the routine as shown in figure L.1b.

**FIRE's sequence parsing routine ignores lower-case letters** This peculiar behavior went unnoticed at first, and FIRE's MD performance in the supervised synthetic data experiments was equally good as that of DREME and MoAn, i.e. it apparently scored among the best previously published methods. However, when applying it together with DREME and Discover to the Oct4 data of X. Chen et al. (2008) (see section 20.5), it aborted with memory errors. Investigating the reasons for the errors revealed that FIRE ignores all lower-case parts of sequences. As the provided ChIP-Seq sequences consisted entirely of lower-case letters, FIRE tried to discover from sets of sequences that were all empty, leading to the aforementioned memory errors. Providing upper-cased versions of ChIP-Seq sequences fixed this problem.

Subsequently, I revisited the synthetic data experiments, and found that the problem was also manifesting itself there. The synthetic sequences used in the supervised experiments consist of lower-case sequence context, and implanted motifs in upper-case. In other words, FIRE ignored everything distracting and thereby cheated, as it only was looking at the pure signal that was to be discovered.

After upper-casing also the synthetic sequences, the MD performance of FIRE dropped significantly, and FIRE was not among the best-performing published methods anymore (see section 17.2).



## L.2 A fishy smell in CHIP-Seq data

Initially, we performed CHIP-Seq in human cell lines for human homologs E2F3 and HELLS of the later published CHIP-Seq'd mouse proteins (E2f3, Hells). After the data was sequenced, I mapped it to the human genome. To our surprise, we found that only few reads actually mapped to the genome: among the samples we had, only 3–8% of reads were seemingly deriving from the human genome. This was the first time, someone had done CHIP-Seq in the wet-lab or computationally analyzed CHIP-Seq data on the MDC campus. Previously, I had only worked with deep-sequencing data of small RNAs, and this experience with mapping small RNA to the genome told me that the observed low fraction of reads mappable to the genome indicated problems. In particular, there were some reads that were quite abundant in our data but that seemingly did not derive from the genome, and that also did not look like obvious sequencing artifacts, for they were neither known adapter sequences nor of low sequence complexity. After checking that I had not made other mistakes in mapping, I proceeded to use NCBI's BLAST web service<sup>1</sup> to search for these highly expressed reads in a non-redundant database of nucleotide sequences. To my amusement, I found various fish species seemingly had sequences that were similar to many of the top-expressed reads in our data. While I could not make much sense of this, my wet-lab colleague when I told him about, after initial puzzlement, suddenly turned rather pale. He realized a mistake that he had committed in his protocol.

It had been the first time that he had performed CHIP-Seq, and he had followed a protocol that he had earlier used for CHIP-Chip. The crucial mistake that he had committed was to add unrelated DNA to the samples, prior to adding cellular extract, in order to cover the beads. This is done in microarray-based CHIP analysis to reduce the amount of unspecifically interacting DNA that sticks to the beads. As microarrays only interrogate sequences that are spotted onto them, the unrelated DNA will compete with unspecific DNA of the organism under question, and is thus beneficial. Unfortunately, when the samples are subsequently sequenced, as in CHIP-Seq, this is quite problematic, as also the unrelated DNA will be sequenced, and will now in fact compete with the specifically bound IPed DNA for sequencing. This led to the low amount of IPed DNA in our sequencing samples that could be mapped to the human genome.

The unrelated DNA typically used for this purpose is salmon sperm DNA, which explains the fishy character of the top-expressed reads.

While we clearly were not able to work with these samples further<sup>2</sup>, what little useful data we had in them later was actually found to still be meaningful: analyzing the human genome mappable reads for enriched regions, I found MLL1 to be by far the strongest enriched target for HELLS, analogously to our later results with the mouse homologs, Mll1 and Hells.

---

<sup>1</sup><http://blast.ncbi.nlm.nih.gov/Blast.cgi>

<sup>2</sup>We contacted the Salmon Genome Sequencing consortium, but they were not interested in these data.



# Bibliography

- Agafonov, D. E., Deckert, J., Wolf, E., Odenwalder, P., Bessonov, S., Will, C. L., Urlaub, H., and Luhrmann, R. (2011). Semiquantitative proteomic analysis of the human spliceosome via a novel two-dimensional gel electrophoresis method. *Mol Cell Biol* 31 (13): 2667–2682.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). *Molecular Biology of the Cell*. 5th ed. Garland Science.
- Amari, S. and Douglas, S. C. (1998). Why natural gradient? In: *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP '98*. Vol. 2: 1213–1216.
- Ambros, V. (2001). microRNAs: tiny regulators with great potential. *Cell* 107 (7): 823–826.
- Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T., and Jewell, D. (2003). MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr Biol* 13 (10): 807–818.
- Anders, G., Mackowiak, S. D., Jens, M., Maaskola, J., Kuntzagk, A., Rajewsky, N., Landthaler, M., and Dieterich, C. (2012). doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res* 40 (Database issue): D180–D186.
- Ao, W., Gaudet, J., Kent, W. J., Muttumu, S., and Mango, S. E. (2004). Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* 305 (5691): 1743–1746.
- Aravin, A. A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M. J., Kuramochi-Miyagawa, S., Nakano, T., Chien, M., Russo, J. J., Ju, J., Sheridan, R., Sander, C., Zavolan, M., and Tuschl, T. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442 (7099): 203–207.
- Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G. J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* 316 (5825): 744–747.
- Aravin, A. A. and Tuschl, T. (2005). Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett* 579 (26): 5830–5840.
- Asp, P., Acosta-Alvear, D., Tsikitis, M., van Oevelen, C., and Dynlacht, B. D. (2009). E2f3b plays an essential role in myogenic differentiation through isoform-specific gene regulation. *Genes Dev* 23 (1): 37–53.

- Ast, G. (2004). How did alternative splicing evolve? *Nat Rev Genet* 5 (10): 773–782.
- Attisano, L. and Wrana, J. L. (2002). Signal transduction by the TGF-beta superfamily. *Science* 296 (5573): 1646–1647.
- Avery, O. T., Macleod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med* 79 (2): 137–158.
- Avilion, A. A., Nicolis, S. K., Pevny, L. H., Perez, L., Vivian, N., and Lovell-Badge, R. (2003). Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev* 17 (1): 126–140.
- Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. (2008). The impact of microRNAs on protein output. *Nature* 455 (7209): 64–71.
- Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27 (12): 1653–1659.
- Bailey, T. L. and Elkan, C. (1995a). The value of prior knowledge in discovering motifs with MEME. In: *Proceedings of the Third International Conference on Intelligent Systems for Molecular biology*. Ed. by C. Rawlings, D. Clark, R. Altman, L. Hunter, T. L. T, and S. Wodak: 21–25.
- (1995b). Unsupervised Learning of Multiple Motifs In Biopolymers Using EM. *Mach. Learn.* 21: 51–80.
- Bailey, T. L. and Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res* 40 (17): e128.
- Bailey, T. L., Williams, N., Mischak, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 34 (Web Server issue): W369–W373.
- Barash, Y., Bejerano, G., and Friedman, N. (2001). A Simple Hyper-Geometric Approach for Discovering Putative Transcription Factor Binding Sites. In: *WABI '01 Proceedings of the First International Workshop on Algorithms in Bioinformatics*.
- Barker, D. D., Wang, C., Moore, J., Dickinson, L. K., and Lehmann, R. (1992). Pumilio is essential for function but not for distribution of the Drosophila abdominal determinant Nanos. *Genes Dev* 6 (12A): 2312–2326.
- Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116 (2): 281–297.
- (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136 (2): 215–233.
- Batista, P. J., Ruby, J. G., Claycomb, J. M., Chiang, R., Fahlgren, N., Kasschau, K. D., Chaves, D. A., Gu, W., Vasale, J. J., Duan, S., Conte Jr, D., Luo, S., Schroth, G. P., Carrington, J. C., Bartel, D. P., and Mello, C. C. (2008). PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol Cell* 31 (1): 67–78.

- Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L., and Hunter, C. P. (2003). Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development* 130 (5): 889–900.
- Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* 3 (1): 1–8.
- Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* 73: 360–363.
- Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics* 41 (1): 164–171. ISSN: 00034851.
- Bechara, E. G., Sebestyén, E., Bernardis, I., Eyra, E., and Valcárcel, J. (2013). RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. *Mol Cell* 52 (5): 720–733.
- Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., and Langley, C. H. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol* 5 (11): e310.
- Behzadnia, N., Golas, M. M., Hartmuth, K., Sander, B., Kastner, B., Deckert, J., Dube, P., Will, C. L., Urlaub, H., Stark, H., and Lührmann, R. (2007). Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal A complexes. *EMBO J* 26 (6): 1737–1748.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57 (1): 289–300.
- Benjamini, Y. and Yekutieli, Y. (2005). False discovery rate controlling confidence intervals for selected parameters. *Journal of the American Statistical Association* 100 (469): 71–80.
- Benos, P. V., Bulyk, M. L., and Stormo, G. D. (2002). Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* 30 (20): 4442–4451.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004). GenBank: update. *Nucleic Acids Res* 32 (Database issue): D23–D26.
- Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27 (2): 573–580.
- Bentley, D. R. (2006). Whole-genome re-sequencing. *Curr Opin Genet Dev* 16 (6): 545–552.
- Bentwich, I. (2005). Prediction and validation of microRNAs and their targets. *FEBS Lett* 579 (26): 5904–5910.

- Berezikov, E. (2011). Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet* 12 (12): 846–860.
- Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R. H. A. (2006). Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* 38 (12): 1375–1377.
- Berninger, P., Gaidatzis, D., van Nimwegen, E., and Zavolan, M. (2008). Computational analysis of small RNA cloning data. *Methods* 44 (1): 13–21.
- Bieda, M., Xu, X., Singer, M. A., Green, R., and Farnham, P. J. (2006). Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Res* 16 (5): 595–605.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Black, A. R., Black, J. D., and Azizkhan-Clifford, J. (2001). Sp1 and krüppel-like factor family of transcription factors in cell growth regulation and cancer. *J Cell Physiol* 188 (2): 143–160.
- Blais, A. and Dynlacht, B. D. (2007). E2F-associated chromatin modifiers and cell cycle control. *Curr Opin Cell Biol* 19 (6): 658–662.
- Boffelli, D., Weer, C. V., Weng, L., Lewis, K. D., Shoukry, M. I., Pachter, L., Keys, D. N., and Rubin, E. M. (2004). Intraspecies sequence comparisons for annotating genomes. *Genome Res* 14 (12): 2406–2411.
- Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R., and Young, R. A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122 (6): 947–956.
- Bozek, K., Rosahl, A. L., Gaub, S., Lorenzen, S., and Herzog, H. (2010). Circadian transcription in liver. *Biosystems* 102 (1): 61–69.
- Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128 (6): 1089–1103.
- Brodersen, P. and Voinnet, O. (2009). Revisiting the principles of microRNA target recognition and mode of action. *Nat Rev Mol Cell Biol* 10 (2): 141–148.
- Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Application* 6: 76–90.
- Burkhardt, D. L. and Sage, J. (2008). Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nat Rev Cancer* 8 (9): 671–682.
- Burset, M. and Guigó, R. (1996). Evaluation of gene structure prediction programs. *Genomics* 34 (3): 353–367.

- Bushati, N. and Cohen, S. M. (2007). microRNA functions. *Annu Rev Cell Dev Biol* 23: 175–205.
- Campo-Paysaa, F., Sémon, M., Cameron, R. A., Peterson, K. J., and Schubert, M. (2011). microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evol Dev* 13 (1): 15–27.
- Cappé, O., Moulines, E., and Rydén, T. (2010). *Inference in Hidden Markov Models*. Springer.
- Caputi, M., Casari, G., Guenzi, S., Tagliabue, R., Sidoli, A., Melo, C. A., and Baralle, F. E. (1994). A novel bipartite splicing enhancer modulates the differential processing of the human fibronectin EDA exon. *Nucleic Acids Res* 22 (6): 1018–1022.
- Cartwright, P., McLean, C., Sheppard, A., Rivett, D., Jones, K., and Dalton, S. (2005). LIF / STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. *Development* 132 (5): 885–896.
- Chambers, I., Colby, D., Robertson, M., Nichols, J., Lee, S., Tweedie, S., and Smith, A. (2003). Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell* 113 (5): 643–655.
- Chen, K. and Rajewsky, N. (2006). Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet* 38 (12): 1452–1456.
- (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* 8 (2): 93–103.
- Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L., and Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133 (6): 1106–1117.
- Chittenden, T., Livingston, D. M., and Kaelin Jr, W. (1991). The T/E1A-binding domain of the retinoblastoma product can interact selectively with a sequence-specific DNA-binding protein. *Cell* 65 (6): 1073–1082.
- Chong, J.-L., Wenzel, P. L., Sáenz-Robles, M. T., Nair, V., Ferrey, A., Hagan, J. P., Gomez, Y. M., Sharma, N., Chen, H., Ouseph, M., Wang, S., Trikha, P., Culp, B., Mezache, L., Winton, D. J., Sansom, O. J., Chen, D., Bremner, R., Cantalupo, P. G., Robinson, M. L., Pipas, J. M., and Leone, G. (2009). E2f1-3 switch from activators in progenitor cells to repressors in differentiating cells. *Nature* 462 (7275): 930–934.
- Clerte, C. and Hall, K. B. (2009). The domains of polypyrimidine tract binding protein have distinct RNA structural preferences. *Biochemistry* 48 (10): 2063–2074.
- Cole, M. F., Johnstone, S. E., Newman, J. J., Kagey, M. H., and Young, R. A. (2008). Tcf3 is an integral component of the core regulatory circuitry of embryonic stem cells. *Genes Dev* 22 (6): 746–755.

- Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., and Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol* 12 (8): R79.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience.
- Crittenden, S. L., Bernstein, D. S., Bachorik, J. L., Thompson, B. E., Gallegos, M., Petcherski, A. G., Moulder, G., Barstead, R., Wickens, M., and Kimble, J. (2002). A conserved RNA-binding protein controls germline stem cells in *Caenorhabditis elegans*. *Nature* 417 (6889): 660–663.
- Czech, B. and Hannon, G. J. (2011). Small RNA sorting: matchmaking for Argonautes. *Nat Rev Genet* 12 (1): 19–31.
- Deckert, J., Hartmuth, K., Boehringer, D., Behzadnia, N., Will, C. L., Kastner, B., Stark, H., Urlaub, H., and Lührmann, R. (2006). Protein composition and electron microscopy structure of affinity-purified human spliceosomal B complexes isolated under physiological conditions. *Mol Cell Biol* 26 (14): 5528–5543.
- Declercq, J., Sheshadri, P., Verfaillie, C. M., and Kumar, A. (2013). Zic3 enhances the generation of mouse induced pluripotent stem cells. *Stem Cells Dev* 22 (14): 2017–2025.
- DeGregori, J. and Johnson, D. G. (2006). Distinct and Overlapping Roles for E2F Family Members in Transcription, Proliferation and Apoptosis. *Curr Mol Med* 6 (7): 739–748.
- Dembo, A., Cover, T. M., and Thomas, J. A. (1991). Information theoretic inequalities. *IEEE Transactions on Information Theory* 37 (6): 1501–1518.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1): 1–38.
- Denli, A. M., Tops, B. B. J., Plasterk, R. H. A., Ketting, R. F., and Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature* 432 (7014): 231–235.
- Dennis, K., Fan, T., Geiman, T., Yan, Q., and Muegge, K. (2001). Lsh, a member of the SNF2 family, is required for genome-wide methylation. *Genes Dev* 15 (22): 2940–2944.
- Diederichs, S. and Haber, D. A. (2006). Sequence variations of microRNAs in human cancer: alterations in predicted secondary structure do not affect processing. *Cancer Res* 66 (12): 6097–6104.
- Dudley, R. M. (Spring 2003). Mathematical Statistics, 18.466 lecture notes. In: MIT OCW (OpenCourseWare). Chap. 3.9. URL: <http://ocw.mit.edu/courses/mathematics/18-466-mathematical-statistics-spring-2003/lecture-notes/>.
- Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.
- Earl, D. J. and Deem, M. W. (2005). Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* 7 (23): 3910–3916.



- Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* 14 (9): 755–763.
- Edgar, L. G., Wolf, N., and Wood, W. B. (1994). Early transcription in *Caenorhabditis elegans* embryos. *Development* 120 (2): 443–451.
- Ehrenreich, I. and Purugganan, M. (2008). Sequence variation of microRNAs and their binding sites in *Arabidopsis thaliana*. *Plant Physiol* 146: 1974–1982.
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* 323 (5910): 133–138.
- Elemento, O., Slonim, N., and Tavazoie, S. (2007). A universal framework for regulatory element discovery across all genomes and data types. *Mol Cell* 28 (2): 337–350.
- Ellington, A. D. and Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature* 346 (6287): 818–822.
- Esteller, M. (2011). Non-coding RNAs in human disease. *Nat Rev Genet* 12 (12): 861–874.
- Evans, T. C. and Hunter, C. P. (2005). Translational control of maternal RNAs. *WormBook*: 1–11.
- Fabian, M. R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem* 79: 351–379.
- Fairbrother, W. G., Holste, D., Burge, C. B., and Sharp, P. A. (2004). Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol* 2 (9): E268.
- Fairbrother, W. G., Yeh, R.-F., Sharp, P. A., and Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes. *Science* 297 (5583): 1007–1013.
- Feber, A., Clark, J., Goodwin, G., Dodson, A. R., Smith, P. H., Fletcher, A., Edwards, S., Flohr, P., Falconer, A., Roe, T., Kovacs, G., Dennis, N., Fisher, C., Wooster, R., Huddart, R., Foster, C. S., and Cooper, C. S. (2004). Amplification and overexpression of E2F3 in human bladder cancer. *Oncogene* 23 (8): 1627–1630.
- Feng, B., Jiang, J., Kraus, P., Ng, J., Heng, J.-C. D., Chan, Y., Yaw, L., Zhang, W., Loh, Y.-H., Han, J., Vega, V. B., Cacheux-Rataboul, V., Lim, B., Lufkin, T., and Ng, H.-H. (2009). Reprogramming of fibroblasts into induced pluripotent stem cells with orphan nuclear receptor Esrrb. *Nat Cell Biol* 11 (2): 197–203.
- Fernandez, A. G., Gunsalus, K. C., Huang, J., Chuang, L.-S., Ying, N., Liang, H.-L., Tang, C., Schetter, A. J., Zegar, C., Rual, J.-F., Hill, D. E., Reinke, V., Vidal, M., and Piano, F. (2005).

- New genes with roles in the *C. elegans* embryo revealed using RNAi of ovary-enriched ORFeome clones. *Genome Res* 15 (2): 250–259.
- Fisher, R. A. (1922). On the interpretation of  $\chi^2$  from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 85 (1): 87–94.
- Fletcher, R. (1970). A New Approach to Variable Metric Algorithms. *Computer Journal* 13 (3): 317–322.
- Foster, C. S., Falconer, A., Dodson, A. R., Norman, A. R., Dennis, N., Fletcher, A., Southgate, C., Dowe, A., Dearnaley, D., Jhavar, S., Eeles, R., Feber, A., and Cooper, C. S. (2004). Transcription factor E2F3 overexpressed in prostate cancer independently predicts clinical outcome. *Oncogene* 23 (35): 5871–5879.
- Fox, M., Urano, J., and Reijo Pera, R. A. (2005). Identification and characterization of RNA sequences to which human PUMILIO-2 (PUM2) and deleted in Azoospermia-like (DAZL) bind. *Genomics* 85 (1): 92–105.
- Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 26 (4): 407–415.
- Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 40 (1): 37–52.
- Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 19 (1): 92–105.
- Galan-Caridad, J. M., Harel, S., Arenzana, T. L., Hou, Z. E., Doetsch, F. K., Mirny, L. A., and Reizis, B. (2007). Zfx controls the self-renewal of embryonic and hematopoietic stem cells. *Cell* 129 (2): 345–357.
- Galgano, A., Forrer, M., Jaskiewicz, L., Kanitz, A., Zavolan, M., and Gerber, A. P. (2008). Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. *PLoS One* 3 (9): e3164.
- García-Blanco, M. A., Jamison, S. F., and Sharp, P. A. (1989). Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. *Genes Dev* 3 (12A): 1874–1886.
- Gartel, A. L., Ye, X., Goufman, E., Shianov, P., Hay, N., Najmabadi, F., and Tyner, A. L. (2001). Myc represses the p21(WAF1/CIP1) promoter and interacts with Sp1/Sp3. *Proc Natl Acad Sci U S A* 98 (8): 4510–4515.
- Georgiev, S., Boyle, A. P., Jayasurya, K., Ding, X., Mukherjee, S., and Ohler, U. (2010). Evidence-ranked motif identification. *Genome Biol* 11 (2): R19.
- Gerber, A. P., Herschlag, D., and Brown, P. O. (2004). Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol* 2 (3): E79.

- Gerber, A. P., Luschnig, S., Krasnow, M. A., Brown, P. O., and Herschlag, D. (2006). Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* 103 (12): 4487–4492.
- Ghildiyal, M. and Zamore, P. (2009). Small silencing RNAs: an expanding universe. *Nat Rev Genet.* 10 (2): 94–108.
- Gibbs, J. W. (1902). *Elementary principles in statistical mechanics: Developed with especial reference to the rational foundation of thermodynamics*. C. Scribner's Sons.
- Girard, A., Sachidanandam, R., Hannon, G. J., and Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442 (7099): 199–202.
- Goldfarb, D. (1970). A Family of Variable Metric Updates Derived by Variational Means. *Mathematics of Computation* 24 (109): 23–26.
- Gönczy, P. and Rose, L. S. (2005). Asymmetric cell division and axis formation in the embryo. *WormBook*: 1–20.
- Gopisetty, G., Xu, J., Sampath, D., Colman, H., and Pudevalli, V. K. (2013). Epigenetic regulation of CD133/PROM1 expression in glioma stem cells by Sp1/myc and promoter methylation. *Oncogene* 32 (26): 3119–3129.
- Griffith, F. (1928). The Significance of Pneumococcal Types. *J Hyg (Lond)* 27 (2): 113–159.
- Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res* 36 (Database issue): D154–D158.
- Gripp, K. W., Hopkins, E., Johnston, J. J., Krause, C., Dobyms, W. B., and Biesecker, L. G. (2011). Long-term survival in TARP syndrome and confirmation of RBM10 as the disease-causing gene. *Am J Med Genet A* 155A (10): 2516–2520.
- Grishok, A., Pasquinelli, A. E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D. L., Fire, A., Ruvkun, G., and Mello, C. C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* 106 (1): 23–34.
- Grivna, S. T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* 20 (13): 1709–1714.
- Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M. C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science* 315 (5818): 1587–1590.
- Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466 (7308): 835–840.
- Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biol* 8 (2): R24.

- Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press.
- Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141 (1): 129–141.
- Han, J., Lee, Y., Yeom, K., Nam, J., Heo, I., Rhee, J.-K., Sohn, S. Y., Cho, Y., Zhang, B.-T., and Kim, V. N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* 125 (5): 887–901.
- Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144 (5): 646–674.
- Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495 (7441): 384–388.
- Hartley, R. V. L. (1928). Transmission of Information. *Bell System Technical Journal* 7 (3): 535–563.
- Hegele, A., Kamburov, A., Grossmann, A., Sourlis, C., Wowro, S., Weimann, M., Will, C. L., Pena, V., Lührmann, R., and Stelzl, U. (2012). Dynamic protein-protein interaction wiring of the human spliceosome. *Mol Cell* 45 (4): 567–580.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38 (4): 576–589.
- Hendrickson, D. G., Hogan, D. J., McCullough, H. L., Myers, J. W., Herschlag, D., Ferrell, J. E., and Brown, P. O. (2009). Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biol* 7 (11): e1000238.
- Hiller, M., Pudimat, R., Busch, A., and Backofen, R. (2006). Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res* 34 (17): e117.
- Hochberg, Y. and Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Stat Med* 9 (7): 811–818.
- Hochberg, Y. (1988). A Sharper Bonferroni Procedure for Multiple Tests of Significance. *Biometrika* 75 (4): 800–802.
- Hochberg, Y. and Tamhane, A. C. (1987). *Multiple comparison procedures*. Wiley.
- Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie* 125: 167–188.

- Hornstein, E. and Shomron, N. (2006). Canalization of development by microRNAs. *Nat Genet* 38 Suppl: S20–S24.
- Houwing, S., Kamminga, L. M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D. V., Blaser, H., Raz, E., Moens, C. B., Plasterk, R. H. A., Hannon, G. J., Draper, B. W., and Ketting, R. F. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* 129 (1): 69–82.
- Huggins, P., Zhong, S., Schiff, I., Beckerman, R., Laptenko, O., Prives, C., Schulz, M. H., Simon, I., and Bar-Joseph, Z. (2011). DECOD: fast and accurate discriminative DNA motif finding. *Bioinformatics* 27 (17): 2361–2367.
- Huntzinger, E. and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* 12 (2): 99–110.
- Hurst, C. D., Tomlinson, D. C., Williams, S. V., Platt, F. M., and Knowles, M. A. (2008). Inactivation of the Rb pathway and overexpression of both isoforms of E2F3 are obligate events in bladder tumours with 6p22 amplification. *Oncogene* 27 (19): 2716–2727.
- Imielinski, M., Berger, A. H., Hammerman, P. S., Hernandez, B., Pugh, T. J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., Sougnez, C., Auclair, D., Lawrence, M. S., Stojanov, P., Cibulskis, K., Choi, K., de Waal, L., Sharifnia, T., Brooks, A., Greulich, H., Banerji, S., Zander, T., Seidel, D., Leenders, F., Ansén, S., Ludwig, C., Engel-Riedel, W., Stoelben, E., Wolf, J., Goparju, C., Thompson, K., Winckler, W., Kwiatkowski, D., Johnson, B. E., Jänne, P. A., Miller, V. A., Pao, W., Travis, W. D., Pass, H. I., Gabriel, S. B., Lander, E. S., Thomas, R. K., Garraway, L. A., Getz, G., and Meyerson, M. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 150 (6): 1107–1120.
- Inoue, A., Yamamoto, N., Kimura, M., Nishio, K., Yamane, H., and Nakajima, K. (2014). RBM10 regulates alternative splicing. *FEBS Lett* 588 (6): 942–947.
- Ivanova, N., Dobrin, R., Lu, R., Kotenko, I., Levorse, J., DeCoste, C., Schafer, X., Lun, Y., and Lemischka, I. R. (2006). Dissecting self-renewal in stem cells with RNA interference. *Nature* 442 (7102): 533–538.
- Iwai, N. and Naraba, H. (2005). Polymorphisms in human pre-miRNAs. *Biochem Biophys Res Commun* 331 (4): 1439–1444.
- Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409 (6819): 533–538.
- Jensen, J. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica* 30 (1): 175–193. ISSN: 0001-5962.
- Jenuwein, T. and Allis, C. D. (2001). Translating the histone code. *Science* 293 (5532): 1074–1080.

- Jiang, J., Chan, Y., Loh, Y.-H., Cai, J., Tong, G., Lim, C.-A., Robson, P., Zhong, S., and Ng, H.-H. (2008). A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat Cell Biol* 10 (3): 353–360.
- Jiang, P., Singh, M., and Collier, H. A. (2013). Computational assessment of the cooperativity between RNA binding proteins and MicroRNAs in Transcript Decay. *PLoS Comput Biol* 9 (5): e1003075.
- Jin, W., Niu, Z., Xu, D., and Li, X. (2012). RBM5 promotes exon 4 skipping of AID pre-mRNA by competing with the binding of U2AF65 to the polypyrimidine tract. *FEBS Lett* 586 (21): 3852–3857.
- Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316 (5830): 1497–1502.
- Johnston, J. J., Teer, J. K., Cherukuri, P. F., Hansen, N. F., Loftus, S. K., NIH Intramural Sequencing Center, Chong, K., Mullikin, J. C., and Biesecker, L. G. (2010). Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am J Hum Genet* 86 (5): 743–748.
- Juliano, C. E., Voronina, E., Stack, C., Aldrich, M., Cameron, A. R., and Wessel, G. M. (2006). Germ line determinants are not localized early in sea urchin development, but do accumulate in the small micromere lineage. *Dev Biol* 300 (1): 406–415.
- Juliano, C., Wang, J., and Lin, H. (2011). Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. *Annu Rev Genet* 45: 447–469.
- Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16 (9): 418–420.
- Kaczynski, J., Cook, T., and Urrutia, R. (2003). Sp1- and Krüppel-like transcription factors. *Genome Biol* 4 (2): 206.
- Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., Welchman, D. P., Zipperlen, P., and Ahringer, J. (2003). Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421 (6920): 231–237.
- Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32 (Database issue): D493–D496.
- Karolchik, D., Kuhn, R. M., Baertsch, R., Barber, G. P., Clawson, H., Diekhans, M., Gardine, B., Harte, R. A., Hinrichs, A. S., Hsu, F., Kober, K. M., Miller, W., Pedersen, J. S., Pohl, A., Raney, B. J., Rhead, B., Rosenbloom, K. R., Smith, K. E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A. S., Haussler, D., and Kent, W. J. (2008). The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res* 36 (Database issue): D773–D779.

- Kedde, M., van Kouwenhove, M., Zwart, W., Oude Vrielink, J. A. F., Elkon, R., and Agami, R. (2010). A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. *Nat Cell Biol* 12(10): 1014–1020.
- Keilwagen, J., Grau, J., Paponov, I. A., Posch, S., Strickert, M., and Grosse, I. (2011). De novo discovery of differentially abundant transcription factor binding sites including their positional preference. *PLoS Comput Biol* 7(2): e1001070.
- Kennedy, B. K., Gotta, M., Sinclair, D. A., Mills, K., McNabb, D. S., Murthy, M., Pak, S. M., Laroche, T., Gasser, S. M., and Guarente, L. (1997). Redistribution of silencing proteins from telomeres to the nucleolus is associated with extension of life span in *S. cerevisiae*. *Cell* 89(3): 381–391.
- Kershner, A. M. and Kimble, J. (2010). Genome-wide analysis of mRNA targets for *Caenorhabditis elegans* FBF, a conserved stem cell regulator. *Proc Natl Acad Sci U S A* 107(8): 3936–3941.
- Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat Genet* 39(10): 1278–1284.
- Kim, N., Tharakaraman, K., Mariño-Ramírez, L., and Spouge, J. L. (2008). Finding sequence motifs with Bayesian models incorporating positional information: an application to transcription factor binding sites. *BMC Bioinformatics* 9: 262.
- Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenko, V. V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 128(6): 1231–1245.
- Knight, S. W. and Bass, B. L. (2001). A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*. *Science* 293(5538): 2269–2271.
- Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models, Principles and Techniques*. MIT Press.
- Kong, L.-J., Chang, J. T., Bild, A. H., and Nevins, J. R. (2007). Compensation and specificity of function within the E2F family. *Oncogene* 26(3): 321–327.
- König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* 17(7): 909–915.
- Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., and Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nat Genet* 37(5): 495–500.
- Krogh, A. (1994). Hidden Markov models for labeled sequences. In: *Proc. 12th IAPR Int. Pattern Recognition Vol. 2 - Conf. B: Computer Vision & Image Processing, Conf. Vol. 2: 140–144*.

- Krol, J., Loedige, I., and Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nature Reviews Genetics* 11 (9): 597–610. ISSN: 1471-0056.
- Ku, C. S., Naidoo, N., Wu, M., and Soong, R. (2011). Studying the epigenome using next generation sequencing. *J Med Genet* 48 (11): 721–730.
- Kuhn, A. N., van Santen, M. A., Schwienhorst, A., Urlaub, H., and Lührmann, R. (2009). Stalling of spliceosome assembly at distinct stages by small-molecule inhibitors of protein acetylation and deacetylation. *RNA* 15 (1): 153–175.
- Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Ann. Math. Stat.* 22: 79–86.
- Kullback, S. (1959). *Information Theory and Statistics*. John Wiley & Sons, Inc.
- Kuroda, T., Tada, M., Kubota, H., Kimura, H., Hatano, S.-y., Suemori, H., Nakatsuji, N., and Tada, T. (2005). Octamer and Sox elements are required for transcriptional cis regulation of Nanog gene expression. *Mol Cell Biol* 25 (6): 2475–2485.
- Kyo, S., Takakura, M., Taira, T., Kanaya, T., Itoh, H., Yutsudo, M., Ariga, H., and Inoue, M. (2000). Sp1 cooperates with c-Myc to activate transcription of the human telomerase reverse transcriptase gene (hTERT). *Nucleic Acids Res* 28 (3): 669–677.
- Lai, E. C., Tomancak, P., Williams, R. W., and Rubin, G. M. (2003). Computational identification of Drosophila microRNA genes. *Genome Biol* 4 (7): R42.
- Lall, S., Grün, D., Krek, A., Chen, K., Wang, Y.-L., Dewey, C. N., Sood, P., Colombo, T., Bray, N., Macmenamin, P., Kao, H.-L., Gunsalus, K. C., Pachter, L., Piano, F., and Rajewsky, N. (2006). A genome-wide map of conserved microRNA targets in *C. elegans*. *Curr Biol* 16 (5): 460–471.
- Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A. A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., Lin, C., Socci, N. D., Hermida, L., Fulci, V., Chiaretti, S., Foà, R., Schliwka, J., Fuchs, U., Novosel, A., Müller, R.-U., Schermer, B., Bissels, U., Inman, J., Phan, Q., Chien, M., Weir, D. B., Choksi, R., De Vita, G., Frezzetti, D., Trompeter, H.-I., Hornung, V., Teng, G., Hartmann, G., Palkovits, M., Di Lauro, R., Wernet, P., Macino, G., Rogler, C. E., Nagle, J. W., Ju, J., Papavasiliou, F. N., Benzing, T., Lichter, P., Tam, W., Brownstein, M. J., Bosio, A., Borkhardt, A., Russo, J. J., Sander, C., Zavolan, M., and Tuschl, T. (2007). A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 129 (7): 1401–1414.
- Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10 (3): R25.
- Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294 (5543): 858–862.



- Lau, N. C., Seto, A. G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D. P., and Kingston, R. E. (2006). Characterization of the piRNA complex from rat testes. *Science* 313 (5785): 363–367.
- Lavigueur, A., Branche, H. L., Kornblihtt, A. R., and Chabot, B. (1993). A splicing enhancer in the human fibronectin alternate ED1 exon interacts with SR proteins and stimulates U2 snRNP binding. *Genes Dev* 7 (12A): 2405–2417.
- Lee, J. T. and Bartolomei, M. S. (2013). X-inactivation, imprinting, and long noncoding RNAs in health and disease. *Cell* 152 (6): 1308–1323.
- Lee, J.-S., Smith, E., and Shilatifard, A. (2010). The language of histone crosstalk. *Cell* 142 (5): 682–685.
- Lee, R. C. and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294 (5543): 862–864.
- Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., and Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425 (6956): 415–419.
- Leibovich, L., Mandel-Gutfreund, Y., and Yakhini, Z. (2010). A structural-based statistical approach suggests a cooperative activity of PUM1 and miR-410 in human 3'-untranslated regions. *Silence* 1 (1): 17.
- Leibovich, L., Paz, I., Yakhini, Z., and Mandel-Gutfreund, Y. (2013). DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. *Nucleic Acids Res* 41 (Web Server issue): W174–W179.
- Leibovich, L. and Yakhini, Z. (2012). Efficient motif search in ranked lists and applications to variable gap motifs. *Nucleic Acids Res* 40 (13): 5832–5847.
- Leung, H. C. M. and Chin, F. Y. L. (2006). Finding motifs from all sequences with and without binding sites. *Bioinformatics* 22 (18): 2217–2223.
- Levenberg, K. (1944). A Method for the Solution of Certain Non-Linear Problems in Least Squares. *The Quarterly of Applied Mathematics* 2: 164–168.
- Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120 (1): 15–20.
- Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456 (7221): 464–469.
- Lieber, D. S., Elemento, O., and Tavazoie, S. (2010). Large-scale discovery and characterization of protein regulatory motifs in eukaryotes. *PLoS One* 5 (12): e14444.
- Lim, L. S., Hong, F. H., Kunarso, G., and Stanton, L. W. (2010). The pluripotency regulator *Zic3* is a direct activator of the *Nanog* promoter in ESCs. *Stem Cells* 28 (11): 1961–1969.

- Lim, L. S., Loh, Y.-H., Zhang, W., Li, Y., Chen, X., Wang, Y., Bakre, M., Ng, H.-H., and Stanton, L. W. (2007). *Zic3* is required for maintenance of pluripotency in embryonic stem cells. *Mol Biol Cell* 18 (4): 1348–1358.
- Lin, T.-H., Murphy, R. F., and Bar-Joseph, Z. (2011). Discriminative motif finding for predicting protein subcellular localization. *IEEE/ACM Trans Comput Biol Bioinform* 8 (2): 441–451.
- Lindgren, B. (1993 and 1998). *Statistical Theory*. 4th ed. Chapman & Hall.
- Liu, H., Cheng, E. H.-Y., and Hsieh, J. J.-D. (2007). Bimodal degradation of MLL by SCF-Skp2 and APC/Cdc20 assures cell cycle execution: a critical regulatory circuit lost in leukemogenic MLL fusions. *Genes Dev* 21 (19): 2385–2398.
- Liu, W., Tanasa, B., Tyurina, O. V., Zhou, T. Y., Gassmann, R., Liu, W. T., Ohgi, K. A., Benner, C., Garcia-Bassets, I., Aggarwal, A. K., Desai, A., Dorrestein, P. C., Glass, C. K., and Rosenfeld, M. G. (2010). PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. *Nature* 466 (7305): 508–512.
- Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* 20 (8): 835–839.
- Liu, X., Brutlag, D. L., and Liu, J. S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*: 127–138.
- Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., Wong, K., Sung, K. W., Lee, C. W. H., Zhao, X.-D., Chiu, K., Lipovich, L., Kuznetsov, V. A., Robson, P., Stanton, L. W., Wei, C.-L., Ruan, Y., Lim, B., and Ng, H.-H. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* 38 (4): 431–440.
- Loughlin, F. E., Mansfield, R. E., Vaz, P. M., McGrath, A. P., Setiyaputra, S., Gamsjaeger, R., Chen, E. S., Morris, B. J., Guss, J. M., and Mackay, J. P. (2009). The zinc fingers of the SR-like protein ZRANB2 are single-stranded RNA-binding domains that recognize 5' splice site-like sequences. *Proc Natl Acad Sci U S A* 106 (14): 5581–5586.
- Lu, J., Fu, Y., Kumar, S., Shen, Y., Zeng, K., Xu, A., Carthew, R., and Wu, C.-I. (2008). Adaptive evolution of newly emerged micro-RNA genes in *Drosophila*. *Mol Biol Evol* 25 (5): 929–938.
- Lu, J., Shen, Y., Wu, Q., Kumar, S., He, B., Shi, S., Carthew, R. W., Wang, S. M., and Wu, C.-I. (2008). The birth and death of microRNA genes in *Drosophila*. *Nat Genet* 40 (3): 351–355.
- Lunter, G., Ponting, C. P., and Hein, J. (2006). Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* 2 (1): e5.

- Maaskola, J. and Rajewsky, N. (2014). Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. *Nucleic Acids Res* 42 (21): 12995–13011.
- MacKay, D. J. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.
- Majoros, W. H., Lekprasert, P., Mukherjee, N., Skalsky, R. L., Corcoran, D. L., Cullen, B. R., and Ohler, U. (2013). MicroRNA target site identification by integrating sequence and binding information. *Nat Methods* 10 (7): 630–633.
- Manber, U. and Myers, G. (1993). Suffix Arrays: A New Method for On-Line String Searches. *SIAM Journal on Computing* 22 (5): 935–948. ISSN: 1095-7111.
- Mao, X. and Hu, G. (2001). Estimation of HMM parameters based on gradients. *Journal of Electronics (China)* 18 (3): 277–280.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in micro-fabricated high-density picolitre reactors. *Nature* 437 (7057): 376–380.
- Marquardt, D. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics* 11 (2): 431–441.
- Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., Guenther, M. G., Johnston, W. K., Wernig, M., Newman, J., Calabrese, J. M., Dennis, L. M., Volkert, T. L., Gupta, S., Love, J., Hannett, N., Sharp, P. A., Bartel, D. P., Jaenisch, R., and Young, R. A. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 134 (3): 521–533.
- Mason, M. J., Plath, K., and Zhou, Q. (2010). Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics* 26 (22): 2826–2832.
- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405 (2): 442–451.
- Mayr, C. and Bartel, D. P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* 138 (4): 673–684.
- Mehta, C. R. and Patel, N. R. (1983). A Network Algorithm for Performing Fisher's Exact Test in  $r \times c$  Contingency Tables. *Journal of the American Statistical Association* 78 (382): pp. 427–434. ISSN: 01621459.

- Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S. D., Gregersen, L. H., Munschauer, M., Loewer, A., Ziebold, U., Landthaler, M., Kocks, C., le Noble, F., and Rajewsky, N. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495 (7441): 333–338.
- Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448 (7153): 553–560.
- Miles, W. O., Tschöp, K., Herr, A., Ji, J., and Dyson, N. J. (2012). Pumilio facilitates miRNA regulation of the E2F3 oncogene. *Genes Dev* 26 (4): 356–368.
- Miranda, K. C., Huynh, T., Tay, Y., Ang, Y., Tam, W., Thomson, A. M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 126 (6): 1203–1217.
- Miska, E. A., Alvarez-Saavedra, E., Abbott, A. L., Lau, N. C., Hellman, A. B., McGonagle, S. M., Bartel, D. P., Ambros, V., and Horvitz, H. R. (2007). Most *Caenorhabditis elegans* microRNAs are individually not essential for development or viability. *PLoS Genet* 3 (12): e215.
- Mitchell, S. A., Spriggs, K. A., Bushell, M., Evans, J. R., Stoneley, M., Le Quesne, J. P. C., Spriggs, R. V., and Willis, A. E. (2005). Identification of a motif that mediates polypyrimidine tract-binding protein-dependent internal ribosome entry. *Genes Dev* 19 (13): 1556–1571.
- Mitsui, K., Tokuzawa, Y., Itoh, H., Segawa, K., Murakami, M., Takahashi, K., Maruyama, M., Maeda, M., and Yamanaka, S. (2003). The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell* 113 (5): 631–642.
- Moore, F. L., Jaruzelska, J., Fox, M. S., Urano, J., Firpo, M. T., Turek, P. J., Dorfman, D. M., and Pera, R. A. R. (2003). Human Pumilio-2 is expressed in embryonic stem cells and germ cells and interacts with DAZ (Deleted in AZoospermia) and DAZ-like proteins. *Proc Natl Acad Sci U S A* 100 (2): 538–543.
- Moré, J. J. and Thuente, D. J. (1994). Line search algorithms with guaranteed sufficient decrease. *ACM Trans. Math. Softw.* 20 (3): 286–307.
- Morris, A. R., Mukherjee, N., and Keene, J. D. (2008). Ribonomic analysis of human Pum1 reveals cis-trans conservation across species despite evolution of diverse mRNA target sets. *Mol Cell Biol* 28 (12): 4093–4103.
- Mount, D. (2004). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press.
- Murata, Y. and Wharton, R. P. (1995). Binding of pumilio to maternal hunchback mRNA is required for posterior patterning in *Drosophila* embryos. *Cell* 80 (5): 747–756.

- Myant, K. and Stancheva, I. (2008). LSH cooperates with DNA methyltransferases to repress transcription. *Mol Cell Biol* 28 (1): 215–226.
- Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320 (5881): 1344–1349.
- Narita, M., Núñez, S., Heard, E., Narita, M., Lin, A. W., Hearn, S. A., Spector, D. L., Hannon, G. J., and Lowe, S. W. (2003). Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell* 113 (6): 703–716.
- Nelson, D. L. and Cox, M. M. (2012). *Lehninger Principles of Biochemistry*. 6th ed. W.H. Freeman.
- Neyman, J. and Pearson, E. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231: 289–337.
- Ng, C. K. L., Li, N. X., Chee, S., Prabhakar, S., Kolatkar, P. R., and Jauch, R. (2012). Deciphering the Sox-Oct partner code by quantitative cooperativity measurements. *Nucleic Acids Res* 40 (11): 4933–4941.
- Nguyen, C. D., Mansfield, R. E., Leung, W., Vaz, P. M., Loughlin, F. E., Grant, R. P., and Mackay, J. P. (2011). Characterization of a family of RanBP2-type zinc fingers that can recognize single-stranded RNA. *J Mol Biol* 407 (2): 273–283.
- Nichols, J., Zevnik, B., Anastassiadis, K., Niwa, H., Klewe-Nebenius, D., Chambers, I., Schöler, H., and Smith, A. (1998). Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell* 95 (3): 379–391.
- Nielsen, S. J., Schneider, R., Bauer, U. M., Bannister, A. J., Morrison, A., O'Carroll, D., Firestein, R., Cleary, M., Jenuwein, T., Herrera, R. E., and Kouzarides, T. (2001). Rb targets histone H3 methylation and HP1 to promoters. *Nature* 412 (6846): 561–565.
- Nishimoto, M., Fukushima, A., Okuda, A., and Muramatsu, M. (1999). The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2. *Mol Cell Biol* 19 (8): 5453–5465.
- Niwa, H., Burdon, T., Chambers, I., and Smith, A. (1998). Self-renewal of pluripotent embryonic stem cells is mediated via activation of STAT3. *Genes Dev* 12 (13): 2048–2060.
- Niwa, H., Miyazaki, J., and Smith, A. G. (2000). Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. *Nat Genet* 24 (4): 372–376.
- Oegema, K. and Hyman, A. A. (2006). Cell division. *WormBook*: 1–40.
- Oeggerli, M., Tomovska, S., Schraml, P., Calvano-Forte, D., Schafroth, S., Simon, R., Gasser, T., Mihatsch, M. J., and Sauter, G. (2004). E2F3 amplification and overexpression is associated with invasive tumor growth and rapid tumor cell proliferation in urinary bladder cancer. *Oncogene* 23 (33): 5616–5623.

- The On-Line Encyclopedia of Integer Sequences*, Sequence A000670 (2013). URL: <http://oeis.org/A000670>.
- Oliphant, A. R., Brandl, C. J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Mol Cell Biol* 9 (7): 2944–2949.
- Olivas, W. and Parker, R. (2000). The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast. *EMBO J* 19 (23): 6602–6611.
- Pak, J. and Fire, A. (2007). Distinct populations of primary and secondary effectors during RNAi in *C. elegans*. *Science* 315 (5809): 241–244.
- Parisi, F., Wirapati, P., and Naef, F. (2007). Identifying synergistic regulation involving c-Myc and sp1 in human tissues. *Nucleic Acids Res* 35 (4): 1098–1107.
- Parisi, T., Yuan, T. L., Faust, A. M., Caron, A. M., Bronson, R., and Lees, J. A. (2007). Selective requirements for E2f3 in the development and tumorigenicity of Rb-deficient chimeric tissues. *Mol Cell Biol* 27 (6): 2283–2293.
- Pauli, A., Rinn, J. L., and Schier, A. F. (2011). Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet* 12 (2): 136–149.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5* 50: 157–175.
- Peng, J. C. and Lin, H. (2013). Beyond transposons: the epigenetic and somatic functions of the Piwi-piRNA mechanism. *Curr Opin Cell Biol* 25 (2): 190–194.
- Persson, H., Kvist, A., Vallon-Christersson, J., Medstrand, P., Borg, A., and Rovira, C. (2009). The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nat Cell Biol* 11 (10): 1268–1271.
- Peterson, K. J., Dietrich, M. R., and McPeck, M. A. (2009). MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *Bioessays* 31 (7): 736–747.
- Piano, F., Schetter, A. J., Mangone, M., Stein, L., and Kemphues, K. J. (2000). RNAi analysis of genes expressed in the ovary of *Caenorhabditis elegans*. *Curr Biol* 10 (24): 1619–1622.
- Piano, F., Schetter, A. J., Morton, D. G., Gunsalus, K. C., Reinke, V., Kim, S. K., and Kemphues, K. J. (2002). Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. *Curr Biol* 12 (22): 1959–1964.
- Piedade, I. da, Tang, M. E., and Elemento, O. (2009). DISPARE: DIScriminative PATtern REfinement for Position Weight Matrices. *BMC Bioinformatics* 10: 388.
- Press, W. H., Teukolsky, S. A., Flannery, B. P., and Vetterling, W. T. (1995). *Numerical Recipes in C*. Cambridge University Press.

- Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33 (Database issue): D501–D504.
- Rabani, M., Kertesz, M., and Segal, E. (2008). Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci U S A* 105 (39): 14885–14890.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77 (2): 257–286.
- Rabiner, L. R. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall.
- Rabinovich, A., Jin, V. X., Rabinovich, R., Xu, X., and Farnham, P. J. (2008). E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. *Genome Res* 18 (11): 1763–1777.
- Rahmann, S. (2003). Dynamic Programming Algorithms for Two Statistical Problems in Computational Biology. In: *Algorithms in Bioinformatics*. Vol. 2812. Lecture Notes in Computer Science: Lecture Notes in Bioinformatics. Springer: 151–164.
- Rajewsky, N. (2006). microRNA target predictions in animals. *Nat Genet* 38 Suppl: S8–13.
- (2011). MicroRNAs and the Operon paper. *J Mol Biol* 409 (1): 70–75.
- Ramchatesingh, J., Zahler, A. M., Neugebauer, K. M., Roth, M. B., and Cooper, T. A. (1995). A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer. *Mol Cell Biol* 15 (9): 4898–4907.
- Rando, O. J. (2012). Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr Opin Genet Dev* 22 (2): 148–155.
- Rappsilber, J., Ryder, U., Lamond, A. I., and Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Res* 12 (8): 1231–1245.
- Redhead, E. and Bailey, T. L. (2007). Discriminative motif discovery in DNA and protein sequences using the DEME algorithm. *BMC Bioinformatics* 8: 385.
- Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science* 290 (5500): 2306–2309.
- Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R. A., and Dynlacht, B. D. (2002). E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev* 16 (2): 245–256.
- Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M., and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4 (8): 651–657.

- Rodriguez, A. J., Seipel, S. A., Hamill, D. R., Romancino, D. P., DI Carlo, M., Suprenant, K. A., and Bonder, E. M. (2005). Seawi—a sea urchin piwi/argonaute family member is a component of MT-RNP complexes. *RNA* 11 (5): 646–656.
- Rouget, C., Papin, C., Boureux, A., Meunier, A.-C., Franco, B., Robine, N., Lai, E. C., Pelisson, A., and Simonelig, M. (2010). Maternal mRNA deadenylation and decay by the piRNA pathway in the early *Drosophila* embryo. *Nature* 467 (7319): 1128–1132.
- Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D. P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 127 (6): 1193–1207.
- Ruby, J. G., Stark, A., Johnston, W. K., Kellis, M., Bartel, D. P., and Lai, E. C. (2007). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res* 17 (12): 1850–1864.
- Saito, K., Nishida, K. M., Mori, T., Kawamura, Y., Miyoshi, K., Nagami, T., Siomi, H., and Siomi, M. C. (2006). Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev* 20 (16): 2214–2222.
- Sanford, J. R., Wang, X., Mort, M., Vanduyne, N., Cooper, D. N., Mooney, S. D., Edenberg, H. J., and Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res* 19 (3): 381–394.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74 (12): 5463–5467.
- Saulière, J., Murigneux, V., Wang, Z., Marquet, E., Barbosa, I., Le Tonquèze, O., Audic, Y., Paillard, L., Roest Crollius, H., and Le Hir, H. (2012). CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. *Nat Struct Mol Biol* 19 (11): 1124–1131.
- Saunders, M. A., Liang, H., and Li, W. (2007). Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci U S A* 104 (9): 3300–3305.
- Sawarkar, R. and Paro, R. (2010). Interpretation of developmental signaling at chromatin: the Polycomb perspective. *Dev Cell* 19 (5): 651–661.
- Sawicka, K., Bushell, M., Spriggs, K. A., and Willis, A. E. (2008). Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem Soc Trans* 36 (Pt 4): 641–647.
- Schauer, I. E. and Wood, W. B. (1990). Early *C. elegans* embryos are transcriptionally active. *Development* 110 (4): 1303–1317.
- Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18 (20): 6097–6100.
- Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J Mol Biol* 188 (3): 415–431.



- Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. *Nature* 455 (7209): 58–63.
- Senti, K.-A. and Brennecke, J. (2010). The piRNA pathway: a fly’s perspective on the guardian of the genome. *Trends Genet* 26 (12): 499–509.
- Seydoux, G. and Dunn, M. A. (1997). Transcriptionally repressed germ cells lack a sub-population of phosphorylated RNA polymerase II in early embryos of *Caenorhabditis elegans* and *Drosophila melanogaster*. *Development* 124 (11): 2191–2201.
- Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation* 24: 647–656.
- Shannon, C. E. (1948). A mathematical theory of Communication. *The Bell System Technical Journal* 27: 379–423, 623–656.
- Shapiro, J. A., Huang, W., Zhang, C., Hubisz, M. J., Lu, J., Turissini, D. A., Fang, S., Wang, H.-Y., Hudson, R. R., Nielsen, R., Chen, Z., and Wu, C.-I. (2007). Adaptive genic evolution in the *Drosophila* genomes. *Proc Natl Acad Sci U S A* 104 (7): 2271–2276.
- Siddharthan, R., Siggia, E. D., and Nimwegen, E. van (2005). PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol* 1 (7): e67.
- Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 13 (2): 238–241.
- Singh, R., Valcárcel, J., and Green, M. R. (1995). Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science* 268 (5214): 1173–1176.
- Sinha, S. (2006). On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics* 22 (14): e454–e463.
- Sinha, S. and Tompa, M. (2000). A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol* 8: 344–354.
- (2002). Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* 30 (24): 5549–5560.
- (2003). YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res* 31 (13): 3586–3588.
- Siomi, M. C., Sato, K., Pezic, D., and Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* 12 (4): 246–258.
- Smith, A. D., Sumazin, P., Das, D., and Zhang, M. Q. (2005). Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* 21 Suppl 1: i403–i412.
- Smith, A. D., Sumazin, P., Xuan, Z., and Zhang, M. Q. (2006). DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci USA* 103 (16): 6275–6280.

- Smith, A. D., Sumazin, P., and Zhang, M. Q. (2005). Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci USA* 102 (5): 1560–1565.
- Sokal, R. R. and Rohlf, F. J. (1969). *Biometry: Principles and Practice of Statistics in Biological Research*. W.H. Freeman & Co Ltd.
- Sönnichsen, B., Koski, L. B., Walsh, A., Marschall, P., Neumann, B., Brehm, M., Alleaume, A.-M., Artelt, J., Bettencourt, P., Cassin, E., Hewitson, M., Holz, C., Khan, M., Lazik, S., Martin, C., Nitzsche, B., Ruer, M., Stamford, J., Winzi, M., Heinkel, R., Röder, M., Finell, J., Häantsch, H., Jones, S. J. M., Jones, M., Piano, F., Gunsalus, K. C., Oegema, K., Gönczy, P., Coulson, A., Hyman, A. A., and Echeverri, C. J. (2005). Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature* 434 (7032): 462–469.
- Staden, R. (1989). Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* 5 (2): 89–96.
- Stark, A., Brennecke, J., Bushati, N., Russell, R. B., and Cohen, S. M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* 123 (6): 1133–1146.
- Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S.-W., Han, M. V., Maeder, M. L., Polansky, B. J., Robson, B. E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D. G., Eastman, D. A., Rice, M., Weir, M., Hahn, M. W., Park, Y., Dewey, C. N., Pachter, L., Kent, W. J., Haussler, D., Lai, E. C., Bartel, D. P., Hannon, G. J., Kaufman, T. C., Eisen, M. B., Clark, A. G., Smith, D., Celniker, S. E., Gelbart, W. M., and Kellis, M. (2007). Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* 450 (7167): 219–232.
- Stoeckius, M., Maaskola, J., Colombo, T., Rahn, H.-P., Friedländer, M. R., Li, N., Chen, W., Piano, F., and Rajewsky, N. (2009). Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression. *Nat Methods* 6 (10): 745–751.
- Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100 (16): 9440–9445.
- Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics* 16 (1): 16–23.
- Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res* 10 (9): 2997–3011.
- Stramer, O. and Tweedie, R. L. (1999). Langevin-Type Models II: Self-Targeting Candidates for MCMC Algorithms. *Methodology And Computing In Applied Probability* 1 (3): 307–328.

- Stroeher, V. L., Kennedy, B. P., Millen, K. J., Schroeder, D. F., Hawkins, M. G., Goszczynski, B., and McGhee, J. D. (1994). DNA-protein interactions in the *Caenorhabditis elegans* embryo: oocyte and embryonic factors that bind to the promoter of the gut-specific *ges-1* gene. *Dev Biol* 163 (2): 367–380.
- Sulston, J. E., Schierenberg, E., White, J. G., and Thomson, J. N. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev Biol* 100 (1): 64–119.
- Sumi, T., Tsuneyoshi, N., Nakatsuji, N., and Suemori, H. (2007). Apoptosis and differentiation of human embryonic stem cells induced by sustained activation of c-Myc. *Oncogene* 26 (38): 5564–5576.
- Sun, L. V., Chen, L., Greil, F., Negre, N., Li, T.-R., Cavalli, G., Zhao, H., Steensel, B. V., and White, K. P. (2003). Protein-DNA interaction mapping using genomic tiling path microarrays in *Drosophila*. *Proc Natl Acad Sci U S A* 100 (16): 9428–9433.
- Sun, L., Lee, D. W., Zhang, Q., Xiao, W., Raabe, E. H., Meeker, A., Miao, D., Huso, D. L., and Arceci, R. J. (2004). Growth retardation and premature aging phenotypes in mice with disruption of the SNF2-like gene, PASG. *Genes Dev* 18 (9): 1035–1046.
- Sutovsky, P. (2003). Ubiquitin-dependent proteolysis in mammalian spermatogenesis, fertilization, and sperm quality control: killing three birds with one stone. *Microsc Res Tech* 61 (1): 88–102.
- Suzuki, A., Raya, A., Kawakami, Y., Morita, M., Matsui, T., Nakashima, K., Gage, F. H., Rodríguez-Esteban, C., and Izpisua Belmonte, J. C. (2006). Nanog binds to Smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells. *Proc Natl Acad Sci U S A* 103 (27): 10294–10299.
- Tacke, R. and Manley, J. L. (1995). The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J* 14 (14): 3540–3551.
- Tadauchi, T., Matsumoto, K., Herskowitz, I., and Irie, K. (2001). Post-transcriptional regulation through the HO 3'-UTR by Mpt5, a yeast homolog of Pumilio and FBF. *EMBO J* 20 (3): 552–561.
- Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126 (4): 663–676.
- Takahashi, Y., Rayman, J. B., and Dynlacht, B. D. (2000). Analysis of promoter binding by the E2F and pRB families in vivo: distinct E2F proteins mediate activation and repression. *Genes Dev* 14 (7): 804–816.
- Tam, W., Lim, C. Y., Han, J., Zhang, J., Ang, Y., Ng, H.-H., Yang, H., and Lim, B. (2008). T-cell factor 3 regulates embryonic stem cell pluripotency and self-renewal by the transcriptional control of multiple lineage pathways. *Stem Cells* 26 (8): 2019–2031.
- Tarasov, V., Jung, P., Verdoodt, B., Lodygin, D., Epanchintsev, A., Menssen, A., Meister, G., and Hermeking, H. (2007). Differential regulation of microRNAs by p53 revealed by massively parallel sequencing: miR-34a is a p53 target that induces apoptosis and G1-arrest. *Cell Cycle* 6 (13): 1586–1593.

- Tenenbaum, S. A., Carson, C. C., Lager, P. J., and Keene, J. D. (2000). Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc Natl Acad Sci U S A* 97 (26): 14085–14090.
- Tokuzawa, Y., Kaiho, E., Maruyama, M., Takahashi, K., Mitsui, K., Maeda, M., Niwa, H., and Yamanaka, S. (2003). Fbx15 is a novel target of Oct3/4 but is dispensable for embryonic stem cell self-renewal and mouse development. *Mol Cell Biol* 23 (8): 2699–2708.
- Tomioka, M., Nishimoto, M., Miyagi, S., Katayanagi, T., Fukui, N., Niwa, H., Muramatsu, M., and Okuda, A. (2002). Identification of Sox-2 regulatory region which is under the control of Oct-3/4-Sox-2 complex. *Nucleic Acids Res* 30 (14): 3202–3213.
- Tompa, M., Li, N., Bailey, T. L., Church, G. M., Moor, B. D., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavese, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., Helden, J. van, Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 23 (1): 137–144.
- Triboulet, R. and Gregory, R. I. (2010). Pumilio turns on microRNA function. *Nat Cell Biol* 12 (10): 928–929.
- Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249 (4968): 505–510.
- Tyagi, S., Chabes, A. L., Wysocka, J., and Herr, W. (2007). E2F activation of S phase promoters via association with HCF-1 and the MLL family of histone H3K4 methyltransferases. *Mol Cell* 27 (1): 107–119.
- Urano, J., Fox, M. S., and Reijo Pera, R. A. (2005). Interaction of the conserved meiotic regulators, BOULE (BOL) and PUMILIO-2 (PUM2). *Mol Reprod Dev* 71 (3): 290–298.
- Vagin, V. V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P. D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* 313 (5785): 320–324.
- Valen, E., Sandelin, A., Winther, O., and Krogh, A. (2009). Discovery of regulatory elements is improved by a discriminatory approach. *PLoS Comput Biol* 5 (11): e1000562.
- Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., McKernan, K., Sidow, A., Fire, A., and Johnson, S. M. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res* 18 (7): 1051–1063.
- van den Berg, D. L. C., Snoek, T., Mullin, N. P., Yates, A., Bezstarosti, K., Demmers, J., Chambers, I., and Poot, R. A. (2010). An Oct4-centered protein interaction network in embryonic stem cells. *Cell Stem Cell* 6 (4): 369–381.
- van den Heuvel, S. and Dyson, N. J. (2008). Conserved functions of the pRB and E2F families. *Nat Rev Mol Cell Biol* 9 (9): 713–724.

- Vens, C., Rosso, M.-N., and Danchin, E. G. J. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* 27 (9): 1231–1238.
- von Bubnoff, A. and Cho, K. W. (2001). Intracellular BMP signaling regulation in vertebrates: pathway or network? *Dev Biol* 239 (1): 1–14.
- von Hippel, P. H. and Berg, O. G. (1986). On the specificity of DNA-protein interactions. *Proc Natl Acad Sci USA* 83 (6): 1608–1612.
- Waddington, C. H. (1942). Canalization of Development and the Inheritance of Acquired Characters. *Nature* 150 (3811): 563–565.
- Wang, H.-B., Liu, G., Zhang, H., Xing, S., Hu, L.-J., Zhao, W., Xie, B., Li, M., Zeng, B.-H., Li, Y., and Zeng, M.-S. (2013). Sp1 and c-Myc regulate transcription of BMI1 in nasopharyngeal carcinoma. *FEBS J* 280 (12): 2929–2944.
- Wang, X., McLachlan, J., Zamore, P. D., and Hall, T. M. T. (2002). Modular recognition of RNA by a human pumilio-homology domain. *Cell* 110 (4): 501–512.
- Wang, X., Juan, L., Lv, J., Wang, K., Sanford, J. R., and Liu, Y. (2011). Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1. *BMC Genomics* 12 Suppl 5: S8.
- Wang, Y., Gogol-Döring, A., Hu, H., Fröhler, S., Ma, Y., Jens, M., Maaskola, J., Murakawa, Y., Quedenau, C., Landthaler, M., Kalscheuer, V., Wieczorek, D., Wang, Y., Hu, Y., and Chen, W. (2013). Integrative analysis revealed the molecular mechanism underlying RBM10-mediated splicing regulation. *EMBO Mol Med* 5 (9): 1431–1442.
- Watanabe, T., Takeda, A., Tsukiyama, T., Mise, K., Okuno, T., Sasaki, H., Minami, N., and Imai, H. (2006). Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev* 20 (13): 1732–1743.
- Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171 (4356): 737–738.
- Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S., and Peterson, K. J. (2009). The deep evolution of metazoan microRNAs. *Evol Dev* 11 (1): 50–68.
- Wickens, M., Bernstein, D. S., Kimble, J., and Parker, R. (2002). A PUF family portrait: 3'UTR regulation as a way of life. *Trends Genet* 18 (3): 150–157.
- Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Statist.* 9 (1): 60–62.
- (1962). *Mathematical statistics*. Wiley New York: 644 p.
- Won, K.-J., Chepelev, I., Ren, B., and Wang, W. (2008). Prediction of regulatory elements in mammalian genomes using chromatin signatures. *BMC Bioinformatics* 9: 547.

- Wu, C.-I., Shen, Y., and Tang, T. (2009). Evolution under canalization and the dual roles of microRNAs: a hypothesis. *Genome Res* 19 (5): 734–743.
- Wu, J., Smith, L. T., Plass, C., and Huang, T. H.-M. (2006). ChIP-chip comes of age for genome-wide functional analysis. *Cancer Res* 66 (14): 6899–6902.
- Xi, S., Zhu, H., Xu, H., Schmidtmann, A., Geiman, T. M., and Muegge, K. (2007). Lsh controls Hox gene silencing during development. *Proc Natl Acad Sci USA* 104 (36): 14366–14371.
- Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434 (7031): 338–345.
- Xu, E. Y., Chang, R., Salmon, N. A., and Reijo Pera, R. A. (2007). A gene trap mutation of a murine homolog of the Drosophila stem cell factor Pumilio results in smaller testes but does not affect litter size or fertility. *Mol Reprod Dev* 74 (7): 912–921.
- Xu, X., Bieda, M., Jin, V. X., Rabinovich, A., Oberley, M. J., Green, R., and Farnham, P. J. (2007). A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. *Genome Res* 17 (11): 1550–1561.
- Yan, Q., Cho, E., Lockett, S., and Muegge, K. (2003). Association of Lsh, a regulator of DNA methylation, with pericentromeric heterochromatin is dependent on intact heterochromatin. *Mol Cell Biol* 23 (23): 8416–8428.
- Yi, F., Pereira, L., and Merrill, B. J. (2008). Tcf3 functions as a steady-state limiter of transcriptional programs of mouse embryonic stem cell self-renewal. *Stem Cells* 26 (8): 1951–1960.
- Ying, Q. L., Nichols, J., Chambers, I., and Smith, A. (2003). BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* 115 (3): 281–292.
- Yosefzon, Y., Koh, Y. Y., Chritton, J. J., Lande, A., Leibovich, L., Barziv, L., Petzold, C., Yakhini, Z., Mandel-Gutfreund, Y., Wickens, M., and Arava, Y. (2011). Divergent RNA binding specificity of yeast Puf2p. *RNA* 17 (8): 1479–1488.
- Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., and Majewski, I. J. (2011). ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res* 39 (17): 7415–7427.
- Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., Stewart, R., Slukvin, I. I., and Thomson, J. A. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318 (5858): 1917–1920.
- Zavolan, M., Kondo, S., Schonbach, C., Adachi, J., Hume, D. A., Hayashizaki, Y., Gaasterland, T., Group, R. I. K. E. N. G., and Members, G. S. L. (2003). Impact of alternative

- initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res* 13 (6B): 1290–1300.
- Zeng, Y., Yi, R., and Cullen, B. R. (2005). Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J* 24 (1): 138–148.
- Zhang, B., Gallegos, M., Puoti, A., Durkin, E., Fields, S., Kimble, J., and Wickens, M. P. (1997). A conserved RNA-binding protein that regulates sexual fates in the *C. elegans* hermaphrodite germ line. *Nature* 390 (6659): 477–484.
- Zhang, H., Kolb, F. A., Brondani, V., Billy, E., and Filipowicz, W. (2002). Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *EMBO J* 21 (21): 5875–5885.
- Zhang, M. Q. (1998). Statistical features of human exons and their flanking regions. *Hum Mol Genet* 7 (5): 919–932.
- Zhang, X., Zhang, J., Wang, T., Esteban, M. A., and Pei, D. (2008). Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. *J Biol Chem* 283 (51): 35825–35833.
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9 (9): R137.
- Ziebold, U., Lee, E. Y., Bronson, R. T., and Lees, J. A. (2003). E2F3 loss has opposing effects on different pRB-deficient tumors, resulting in suppression of pituitary tumors but metastasis of medullary thyroid carcinomas. *Mol Cell Biol* 23 (18): 6542–6552.





# Statutory Declaration

Hereby, I testify that this thesis is the result of my own work and research, except for the references given in the bibliography.

Jonas Maaskola

Berlin, June 2014



# Zusammenfassung

Die Mustersuche in Sequenzdaten ist ein Standardproblem der Bioinformatik, der Anwendung von Rechenmethoden in der Biologie. Klassische Verfahren der Mustersuche stützen sich vorrangig auf Methoden des maschinellen Lernens, die üblicherweise auf probabilistischer Modellierung der Sequenzen basieren. Die Einführung neuer Methoden zur Sequenzierung von DNS und RNS im Laufe des letzten Jahrzehnts sorgt für eine Flut von Daten, die motivieren, innovative Lösungen zur automatisierten Analyse zu entwickeln.

Die vorliegende Dissertation beschreibt eine Untersuchung diskriminativer Lernmethoden der Sequenzanalyse mit Anwendung zur Mustersuche in Nukleinsäuresequenzen. Der grundlegende Ansatz diskriminativer Verfahren zur Mustersuche besteht darin, solche Muster aufzuspüren, die in einem Satz von Sequenzen häufiger vorliegen als in einem anderen, oder deren Häufigkeit in mehreren Sequenzsätzen variiert. Es gibt vielfältige Maße zur Quantifizierung relativer Anreicherung solcher Art. Eine Anzahl von Publikationen beschreibt diskriminative Mustersuchmethoden, die sich nicht nur in der Wahl der Zielfunktion unterscheiden, sondern unter anderem auch in der Modellierung der Sequenzen, was es erschwert, die Nützlichkeit verschiedener Maße zu vergleichen.

Diese Dissertation bespricht klassische Verfahren der Sequenzanalyse und beschreibt darauf aufbauend eine flexible Methode zur Mustersuche, die die Wahl verschiedener Zielfunktionen zulässt. Die Leistungsfähigkeit der verschiedenen Zielfunktionen in der beschriebenen Methode und der anderer, bereits publizierter Methoden wird sorgfältig analysiert mit Hilfe von umfassenden, synthetisch erzeugten Daten. Insbesondere erlaubt diese Auswertung auch den Vergleich der Vor- und Nachteile diskriminativer und nicht-diskriminativer Lernmethoden. Dabei stellt sich heraus, dass einige der in der vorliegend beschriebenen Methode implementierten diskriminativen Zielfunktionen wesentlich bessere Ergebnisse erzielen als bisher veröffentlichte Methoden. In der Fähigkeit Muster zu entdecken, sind einige der Zielfunktionen in der betrachteten Aufgabenstellung sehr nahe am theoretisch erreichbaren Optimum. Dies zeigt der Vergleich mit der Musterwiedererkennung, dem Bestimmen von Mustervorkommen, wenn das Muster bereits bekannt ist. Unter diesen Zielfunktionen sticht die gegenseitige Information (mutual information), ein Maß aus der Informationstheorie, heraus, da sie sich sowohl zur Optimierung probabilistischer, wie auch diskreter Sequenzmodelle eignet, sie die Analyse von Kontrasten mit mehr als zwei Bedingungen erlaubt, und sie außerdem geeignete Generalisierungen bietet um Modelle mehrerer Muster zu finden.

Schließlich wird die Nützlichkeit und realistische Anwendbarkeit der vorgestellten Methode unter Verwendung der gegenseitigen Information dargestellt. Zahlreiche publizierte, mit unterschiedlichen Technologien erzeugte, biologische Datensätze werden analysiert. Dies umfasst Daten einer Familie von RNS-bindenden Proteinen in verschiedenen Spezies sowie Daten von Transkriptionsfaktoren, die von zentraler Bedeutung für die Regulation embryonaler Stammzellen höherer Säugetiere sind. Neue Erkenntnisse ergeben sich für einen Alternative-Splicing-Faktor, für den Muster gefunden werden, die als Splicing-relevant bekannt sind, deren Bedeutung für den untersuchten Faktor allerdings bisher noch nicht vergleichbar gut belegt worden ist.