GENE REGULATORY ELEMENT PREDICTION WITH BAYESIAN NETWORKS

VIPIN NARANG

NATIONAL UNIVERSITY OF SINGAPORE

2008

GENE REGULATORY ELEMENT PREDICTION WITH BAYESIAN NETWORKS

VIPIN NARANG

(M.S. Research (Electrical Engineering) , I.I.T. Delhi)

(B. Tech. (Electrical Engineering), I.I.T. Delhi)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2008

ACKNOWLEDGEMENTS

I wish to sincerely thank my advisors Dr. Wing Kin Sung and Dr. Ankush Mittal. Dr. Sung's constant interest in this research and regular meetings and discussions with him have been very valuable. Many of the ideas in this thesis were generated and refined through these discussions. His concern in ensuring high quality of the work has led to many improvements in both the work and the presentation. He has been very generous in giving his time whenever I wanted and prompt in giving his reviews. He has always been very supportive throughout my PhD and tolerant towards my shortcomings.

Dr. Ankush introduced and guided me in the subjects of Bayesian networks and bioinformatics and helped me to to obtain the research direction early on. He extended himself just as an elder brother to share with me his experience in conducting research and in dealing with the research environment and helped me through many difficult times. Several meetings and regular communications with him and his own example were helpful in giving focus and direction to this work. Without his help none of the publications from this work would have been possible.

I owe my deepest gratitude to Dr. Krishnan V. Pagalthivarthi, my most well wishing teacher and guide, who took the entire responsibility and personal difficulties for training me and guiding me throughout my research career. I had neither any clue nor capacity to pursue graduate studies. Since my B. Tech. days, enormous amounts of his time and effort have gone into cultivating me as a sincere student and taking me through every single step. His personal concern prior to and throughout this thesis work has made it materialize. His example as a very dedicated and caring teacher has left a deep impression on me. I am also indebted to him for giving me a meaningful purpose and vision for using this doctoral study.

I am grateful to my friend Sujoy Roy for being a great support and well wisher althroughout my stay at NUS. He is a very sincere student and I have benefitted in many ways from his association. He always extended himself in times of need and also gave valuable suggestions for the improvement of this thesis. I also wish to thank my friends Akshay, Amit Kumar, Sumeet, Anjan, Pankaj, Girish, Ganesh, Kalyan and others who have helped and supported me here.

Thought provoking discussions with my colleague Rajesh Chowdhary on Bayesian networks and gene regulation were valuable in deepening my understanding of these subjects.

I sincerely thank my parents, my elder brother Nitin, and my Masters thesis advisor Prof. M. Gopal for their sacrifices to support me and encouraging my pursuit of graduate studies.

Vipin Narang

TABLE OF CONTENTS

ACKNOWLEDGEMENTS III

TABLE OF CONTENTS V

SUMMARY VII

LIST OF TABLES IX

LIST OF FIGURES XI

LIST OF SYMBOLS XIX

LIST OF ACRONYMS XXI

PUBLICATIONS XXIII

CHAPTER - I 1

INTRODUCTION 1

I-1 BACKGROUND 1
I-2 MOTIVATION FOR PRESENT RESEARCH 9
I-3 NATURE OF THE PROBLEM 16
I-4 RESEARCH OBJECTIVES 21
I-5 ORGANIZATION OF THE THESIS 28

CHAPTER - II 29

LITERATURE REVIEW 29

II-1 DETECTION OF DNA MOTIFS 29
II-2 GENERAL PROMOTER MODELING AND TRANSCRIPTION START SITE PREDICTION 33
II-3 MODELING AND DETECTION OF CIS-REGULATORY MODULES 35

CHAPTER - III 39

PRELIMINARIES 39

III-1 STOCHASTIC MODEL OF THE GENOME 39
III-2 COMPUTATIONAL MODELING OF PROTEIN-DNA BINDING SITES (MOTIFS) 42
III-3 BAYESIAN NETWORKS 46
III-4 MEASURES OF ACCURACY 51

CHAPTER - IV 55

DETECTION OF LOCALIZED MOTIFS 55

IV-1 PROBLEM DEFINITION 56
IV-2 SCORING FUNCTION 57
IV-3 COMBINED SCORE 62
IV-4 ALGORITHM 63
IV-5 IMPLEMENTATION 67

IV-6 RESULTS 68
IV-6.1 Analysis of the scoring function 68
IV-6.2 Performance on Simulated datasets 71
IV-6.3 Performance on Real datasets 75
IV-7 CONCLUSIONS 81

CHAPTER - V 83

GENERAL PROMOTER PREDICTION 83

V-1 INTRODUCTION 83
V-2 STRUCTURE OF HUMAN PROMOTERS 85
V-3 OLIGONUCLEOTIDE POSITIONAL DENSITY 88
V-4 BAYESIAN NETWORK MODEL FOR GENERAL PROMOTER PREDICTION 91
V-4.1 The Promoter Model 91
V-4.2 Naïve Bayes Classifier Representation 94
V-4.3 Modeling and Estimation of Positional Densities 95
V-5 INFERENCE OVER LONG GENOMIC SEQUENCES 98
V-6 IMPLEMENTATION 100
V-7 RESULTS 101
V-7.1 Prominent Features Correspond to Well-Known Transcription Factor Binding Motifs 101
V-7.2 Results of TSS Prediction 102
V-8 CONCLUSIONS 110

CHAPTER - VI 113

CIS-REGULATORY MODULE PREDICTION 113

VI-1 MODULEXPLORER CRM MODEL 114
VI-2 DATA 116
VI-3 METHODS 119
VI-4 TRAINING OF MODULEXPLORER 130
VI-5 PAIRWISE TF-TF INTERACTIONS LEARNT DE-NOVO BY THE MODULEXPLORER 132
VI-6 GENOME WIDE SCAN FOR NOVEL CRMS 137
VI-7 FEATURE BASED CLUSTERING OF CRMS 143
VI-8 IMPLICATIONS OF MODULEXPLORER 161

CHAPTER - VII 163

CONCLUSIONS AND FUTURE WORK 163

APPENDIX 179

SUPPLEMENTARY FIGURES 189

REFERENCES 207

SUMMARY

While computational advances have enabled sequencing of genomes at a rapid rate, annotation of functional elements in genomic sequences is lagging far behind. Of particular importance is the identification of sequences that regulate gene expression. This research contributes to the computational modeling and detection of three very important regulatory elements in eukaryotic genomes, *viz.* transcription factor binding motifs, gene promoters and cis-regulatory modules (enhancers or repressors). Position specificity of transcription factor binding sites is the main insight used to enhance the modeling and detection performance in all three applications.

The first application concerns *in-silico* discovery of transcription factor binding motifs in a set of regulatory sequences which are bound by the same transcription factor. The problem of motif discovery in higher eukaryotes is much more complex than in lower organisms for several reasons, one of which is increasing length of the regulatory region. In many cases it is not possible to narrow down the exact location of the motif, so a region of length ~1kb or more needs to be analyzed. In such long sequences, the motif appears “subtle” or weak in comparison with random patterns and thus becomes inaccessible to any motif finding algorithm. Subdividing the sequences into shorter fragments poses difficulties such as choice of fragment location and length, locally over-represented spurious motifs, and problems associated with compilation and ranking of the results. A novel tool, LocalMotif, is developed in this research to detect biological motifs in long regulatory sequences aligned relative to an anchoring point such as the transcription start site or the center of the ChIP sequences. A new scoring measure called spatial confinement score is developed to accurately demarcate the interval of localization of a motif. Existing scoring measures including over-representation score and relative entropy score are reformulated within the framework of information theory and combined with spatial confinement score to give an overall measure of the goodness of a motif. A fast algorithm finds the best localized motifs using the scoring function. The approach is found useful in detecting biologically relevant motifs in long regulatory sequences. This is illustrated with various examples.

Computational prediction of eukaryotic promoters is another tough problem, with the current best methods reporting less than 35% sensitivity and 60% ppv[1]. A novel statistical modeling and detection framework is developed in this dissertation for

[1] Transcription start site prediction accuracy on ENCODE regions of the human genome within ±250 bp error [Bajic et al. (2006)].

promoter sequences. A number of exisiting techniques analyze the occurrence frequencies of oligonucleotides in promoter sequences as compared to other genomic regions. In contrast, the present approach studies the positional densities of oligonucleotides in promoter sequences. A statistical promoter model is developed based on the oligonucleotide positional densities. When trained on a dataset of known promoter sequences, the model automatically recognizes a number of transcription factor binding sites simultaneously with their occurrence positions relative to the transcription start site (TSS). The analysis does not require any non-promoter sequence dataset or modeling of background oligonucleotide content of the genome. Based on this model, a continuous naïve Bayes classifier is developed for the detection of human promoters and transcription start sites in genomic sequences. Promoter sequence features learnt by the model correlate well with known biological facts. Results of human TSS prediction compare favorably with existing 2nd generation promoter prediction tools.

Computational prediction of cis-regulatory modules (CRM) in genomic sequences has received considerable attention recently. CRMs are enhancers or repressors that control the expression of genes in a particular tissue at a particular development stage. CRMs are more difficult to study than promoters as they may be located anywhere up to several kilo bases upstream or downstream of the gene's TSS and lack anchoring features such as the TATA box. The current method of CRM prediction relies on discovering clusters of binding sites for a set of cooperating transcription factors (TFs). The set of cooperating TFs is called the regulatory code. So far very few (precisely three) regulatory codes are known which have been determined based on tedious wet lab experiments. This has restricted the scope of CRM prediction to the few known module types. The present research develops the first computational approach to learn regulatory codes de-novo from a repository of CRMs. A probabilistic graphical model is used to derive the regulatory codes. The model is also used to predict novel CRMs. Using a training data of 356 non-redundant CRMs, 813 novel CRMs have been recovered from the Drosophila melanogaster genome regulating gene expression in different tissues at various stages of development. Specific regulatory codes are derived conferring gene expression in the drosophila embryonic mesoderm, the ventral nerve cord, the eye-antennal disc and the larval wing imaginal disc. Furthermore, 31 novel genes are implicated in the development of these tissues.

LIST OF TABLES

Table IV-1. Results of using LocalMotif to analyze simulated sequences of length 3000 bp containing a planted (7,1) motif ATGCATG – five top scoring motifs and their predicted localization intervals are reported. 69

Table IV-2. Ranges of parameters studied in simulated short sequence datasets. 71

Table IV-3. Accuracy of motif detection in synthetic long sequence datasets. 75

Table V-1. Results of cross-validation studies in the training of BayesProm. The complete dataset of 1796 human promoter sequences was randomly divided into 1436 training sequences (80%) and 360 validation sequences (20%). Five such uncorrelated cross-validation sets were generated. A negative set of 5000 human exon and 3' UTR sequences obtained from Genbank was used simultaneously for testing. 101

Table VI-1. Overlap of novel CRMs predicted by Modulexplorer with CRMs predicted in previous computational studies. 140

Table VI-2. Clusters of CRMs sharing a common regulatory code (motifs) obtained using iterative frequent itemset mining. Five major clusters are listed with their (i) predominant tissue and stage of expression, (ii) number of known and predicted CRM target genes, (iii) number of predicted CRM target genes with validation, (iv) number of validated genes which are novel for their role in development, and (v) false positive rate of the regulatory code on other training CRMs and random background sequences. 145

LIST OF FIGURES

Figure I-1. Annotated DNA sequence of the 5’ region of the human PAX3 gene [Macina et al. (1995), Okladnova et al. (1999), Barber et al. (1999)]. Notable features shown include (i) promoter region, (ii) transcription start site, (iii) transcription factor binding sites such as TATA box, CAAT box, AP-1, AP-2, SP1, (iv) repressor element, (v) nucleotide repeats, (vi) 5’ untranslated region (UTR), (vii) coding sequence with its amino acid translations, (viii) exon, (ix) intron, and (x) splice site. 3

Figure I-2. The locations of gene coding and noncoding regions and the promoter in a DNA strand. The promoter region is present surrounding the start of (and mostly upstream of) the transcript region. Other elements such as enhancer may be present far distant from the transcription start site..........4

Figure I-3. Formation of pre-initiation complex through the binding of transcription factors to DNA nearby the transcription start site [Pederson et al. (1999)]..........6

Figure I-4. Several genomic features are currently being computationally annotated in the human genome in the ENCODE project. The present research focuses on three features in the regulatory sequence track: transcription start sites, transcription factor binding sites (motifs) and enhancers (cis-regulatory modules). 10

Figure I-5. The “Genomes to Life” program of the U.S. Department of Energy [Frazier et al. (2003)] plans for the next 10 years to use DNA sequences from microbes and higher organisms, including humans, as starting points for systematically tackling questions about the essential processes of living systems. Advanced technological and computational resources will help to identify and understand the underlying mechanisms that enable organisms to develop, survive, carry out their normal functions, and reproduce under myriad environmental conditions..........11

Figure I-6. Applications of the present research in current bioinformatics context..........12

Figure I-7. Transcription factor binding motifs, promoters and CRMs are all associated with a notion of position specificity. 15

Figure I-8. Discovering (6,1) motifs within a set of N sequences $S_1, S_2, \ldots, S_N$ of length L. In (a) the random pattern TTTAAA is seen to eclipse the real motif TTGACA when the complete sequence is analyzed, but in (b) the real motif TTGACA becomes dominant when only the local interval (p_1,p_2) is considered. 23

Figure I-9. Difference between the distribution of binding sites of (a) a localized motif, and (b) a spurious motif. While both may appear over-represented in a local sequence interval, localized motifs have a prominent region of confinement within the entire sequence length. 23

Figure I-10. An illustration of the difficulties in analyzing sub-intervals of long regulatory sequences – for short intervals, motifs A and C are missed, and for long intervals the motifs may become weak. 24

Figure II-1. Computational models for cis-regulatory modules: (a) homotypic cluster of TFBS [Markstein et al. (2002)], (b) heterotypic cluster of TFBS [Berman et al. (2002)], (c) hidden Markov model [Frith et al. (2001)], (d) statistical model of Gupta and Liu (2005), (e) discriminatory Bayesian network model of Segal and Sharan (2005). 36

Figure III-1. Finite state machine visualization of a first order Markov model for sequence background. 41

Figure III-2. A small sample of binding sites for the transcription factor NF-Y. 43

Figure III-3. Single-letter IUPAC codes for representing degeneracy of nucleotides. 43

Figure III-4. Positional weight matrix developed from the collection of NF-Y TFBS in Figure III-2. 44

Figure III-5. A Bayesian network for modeling the causes of heart disease. 47

Figure III-6. Conditional probability table (CPT) for the node “obesity” in the Bayesian network of Figure III-5. 48

Figure III-7. The Receiver Operating Characteristics (ROC) curve. 53

Figure IV-1. Discovering (6,1) motifs within a set of N sequences $S_1, S_2, \ldots, S_N$ each of length L. The random pattern TTTAAA is seen to eclipse the real motif TTGACA. 57

Figure IV-2. Illustration of how spatial confinement score finds the shortest interval encompassing the maximum proportion of TFBS – though interval A has higher density of TFBS, its score is lower since a large proportion of TFBS still lie outside it. 62

Figure IV-3. The LocalMotif algorithm....64

Figure IV-4. Contours showing (a) the total score, (b) over-representation score, and (c) spatial confinement score of the motif ATGCATG in different position intervals (p_1,p_2) of the planted motif sequences....70

Figure IV-5. Performance of MEME, Weeder and Localmotif in simulated short sequence datasets with (a) varying sequence length, L, (b) varying percentage, k, of sequences containing motif instances. 73

Figure IV-6. Accuracy of LocalMotif's interval predictions. 73

Figure IV-7. Motifs discovered by MEME and LocalMotif in Drosophila promoters. 76

Figure IV-8. Variation of sensitivity and false positive rate of Localmotif's predictions in long regulatory sequences upstream of the TSS as the number of predicted motifs is increased. 79

Figure IV-9. Distribution of forkhead binding sites relative to ER binding sites. 80

Figure IV-10. Motifs discovered by MEME, Weeder and LocalMotif in ERE dataset. 80

Figure V-1. Positional densities of the TATA box and CAAT box binding sites in a set of 1796 promoter sequences obtained from the eukaryotic promoter database. 86

Figure V-2. An illustration of the positional density of the oligonucleotide TATAAA, obtained using 1796 human promoter sequences in EPD. The TSS is located at position 0. The curve indicates the probability of observing the oligonucleotide TATAAA at various positions upstream and downstream of the TSS. 90

Figure V-3. (a) Relationship between positional density definition and training promoter sequences, (b) modeling a nucleotide sequence, S, for promoter inference (Equation 5.4)....92

Figure V-4. The naïve Bayes classifier for promoter prediction. 95

Figure V-5. Using naïve Bayes classifier to detect promoter region and TSS in long genomic sequences. 99

Figure V-6. Important consensus sequences recognized by the naïve Bayes model....103

Figure V-7. ROC curve showing the TSS prediction performance of BayesProm and Eponine on Genbank dataset. In case A, TSS predictions within ±200 bp of the annotated TSS were considered correct, while in case B, this range was extended to ±1000 bp. Eponine is seen to be highly specific, while BayesProm has high sensitivity. 106

Figure V-8. Density of true predictions relative to the annotated TSS on Genbank dataset. Both Eponine and BayesProm report a histogram peak at zero distance, indicating the accuracy of these softwares. Eponine is seen to be highly specific but less sensitive, while BayesProm is moderately specific but highly sensitive...106

Figure V-9. Predictions of regulatory regions in the human β globin locus on chromosome 11 (Genbank accession no. U01317) using (a) Hidden Markov Model by Crowley et al. (1997), (b) BayesProm, showing only predictions above threshold of –10, and (c) Interpolated Markov Chain model by Ohler et al. (1999). It is observed that the HMM in (a) can only predict the locus control regions, while BayesProm accurately predicts five of the six transcription start sites with very few false positives.........................108

Figure V-10. ROC curve showing the evaluation of BayesProm and several 2nd generation promoter prediction tools on chromosome 22 dataset. The test criterion was same as that used by Scherf et al. (2001)..109

Figure VI-1. The Modulexplorer pipeline to learn a CRM model from a repository of uncharacterized CRMs and background sequences, and to use the model for predicting novel CRMs is shown in (a). Also shown are the validations that have been conducted in this study to verify the model and the novel CRMs predicted by the model..114

Figure VI-2. The Modulexplorer Bayesian network model. The model describes a CRM as a cluster of multiple interacting TFBS with distance and order constraints. The nodes D_i are the dyad motifs representing the TFBSs. They have states 0 or 1 according to whether the motif is absent or present in the CRM. The CRM is their common effect or hypothesis, represented as the child node. Each dyad motif D_i has two monad components $\left(M_{i1}, M_{i2}\right)$ with a spacer of 0 to 15 bp. These monads are represented by individual nodes M_{i1}, M_{i2} having states 0 or 1, *i.e.* present or absent, and are related to the dyad node D_i by a noisy-AND relationship. The spacer length (or distance), discretized as low or high, is modeled by the node d_i. Furthermore each D_i is associated with an *order* either left or right according to whether M_{i1} appears to the left or to the right of M_{i2} in the CRM. 115

Figure VI-3. (a) From a total of 619 experimental CRM sequences obtained from the REDfly database, 205 redundant CRMs were discarded, 58 long CRMs (>3.5 kbp) were used as a testing set and remaining 356 form the training set. The length distribution of the 356 training CRMs is shown in (b). Most CRMs are between 200 to 1200 bp long with 1040 bp as the median length. The functional diversity among the training CRMs is shown in (c) and (d). Out of the 356 training CRMs, 302 are expressed in the embryo stage, 193 in the larva stage, and 86 in the adult fly. Among the 302 CRMs expressed in the embryo, 87 are expressed in the blastoderm stage (stages 3-5) and 205 in the post-blastoderm stages (stages 6 to 16). Categorization of the 205 post-blastoderm CRMs in terms of the developing organ system where they express is shown in (d). The integumentary system (ectoderm), imaginal precursor (wing disc, retinal disc etc.), nervous system, digestive system (abdomen) and muscle system are over-represented classes among the known CRMs 118

Figure VI-4. Drosophila CRMs have high redundancy of transcription factor binding sites. The number of binding sites per transcription factor in a CRM is shown in (a) for 19 CRMs having full experimental TFBS annotation (average 5.4 binding sites per TF) and 136 partially annotated CRMs (average 3.6 binding sites per TF). The fluffy tail test (FTT) scores [Abnizova et al. (2005)] for these sequences are shown in (b). The sequences were repeatmasked before computing the FTT to eliminate tandem repeats that may erroneously cause a high FTT value. FTT scores of most CRMs are greater than 2.0, indicating significant redundancy. The FTT scores of fully and partially annotated CRMs are similar, indicating that partially annotated CRMs may have greater redundancy than observed in the partial annotation. The full annotation of 19 CRMs is shown in (c)..................120

Figure VI-5. Over the next three pages, the figure illustrates the novel procedure used in Modulexplorer for characterizing TFBSs de-novo in a CRM...122

Figure VI-6. Potentials $\Pr\left(D_i \middle| M_{i,1}, M_{i,2}\right)$ factorized using the hidden nodes B_i. 127

Figure VI-7. The TFBSs in Drosophila CRMs appear as repeated or redundant sites. Modulexplorer locates these redundant sites as potential TFBSs. The receiver-operating characteristic of predicting TFBSs using redundant sites in 19 fully annotated CRMs is shown in (a). Here sensitivity (y-axis) refers to the % of nucleotides in TFBSs that are overlapped by some redundant site, while false positive rate (x-axis) refers to the % of nucleotides in a redundant site that do not match any TFBS. The maximum effectiveness of TFBS characterization in each of the 19 CRMs is shown in (b), which is the point in the ROC curve where Matthew's correlation coefficient is maximized. At this maximum effectiveness, the visual overlap between the TFBS sites (blue boxes) and the redundant sites (red boxes) in each CRM is shown in (c)...131

Figure VI-8. Performance of the Modulexplorer in discriminating between CRM and background sequences. Modulexplorer's performance is compared with two other methods: a Markov model (orders 2 to 6) and the HexDiff algorithm [Chan and Kibler (2005)]. The original Hexdiff algorithmuses (6,0) motifs, but it was extended in this comparison to try several different (l,d) motifs. Discrimination achieved between training CRMs and exon sequences in 10-fold cross-validation is shown in (a). The ROC shows that all three methods could easily discriminate CRMs from exons. Discrimination between CRMs and non-coding sequences (intron+intergenic) is shown in (b). Here Markov model shows no discrimination, HexDiff has marginal discrimination, while Modulexplorer achieves maximum discrimination. Modulexplorer was further evaluated on a separate testing set of 58 CRMs. The number of CRMs of different types in the test set according to their stage and tissue of expression is shown in (c). The performance of Modulexplorer on this test set, shown in (d), is similar to the training performance. 133

Figure VI-9 Dyad motifs in Modulexplorer most closely resembling the binding sites of known TFs. 134

Figure VI-10 Pairwise interactions between 61 different TFs learnt de-novo by the Modulexplorer probability model. Based on the interaction matrix, the TFs were hierarchically clustered. Six functionally related groups of TFs were formed: (1) cofactors of twist in mesoderm and nervous system development, (2) TFs involved in imaginal disc development, (3) the antennapedia complex, (4) TFs expressed in the blastoderm, (5) TFs for eye development and (6) a miscellaneous set of TFs. Five distinct clusters are seen in the interaction matrix. Three of the clusters contain mixed set of TFs from groups 1-4, while two other clusters correspond to the TF groups 5 and 6. 135

Figure VI-11. Summary of Modulexplorer's whole genome CRM predictions: (a) A stringent score threshold was used for shortlisting predicted CRM windows such that the false positive rate is about 0.1%. (b) A total of 1298 windows were predicted above the chosen threshold, out of which 813 are novel predictions. (c) The predicted CRMs are significantly over-represented in the promoter and upstream intergenic regions. (d) This is the list of level 3 gene ontology (GO) categories statistically over-represented in the target genes of the predicted CRMs. They show enrichment in development and regulatory functions (Bonferroni corrected P-values of the GO associations are shown alongside). 139

Figure VI-12. The 619 known REDfly CRMs, the 813 CRM windows predicted by Modulexplorer and a set of 813 randomly distributed segments were analyzed for their clustering around genes. A 50 kb long sliding window was scanned over the genome. The number of windows which contained one or more CRMs or random segments is shown below. The histogram shows the number of CRMs or random segments in the window on *x*-axis and the number of such windows on *y*-axis. The known and predicted CRMs come across in clusters of 3 to 4 CRMs in a window, whereas the randomly distributed segments are not usually clustered..142

Figure VI-13. The GC content of the predicted CRMs is similar to that of the known CRMs and higher in general compared to intron and intergenic sequences. 142

Figure VI-14. Cluster of CRMs controlling target gene expression in the embryonic mesoderm, and their regulatory code....................................146

Figure VI-15. BDGP in-situ expression images for the target genes of novel CRMs in the mesoderm cluster..147

Figure VI-16. Matches of the mesoderm regulatory code motifs within the dpp 813 bp enhancer are shown by underlines. For comparison the known TFBS in this enhancer, available only for the first 600 bp, are shown in red color text. Out of 32 matches of the regulatory code motifs in first 600 bp, 26 overlapped known TFBS. 148

Figure VI-17. Cluster of CRMs controlling target gene expression in the embryonic ventral nerve cord, and their regulatory code. 151

Figure VI-18. BDGP in-situ expression images for the target genes of novel CRMs in the ventral nerve cord cluster. 152

Figure VI-19. Cluster of CRMs controlling target gene expression in the embryonic eye-antennal disc, and their regulatory code. 154

Figure VI-20. BDGP in-situ expression images for the target genes of novel CRMs in the eye-antennal disc cluster....155

Figure VI-21. List of novel CRMs separated from the AT-rich clusters which control target gene expression in the blastoderm embryo. 157

Figure VI-22. BDGP in-situ expression images for the target genes of novel CRMs in the blastoderm cluster....158

Figure VI-23. Binding sites for 10 blastoderm TFs were searched in the region -5000 to +5000 around the 98 predicted blastoderm CRMs. The CRMs are in the location 0 to 1000. In the CRM region the binding sites were over-represented by a factor of around 2. The y-axis shows the total number of binding sites found in the window in all 98 CRMs. 159

LIST OF SYMBOLS

A	Anchor point / Alignment score
b	A nucleotide base ($b \in \{A, C, G, T\}$)
B	Background model
c	Binding site concentration within a position interval
$^{n}C_{k}$	Number of combinations = $n!/(k!(n-k)!)$
d	Number of allowed mismatches in a motif
$D(.)$	Kullback-Leibler distance
e	Number of expected occurrences / Estimated proportion
$E[.]$	Expectation operator
$f(.)$	Probability density function
f	Frequency
G	Number of components in a Gaussian mixture
H	A hypothesis
i, j, k	Indices
I	Position interval
K	An oligonucleotide of length l
l	Length of a motif
L	Length of a sequence
$L(.)$	Likelihood function
n	Number of instances, occurrences or counts
N	Number of sequences
M	A motif
p	Position or probability

Pr(.)	Probability
q	Order of the Markov model
S	A nucleotide (DNA) sequence
s	Step size (refer Section IV-4.2), or an index over
X	A general random variable
w	Weights in a PWM
Z	Z-score
α	Mixing proportion of a component in Gaussian mixture
λ	Likelihood ratio test statistic
$\pi, \bar{\pi}$	Promoter, Non-promoter
μ	Mean of a Gaussian density
σ	Variance of a Gaussian density
θ	Set of parameters of a probability model
ϕ	Gaussian density (pdf)

IUPAC codes for degenerate nucleic acids

A - adenosine
C - cytidine
G - guanine
T - thymidine
U - uridine
R - G A (purine)
Y - T C (pyrimidine)
K - G T (keto)

M - A C (amino)
S - G C (strong)
W - A T (weak)
B - G T C
D - G A T
H - A C T
V - G C A
N - A G C T (any)

LIST OF ACRONYMS

AIC	Akaike Information Criterion
BLAST	Basic Local Alignment Search Tool
CC	Cross-correlation Coefficient
cDNA	Complementary DNA
CPT	Conditional Probability Table
CRM	Cis-Regulatory Module
DAG	Directed Acyclic Graph
DCRD	Drosophila Cis-Regulatory Database
DNA	Deoxyribonucleic acid
EM	Expectation Maximization algorithm
EPD	Eukaryotic Promoter Database
FN	False Negative
FP	False Positive
GO	Gene Ontology
HMM	Hidden Markov Model
IUPAC	International Union of Pure and Applied Chemistry
KL	Kullback-Leibler distance
MEME	Multiple EM for Motif Elicitation [Bailey et al. (1994)]
Npv	Negative Predictive Value
ORS	Over-representation Score
pdf	Probability density function
Ppv	Positive Predictive Value
PWM	Positional Weight Matrix

RES	Relative Entropy Score
ROC	Receiver Operating Characterisitics
RR	Rejection Region
SCS	Spatial Confinement Score
Se	Sensitivity
Sp	Specificity
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TN	True Negative
TP	True Positive
TSS	Transcription Start Site

PUBLICATIONS

The following papers have been published / submitted from this research thesis:

1. Narang, V., Sung, W.K., and Mittal, A. (2005). "Computational modeling of oligonucleotide positional densities for human promoter prediction." *Artificial Intelligence in Medicine*, **35**(1-2), 107-119.

2 Narang, V., Mittal, A., Sung, W.K. (2005). "Discovering weak motifs through binding site distribution analysis." *12th International Conference on Biomedical Engineering (ICBME 2005), Singapore, December 7-10, 2005*.

3. Narang, V., Sung, W.K., and Mittal, A. (2006). "Bayesian network modeling of transcription factor binding sites." in: *Bayesian Network Technologies: Applications and Graphical Models*, A. Mittal and A. Kassim, eds., Idea Group Publishing, Pennsylvania, USA.

4. Narang, V., Sung, W.K., and Mittal, A. "LocalMotif - an in silico tool for detecting localized motifs in regulatory sequences." *18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2006), Washington D.C.,USA, November 13-15, 2006*, 791-799.

5. Narang, V., Sung, W.K., and Mittal, A. (2006). "Computational annotation of transcription factor binding sites in D. melanogaster developmental genes." *Genome Informatics*, **17**(2), 14-24.

6. Narang, V., Sung, W.K., and Mittal, A. (2007). "Localized motif discovery in metazoan regulatory sequences." *Under submission*.

7. Narang, V., Mittal, A., and Sung, W.K. (2008). "Probabilistic Graphical Modeling of Cis-Regulatory Codes Governing Drosophila Development," *Under submission*.

CHAPTER - I

INTRODUCTION

I-1 Background

Over the last few years, computational biology research has contributed significantly to the advancement of molecular biology. High throughput genome sequencing has provided us with the complete genomes of several multicellular species from microbes to human beings. The current significant challenge is to annotate functional elements in these genomes and to understand how the vast amount of information contained in the genome is processed in living systems. One of the ultimate aims is to understand the process of development, i.e. how a living organism grows from a single cell to an adult, and how cells which are identical in the beginning differentiate into different tissues. This dissertation addresses some of these problems. First a brief description of some basic concepts of molecular biology is provided in this section to establish a ground for introducing the present research problem.

I-1.1 The Genetic Code

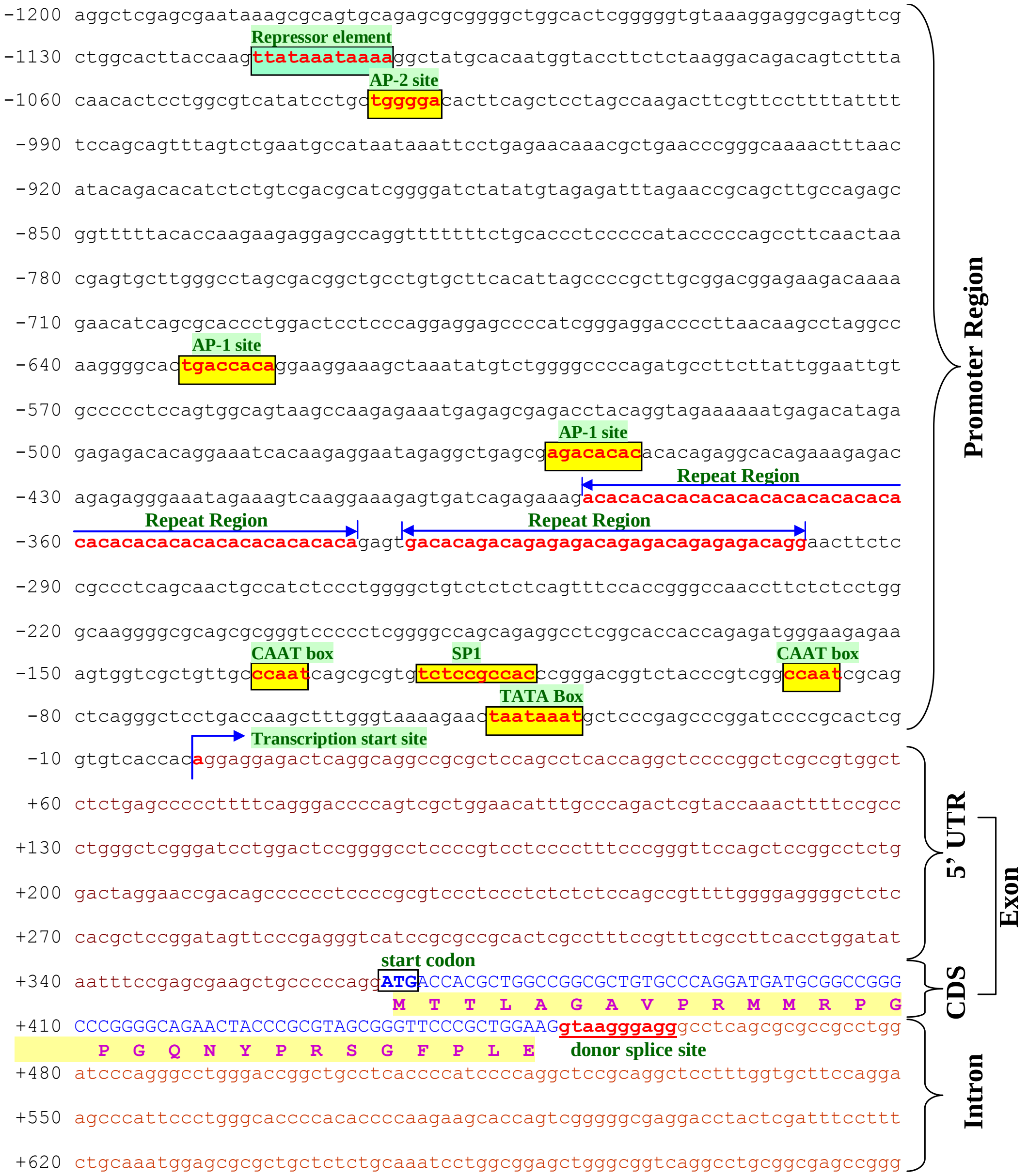
Every living organism's body is made up of microscopic units called cells. Majority of cellular structures are manufactured from proteins, which are complex macromolecules of amino acids. Most of the activities within a cell are also carried out by specific proteins. Each cell contains within its nucleus all the instructions needed to manufacture (or express) all of these proteins in the form of genetic code. In addition, the mechanism to express a protein at the exact time and location (e.g. during development) or whenever needed by the cell is also programmed within the genetic code.

The genetic code exists in the form of very long macromolecular chains called DNA (deoxyribonucleic acid). DNA is composed of four nitrogenous bases viz. Adenine, Cytosine, Guanine, and Thymine (in short A, C, G and T), which are covalently bonded to a backbone of deoxyribose-phosphate to form a DNA strand. Two complementary strands pair up to form a double helical structure where Gs pair with Cs and As with Ts. The two strands are held together by hydrogen bonding between the bases, forming base pairs (bp). The specific ordering of the four bases is responsible for the information content of the DNA. An organism's complete set of DNA is called its *genome*. Genomes vary widely in size. The human genome is approximaltely 3 billion bp long.

A *gene* is a portion of the genome which encodes the amino acid sequence of a protein product. Only a small fraction of the genome is covered by *genes*. The human genome is estimated to contain 30,000 to 40,000 genes. The gene DNA sequence maps to the protein amino acid sequence through the genetic code. In the genetic code each triplet of nucleotides (called ‘codon’) maps to a particular single amino acid. A protein encoding segment is a sequence of codons called coding sequence (CDS) or *exon*. An example of a gene region within the human genome is shown in Figure I-1. The coding sequence is marked in blue color with the encoded amino acids shown below it. Figure I-1 also shows a number of other features in the gene apart from the coding sequences. These include introns, untranslated region (UTR), promoter, etc., which are described in the following section. A block-diagram of the gene region shown in Figure I-1 is provided in Figure I-2 in order to illustrate the functional divisions of the gene region.

I-1.2 Gene Expression

The process of manufacturing proteins from the genetic code in DNA is called *gene expression*. This process is described by the central dogma of molecular biology, which states that the genetic code is utilized to manufacture the encoded protein within



```
-1200 aggctcgagcgaataaagcgcagtgcagagcgcggggctggcactcggggggtgtaaaggaggcgagttcg
                    Repressor element
-1130 ctggcacttaccaagttataaataaaaggctatgcacaatggtaccttctctaaggacagacagtcttta
                                      AP-2 site
-1060 caacactcctggcgtcatatcctgctggggacacttcagctcctagccaagacttcgttccttttatttt
 -990 tccagcagtttagtctgaatgccataataaattcctgagaacaaacgctgaacccgggcaaaactttaac
 -920 atacagacacatctctgtcgacgcatcggggatctatatgtagagatttagaaccgcagcttgccagagc
 -850 ggtttttacaccaagaagaggagccaggtttttttctgcaccctcccccataccccccagccttcaactaa
 -780 cgagtgcttgggcctagcgacggctgcctgtgcttcacattagccccgcttgcggacggagaagacaaaa
 -710 gaacatcagcgcaccctggactcctcccaggaggagccccatcgggaggaccccttaacaagcctaggcc
             AP-1 site
 -640 aaggggcactgaccacaggaaggaaagctaaatatgtctggggccccagatgccttcttattggaattgt
 -570 gccccctccagtggcagtaagccaagagaaatgagagcgagacctacaggtagaaaaaatgagacatagа
                                                  AP-1 site
 -500 gagagacacaggaaatcacaagaggaatagaggctgagcgagacacacacacagaggcacagaaagagac
                                                           Repeat Region
 -430 agagagggaaatagaaagtcaaggaaagagtgatcagagaaagacacacacacacacacacacacacaca
           Repeat Region                        Repeat Region
 -360 cacacacacacacacacacacacagagtgacacagacagagagacagagacagagagacaggaacttctc
 -290 cgccctcagcaactgccatctccctggggctgtctctctcagtttccaccgggccaaccttctctcctgg
 -220 gcaaggggcgcagcgcgggtcccccctcggggccagcagaggcctcggcaccaccagagatgggaagagaa
                    CAAT box          SP1                        CAAT box
 -150 agtggtcgctgttgcccaatcagcgcgtgtctccgccacccgggacggtctacccgtcggccaatcgcag
                                             TATA Box
  -80 ctcagggctcctgaccaagctttgggtaaaagaactaataaatgctcccgagcccggatccccgcactcg
                    Transcription start site
  -10 gtgtcaccacaggaggagactcaggcaggccgcgctccagcctcaccaggctccccggctcgccgtggct
  +60 ctctgagccccctttcagggaccccagtcgctggaacatttgcccagactcgtaccaaacttttccgcc
 +130 ctgggctcgggatcctggactccggggcctccccgtcctcccctttcccgggttccagctccggcctctg
 +200 gactaggaaccgacagccccccctccccgcgtccctccctctctctccagccgttttggggagggggctctc
 +270 cacgctccggatagttcccgagggtcatccgcgccgcactcgcctttccgtttcgccttcacctggatat
                                  start codon
 +340 aatttccgagcgaagctgccccccaggATGACCACGCTGGCCGGCGCTGTGCCCAGGATGATGCGGCCGGG
                                   M  T  T  L  A  G  A  V  P  R  M  M  R  P  G
 +410 CCCGGGGCAGAACTACCCGCGTAGCGGGTTCCCGCTGGAAGgtaagggagggcctcagcgcgccgcctgg
          P  G  Q  N  Y  P  R  S  G  F  P  L  E     donor splice site
 +480 atcccagggcctgggaccggctgcctcaccccatccccaggctccgcaggctcctttggtgcttccagga
 +550 agcccattccctgggcaccccacaccccaagaagcaccagtcgggggcgaggacctactcgatttccttt
 +620 ctgcaaatggagcgcgctgctctctgcaaatcctggcggagctgggcggtcaggcctgcggcgagccggg
```

The nucleotides are color coded as follows:

```
Black  ( ACGTacgt ) : 5' regulatory sequence
Brown  ( ACGTacgt ) : 5' untranslated region (UTR) in the first exon
Blue   ( ACGTacgt ) : Coding sequence (CDS) in the exon
Orange ( ACGTacgt ) : Intron sequence
Red    ( ACGTacgt ) : Some specific feature - marked with a comment
```

Figure I-1. Annotated DNA sequence of the 5' region of the human PAX3 gene [Macina et al. (1995), Okladnova et al. (1999), Barber et al. (1999)]. Notable features shown include (i) promoter region, (ii) transcription start site, (iii) transcription factor binding sites such as TATA box, CAAT box, AP-1, AP-2, SP1, (iv) repressor element, (v) nucleotide repeats, (vi) 5' untranslated region (UTR), (vii) coding sequence with its amino acid translations, (viii) exon, (ix) intron, and (x) splice site.

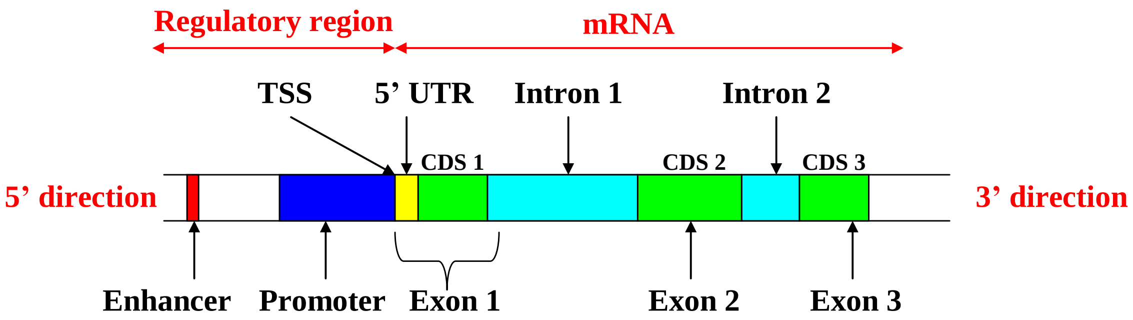


Figure I-2. The locations of gene coding and noncoding regions and the promoter in a DNA strand. The promoter region is present surrounding the start of (and mostly upstream of) the transcript region. Other elements such as enhancer may be present far distant from the transcription start site.

the cell in two steps – (i) *transcription*, or creating a copy of the gene in the form of a RNA molecule, and (ii) *translation*, or decoding the RNA to amino acid sequence through the genetic code. The transcription step is required because the genetic material is physically separated from the site of protein synthesis in the cytoplasm in the cell. The DNA is not directly translated into protein, but an intermediary molecule called RNA is made, which is an exact copy of the DNA. The RNA moves out of the nucleus into the cytoplasm, where it is translated by ribosomes to manufacture the protein.

In eukaryotes, the protein coding genes are transcribed by the RNA-polymerase II enzyme. Transcription initiates at a specific base pair location, called the transcription start site (TSS), as shown in Figure I-1 and Figure I-2. The portion of the gene downstream of the TSS (i.e., in the 3' direction) is transcribed to form the messenger RNA (mRNA). As shown in Figure I-2, in the transcribed sequence (both DNA and mRNA), the coding sequence (CDS) does not exist as a single continuous sequence but is interspersed with gaps called introns. Introns are removed or spliced from the mRNA before the translation step. This is called RNA splicing. The first codon is also often preceded by an untranslated region (5' UTR), whose function is to lend stability to the mRNA. The base positions on the gene are indexed relative to the TSS, which is referred to as position +1. Positions downstream (in 3' direction) of the TSS have a positive index while those upstream (in 5' direction) have a negative index. Figure I-1 shows the -1200 to +700 bp region of the human paired-box gene 3 (PAX3).

I-1.3 Regulation of Gene Expression

The control or regulation of gene expression dictates when, where (in what tissue(s)) and how much quantity of a particular protein is produced. This decides the development of cells and their responses to external stimuli. The detailed working of this control mechanism is still unknown to us. The most important mechanism of control is through regulating the transcription process, i.e. whether or not the transcription of a gene is initiated. In eukaryotic cells, the RNA-polymerase II is incapable of initiating transcription on its own. It does so with the assistance of a number of proteins called transcription factors (TFs). TFs bind to the DNA sequence and interact to form a pre-initiation complex (PIC) as shown in Figure I-3. The RNA-polymerase II is recruited in the PIC, and thus transcription begins. Thus the crucial point of the regulation mechanism is binding of TFs to DNA. Disruptions in gene regulation are often linked to a failure of the TF binding, either due to mutation of the DNA binding site, or due to mutation of the TF itself.

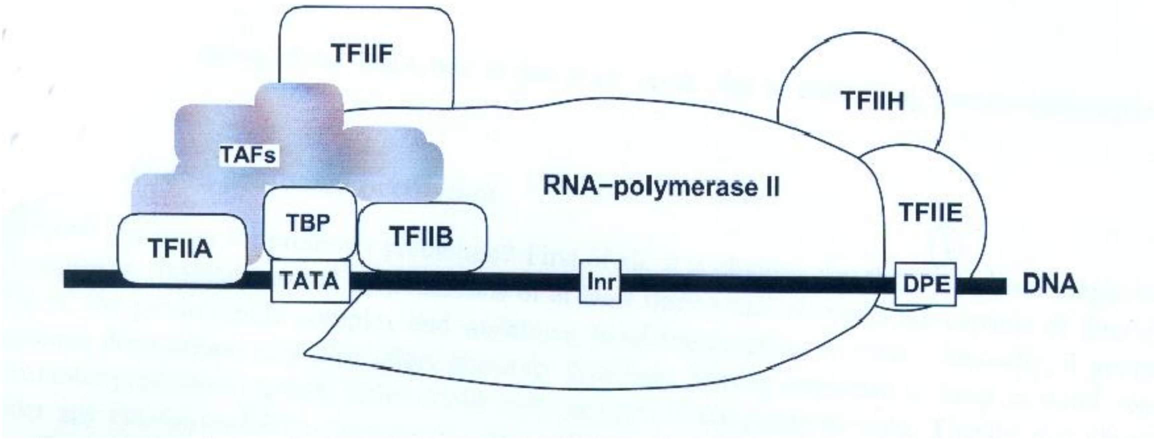


Figure I-3. Formation of pre-initiation complex through the binding of transcription factors to DNA nearby the transcription start site [Pederson et al. (1999)].

I-1.4 Nature of Protein-DNA Binding

TFs have the affinity of binding to a specific DNA sequence. The binding sequence is usually between 5-20 bp long and is identified experimentally. Interestingly not all bases are found to be equally important for effective binding. While some base positions can be substituted without affecting the affinity of the binding, in other positions a base substitution can completely obliterate the binding. A *consensus sequence* or ***motif*** represents the common features of the effective binding site sequence. The TF has high affinity for sequences that match this consensus pattern, and relatively low affinity for sequences different from it. A numerical way of characterizing the binding preferences of a TF is the *positional weight matrix* (PWM) (see section III-2.2), which shows the degree of ambiguity in the nucleotide at each binding site position.

The ambiguity of TF binding appears to be intentional in nature as a way of controlling gene expression. Variable affinity of the TF to different DNA sites causes a kinetic equilibrium exists between TF concentration and occupancy (i.e. which binding sites are actually occupied with the TF *in-vivo*). This provides a mechanism of controlling the transcription of the genes.

I-1.5 Cis-Regulatory Sequences

The DNA sequences where TFs bind in order to regulate gene expression are known as cis-regulatory sequences. The DNA region immediately upstream of the TSS (i.e., in the 5' direction with negative position index) is usually is the center of such activity and is known as the *promoter* (Figure I-2). For example in Figure I-1, the -2000 to -1 sequence marked in black color is the promoter. The binding sites for various TFs within the promoter have been marked with yellow outlines. The promoter contains binding sites for TFs that directly interact with RNA polymerase II to promote transcriptional initiation. The structure and functioning of eukaryotic promoters has been discussed by several reviewers [Werner (1999), Pederson et al. (1999), Zhang (2002)]. The main functional elements within the promoter are the transcription factor binding sites, while the rest of the sequence is nonfunctional and meant to separate the binding sites at an appropriate distance.

There are other cis-regulatory sequences apart from the promoter which enhance or repress the transcription activity. The *cis-regulatory module* (CRM, *enhancer* or *repressor*) is a short sequence that stimulates transcriptional initiation while located at a considerable distance from the TSS. CRMs are often involved in inducing tissue-specific or temporal expression of genes. A CRM may be 100-1000 bp in length and contains several closely arranged TFBS. Thus a CRM resembles the promoter in its composition and the mechanism by which it functions. However a CRM typically contains higher density of TFBS than the promoter, has repetitive TFBS, and involves greater level of cooperative or composite interactions among the TFs. The activity of a CRM is interesting as it can control gene expression from any location or strand orientation. The present understanding of its mechanism is that TFs bound at the CRM interact directly with TFs bound to the promoter sites through the coiling or looping of DNA.

I-1.6 Transcriptional Regulation of Development

One of the most intriguing applications of the study of gene regulation is in understanding the process of development. Development refers to the process of growth of a multicellular organism from a single cell to adult. This dissertation focuses on Drosophila melanogaster (fruit fly) which is a model organism for studying development. Drosophila development occurs in a series of stages including embryo, three larval stages, a pupal stage, and finally the adult stage. The embryo development is further divided into 16 stages (Bownes stages). The single celled zygote first undergoes multiple divisions of the nucleus (stages 1-3). The early Drosophila embryo exists as a single cell with multiple nuclei, called syncytial blastoderm (stage 4). The cytoplasm then gradually divides to form multiple mononucleate cells, forming the cellular blastoderm (stage 5-6). The next stage is gastrulation (stage 7) where separation of different tissues begins to manifest and the rough body plan of the larval structures is established. In subsequent stages (stage 8-16) the cells divide and differentiate further till morphologically distinct organs are formed.

The process by which cells which were similar in the beginning start specializing into specific types or tissues is called *differentiation*, which is at the heart of development. Differentiation is the result of a complex network of gene expression accomplished largely through transcriptional control. A number of genes expressed in the developmental phase encode transcription factors (TFs). The TFs operate in a hierarchical fashion so that TFs released at one stage lead to the expression of genes that release TFs for the next stage. At each stage the complexity of expression pattern increases. A crucial mechanism behind differentiation is the non-uniform distribution of TFs in the embryo cells. The early syncytial blastoderm embryo contains several TFs derived from the mother, which are non-uniformally distributed through the embryo along both anterior-posterior and dorsal-ventral axes. At any given location, various TFs are present in different concentrations. Depending on the TF concentrations, specific CRMs are activated to express or repress their target genes. This results in differential expression of the zygotic genes in different locations. The network of differential gene expression continues, ultimately leading to tissue differentiation. The interaction between TFs and CRMs is thus a fundamental mechanism that controls development.

I-2 Motivation for Present Research

I-2.1 Scope of the present research

As the complete DNA sequences of genomes for many organisms including microbes, plants, animals and human beings have become available, the first task is to annotate these genomic sequences [Stein (2001)]. Annotation refers to locating important functional elements such as genes (introns and exons), transcription start sites, translation start sites, splice sites, polyadenylation sites, gene promoters, etc. on the genomic sequence. For processing the voluminous genomic data, laborious and time consuming experimental techniques alone are insufficient. Computational methods are playing an important role in the ongoing task of detecting and annotating functional signals in genomic sequences. For instance computationally annotated features in the ENCODE project [Encode (2004)] are shown in Figure I-4.

This research work aims at improving the computational modeling and detection of three very important signals – transcription factor binding motif, promoter (transcription start site) and cis-regulatory module (CRM or enhancer). The significance of this problem in current bioinformatics research is highlighted by the fact that the computational investigation of DNA motifs, promoters and CRMs is listed as one of the important computational biology research goal for the next few years in the “Genomes to Life” program (Figure I-5) of the U.S. Department of Energy [Frazier et al. (2003)].

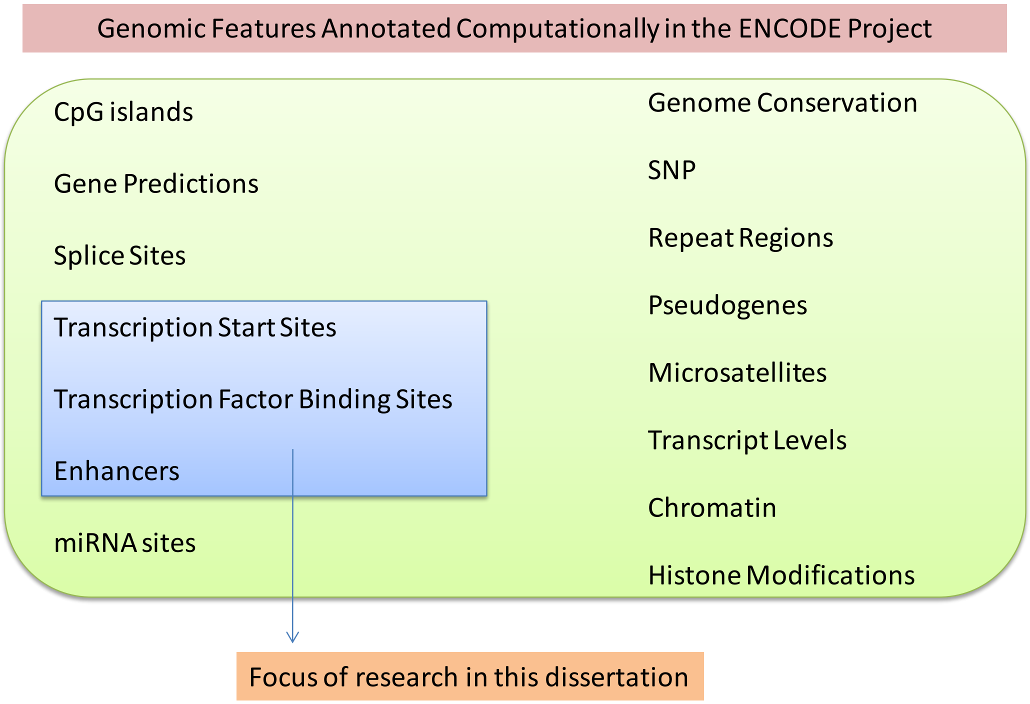


Figure I-4. Several genomic features are currently being computationally annotated in the human genome in the ENCODE project. The present research focuses on three features in the regulatory sequence track: transcription start sites, transcription factor binding sites (motifs) and enhancers (cis-regulatory modules).

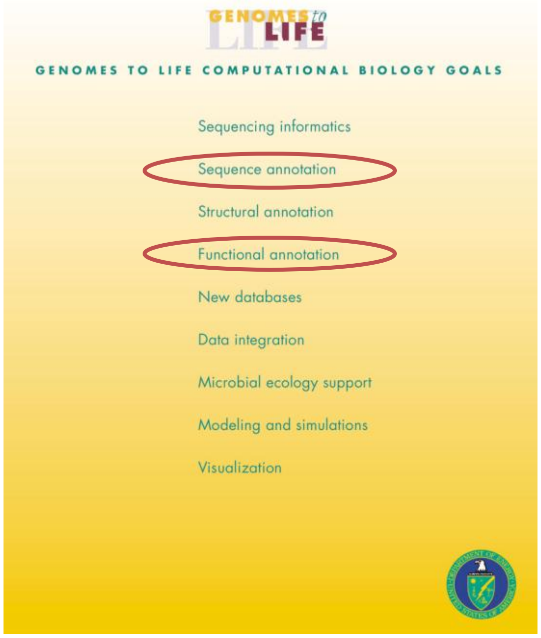


Figure I-5. The "Genomes to Life" program of the U.S. Department of Energy [Frazier et al. (2003)] plans for the next 10 years to use DNA sequences from microbes and higher organisms, including humans, as starting points for systematically tackling questions about the essential processes of living systems. Advanced technological and computational resources will help to identify and understand the underlying mechanisms that enable organisms to develop, survive, carry out their normal functions, and reproduce under myriad environmental conditions.

I-2.2 Relevance of the present research

Computational prediction of promoters (transcription start site) transcription factor binding motifs, and cis-regulatory modules (CRMs or enhancers) has specific relevance in the current bioinformatics research. Reliable computational prediction of promoters and transcription start sites (TSS) is currently required in automated gene discovery. Gene annotation is currently incomplete in a number of sequenced genomes.

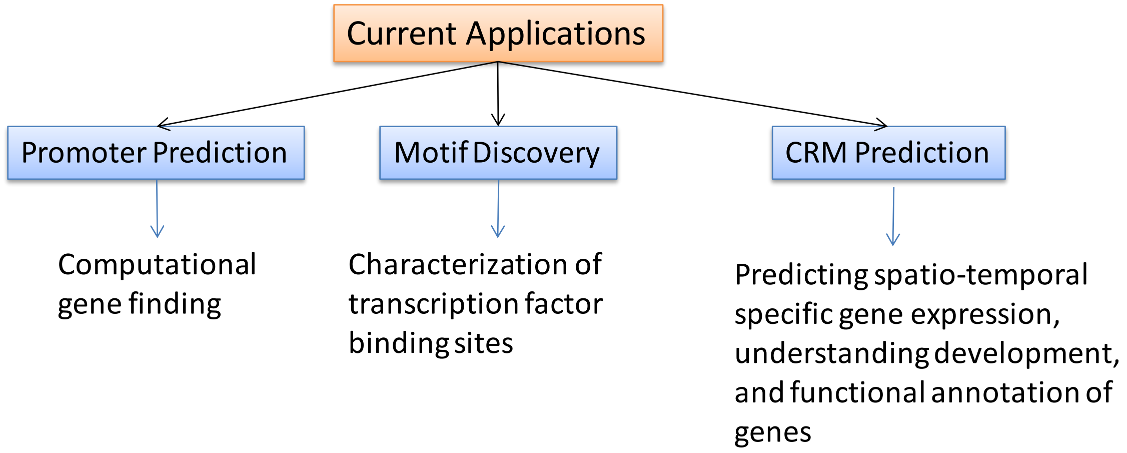

Current Applications
Promoter Prediction
Motif Discovery
CRM Prediction
Computational gene finding
Characterization of transcription factor binding sites
Predicting spatio-temporal specific gene expression, understanding development, and functional annotation of genes

Figure I-6. Applications of the present research in current bioinformatics context.

Though genes can usually be mapped using cDNA and homology with existing annotations, genes with no cDNA transcripts or close homolog must be mapped by computational gene-finding. In fact, a majority of genes are currently annotated using computational gene prediction. While gene finding algorithms can predict introns and exons with about 80% accuracy [Guigo et al. (2006)], the locations of TSS and splice sites are still difficult to predict, with none of the existing methods reporting more than 45% accuracy [Guigo et al. (2006)]. The accuracy of TSS prediction is particularly low at around 35% sensitivity [Bajic et al. (2006)] and a large number of false positives [Fickett and Hatzigeorgiu (1997), Werner (2003)]. This causes the gene-finding algorithm to produce wrong partitioning of exons in obtaining the overall gene structure. Accurate TSS prediction to locate the 5' end of genes and first exons will be clearly helpful.

The identification of transcription factor binding motifs is one of the most basic requirements for understanding gene regulatory mechanisms. Although many TFs are known, specific binding motifs have been fully characterized for only few of them in databases such as TRANSFAC [Matys et al. (2003)] or JASPAR [Sandelin et al. (2004)]. The motifs in these databases are derived from their experimentally determined DNA binding sequences using DNAse footprinting [Brenowitz et al. (1986)]. However DNAse footprinting is costly, laborious and time consuming, and therefore it can be performed only for a few binding sequences. *In-silico* methods have long been used to supplement the experimental approach. The *in-silico* approach analyzes a set of several sequences that possibly contain binding sites for the same protein factor. A large amount of such sequence data is now available through high throughput ChIP technologies (ChIP-Chip, ChIP-PET, ChIP-Seq, etc.), promoters of co-regulated genes identified by microarray, and upstream regions of orthologous genes from closely related species. Still the binding site is difficult to distinguish from the surrounding DNA as it is short in length (5-20 bp) and contains various mutations. Thus reliable computational algorithms are required to search for the common conserved motif. Characterization and detection of biologically meaningful motifs is a long standing research problem in computational biology.

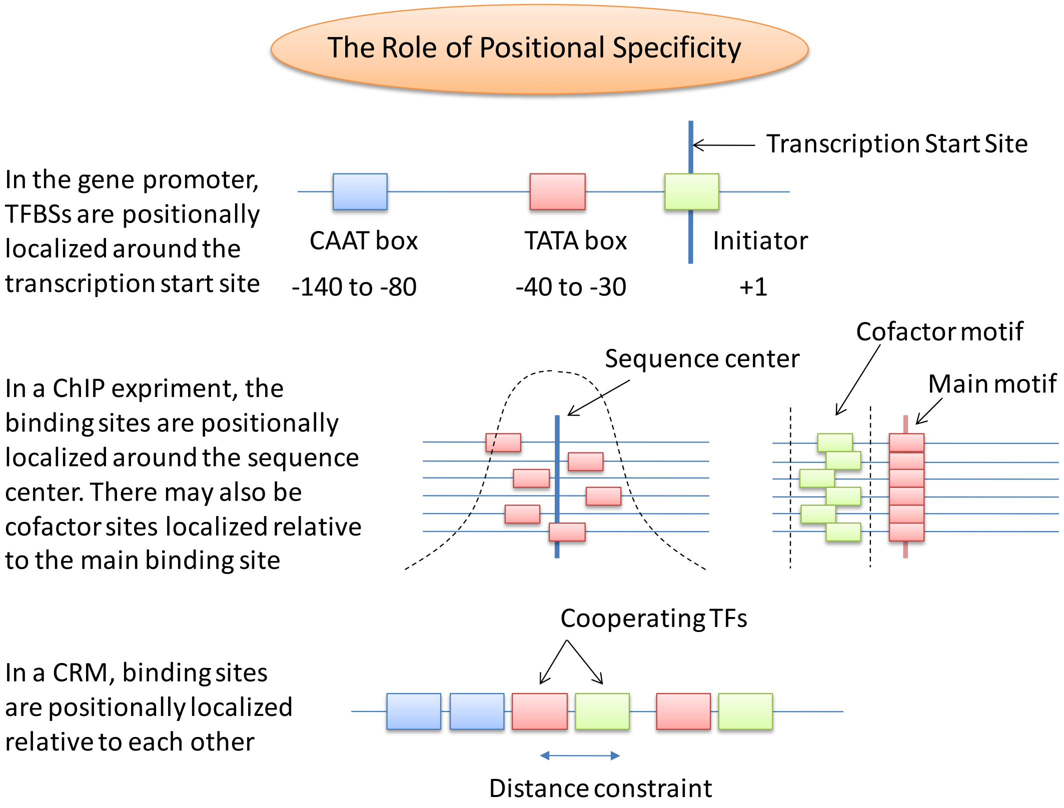
A recent paradigm in the modeling and detection of regulatory regions, especially in higher eukaryotes, is the study of clusters of binding sites for multiple TFs that act in concert [Crowley (1997), Wasserman and Fickett (1998), Frech et al. (1998), etc.]. Though potential TFBS occur with high frequency in the genome, a significant proportion of them are nonfunctional [Euskirchen and Snyder (2004)]. The reason is that TFs function collectively and not individually. Cis-regulatory modules (CRMs) [Arnone and Davidson (1997)] are one such type of autonomous units to which a set of TFs bind cooperatively. Their annotation is especially important for understanding spatio-temporal specific gene expression in the developmental genes in higher eukaryotes. Detection of

CRMs has received particular attention in Drosophila melanogaster and human genomes [Gallo et al. (2006), Sharan et al. (2004)]. CRM prediction also has potential application in determining the functional annotation of uncharacterized genes. Many newly sequenced genes in various species have no functional annotation and the sequence analysis of their protein product also gives no clue on their function. As CRMs are often responsible for context-specific gene expression, *in-silico* functional annotation may be possible by identifying specific CRMs controlling these genes. For instance, novel mucle specific genes could be identified through computational identification of muscle specific CRMs near those genes [Frech et al. (1998)].

I-2.3 Position information in the modeling of regulatory elements

The tasks of modeling and detection are closely related. Accurate modeling is necessary for producing a robust computational detection method, which requires taking into account the underlying biological mechanism. The present research improves upon the previous studies by incorporating a crucial biological aspect, namely position and order of the functional elements, into the computational model.

It is interesting to note that the computational modeling of transcription factor binding motifs, promoters and CRMs are all associated with a notion of position specificity (Figure I-7). Functional binding sites are often found proximal to and at a specific distance from genomic features such as TSS, splice site or a related binding site. In fact, TFBS in the promoter are positioned carefully with respect to each other and the TSS [Werner (1999)]. In ChIP experiments, the binding sites for the immunoprecipitated TF are concentrated around the center of the ChIP sequence. Additionally cofactor binding sites may be located at specific positions around the main TF binding sites.



In the gene promoter, TFBSs are positionally localized around the transcription start site

In a ChIP expriment, the binding sites are positionally localized around the sequence center. There may also be cofactor sites localized relative to the main binding site

In a CRM, binding sites are positionally localized relative to each other

Figure I-7. Transcription factor binding motifs, promoters and CRMs are all associated with a notion of position specificity.

Similarly in CRMs, the TFBS occur in a preferred order and distance with respect to each other [Bailey et al. (2003), Sinha et al. (2003)]. These characteristics have not been adequately exploited in the modeling and detection of these features. The present research develops computational approaches / models that effectively integrate the positional information associated with these features.

I-2.4 Bayesian network modeling

With respect to the modeling framework, the present research relies upon probabilistic modeling using Bayesian networks [Jensen (2001)]. Although the genomic sequence is a fixed deterministic sequence, on the functional level its composition and the mechanism of its expression are stochastic in nature. DNA sequences are tolerant to mutations and displacement of functional elements. Therefore uncertainty based modeling is possible.

Within the framework of probability models, Bayesian networks appear attractive for the modeling of genomic data due to several inherent advantages. Bayesian networks can easily and intuitively incorporate knowledge of the biological mechanism into the model, where causal relationships among the variables of interest can be defined both qualitatively and quantitatively. The Bayesian network model is transparent in contrast to neural networks or SVM, for example inspection of the model parameters directly reveals the probabilisitic relationships among the variables. This helps gain understanding about the problem domain and reveals new knowledge.

Bayesian networks have also shown superior performance as a computational machine learning tool. Bayesian networks can easily integrate prior expert knowledge into the model, which is an inheritance from the Bayesian statistical framework. Thus reliable inference can be made using a Bayesian network even using small training datasets, and overfitting of data to the model can be avoided, ensuring that the learnt model is more representative of the true population. Both continuous and discrete variables can coexist in a Bayesian network. The present research benefits from the the above advantages offered by Bayesian networks.

I-3 Nature of the Problem

The present research involves three related problems, *viz.* (1) detection of DNA motifs, (2) general promoter modeling and transcription start site prediction, and (3) modeling and detection of cis-regulatory modules. The computational nature of each of these problems and the challenges therein are discussed below.

I-3.1 Detection of DNA Motifs

In-silico detection of protein-DNA binding motifs involves analyzing a set of several sequences that possibly contain binding sites for the same protein factor. The binding sites are unknown in each of the input sequences, but they are conspicuous in the sequence set as similar repeating patterns. The problem is however nontrivial as the binding sites are short in length (5-20 bp) and contain various mutations. The computational algorithm searches for a common conserved pattern called the *motif*.

Though multiple alignment tools such as CLUSTALW [Thompson et al. (1994)], ITERALIGN [Brocchieri et al. (1998)] or PROBE [Neuwald et al. (1997)] could be used to detect a conserved pattern or block within the given set of sequences, detection of motifs is more difficult since they are short, lesser conserved and randomly distributed patterns. A specialized computational algorithm for motif detection has three aspects [Friberg et al. (2005)]:

(i) *The motif model:* The motif is represented by a computational model which represents the nature of protein-DNA binding and the similarities and variabilities among the individual binding sites. Common examples are (*l*,*d*) motif model [Pevzner and Sze (2000)] and positional weight matrix (PWM) [Stormo (2000)].

(ii) *The scoring function:* It is a numerical score to measure the prominence or conservation of a motif in the given set of sequences, usually against a background model. For example the Z-score [Tompa (1999)] and relative entropy score [Stormo (2000)].

(iii) *The algorithm:* In accordance with the motif model and the scoring function, the computational algorithm searches for the best candidate motif within the given set of sequences by a strategy such as exhaustive search [Staden (1989)], heuristic search [Pevzner and Sze (2000)], greedy search [Hertz and Stormo (1999)], multiple sequence alignment [Tharakaraman et al. (2005)], Gibbs sampling [Lawrence et al. (1993)], etc. The best scoring candidate is reported as the desired motif.

Each of the above aspects contributes to the performance of the motif finding algorithm. Furthermore a motif finding algorithm must address the challenges of time and memory complexity, noisy input in the form of spurious sequences which do not contain a binding site, conservation of the motif against random patterns [Keich and Pevzner (2002a,b)], accuracy of the background model, and competition among multiple motifs in the input sequences.

I-3.2 General Promoter Modeling and Transcription Start Site Prediction

Computational promoter prediction involves differentiating promoter *versus* non-promoter regions in a given genomic sequence, and predicting the locations of transcription start sites (TSS). The main conserved functional elements within a promoter sequence are short length TFBS, while the rest of the sequence follows the random genomic background. A promoter sequence is therefore hard to distinguish from the rest of the genome. Also, promoter sequences hardly show any sequence similarity among themselves even for closely related genes, thus sequence similarity searching (such as BLAST) is ineffective in detecting promoter sequences. Specialized computational promoter prediction algorithms are thus required.

A computational promoter prediction algorithm must rely on two aspects: (i) recognition of TFBS (motifs), and (ii) modeling the combinations and context of these TFBS within promoter sequences. While a lot is known about various TFBS (motifs) that play an active role in eukaryotic poly-II promoters [Latchman (2003)], yet detection of individual TFBS is insufficient for detecting the promoter. For instance, about 30% of the human promoters contain a conserved binding site known as TATA box upstream of the TSS. However binding sites for TATA box occur on an average once every 1000 bp, and thus it is insufficient in itself to characterize the promoter. The crucial aspect is to model the context of several TFBS within the promoter. This is where the difficulty in constructing a general computational model arises. A great amount of diversity and complexity is observed in the organization of TFBS in promoters. There is no general universal concept known to be applicable to all promoters. There are thousands of transcription factors and their corresponding binding sites, with highly variable contexts observed among different promoter sequences, making the modeling very difficult.

I-3.3 Modeling and Detection of Cis-Regulatory Modules

Computational modeling and prediction of CRMs poses greater challenge than promoters as (i) available data and biological information on CRMs is far less as compared to promoters [Gallo et al. (2006)], (ii) different CRMs are extremely varied in composition and their organization is even lesser understood than promoters [Arnone and Davidson (1997)], (iii) CRMs are intrinsically more difficult to model and predict than promoters since they may be located at any distance from the TSS and lack conserved anchoring features such as TATA box, CAAT box etc. which are found in promoters.

CRMs have been most widely studied in the genome of Drosophila melanogaster (fruit fly). Two kinds of CRMs have generally been observed in Drosophila – *homotypic* CRMs which are composed of multiple binding sites for a single TF, and *heterotypic* CRMs which have binding sites for more than one TF. It is currently understood that the gene expression pattern (i.e. the region/tissue and the stage of gene expression) directed by a CRM depends upon the specific set of TFs that bind to the CRM. A set of TFs that cooperatively bind to a CRM is called a "regulatory code". The regulatory codes are specific as only certain TFs can cooperate in the same regulatory event. For example the TFs *bicoid*, *caudal*, *hunchback*, *knirps*, *Kruppel*, *giant*, *tailless*, etc are known to regulate gene expression in the blastoderm embryo [Berman et al. (2002); Schroeder et al. (2004)]. Whereas the set of TFs *dorsal*, *twist*, *su(H)*, etc. governs gene expression in the embrynoic neuroectoderm [Markstein et al. (2004)]. CRMs are defined to be of different "types" according to their specific regulatory codes. CRMs of the same type will express in the same tissue and developmental stage.

The current computational techniques model CRMs of a specific type as a sequence of fixed length (such as 700bp) in which the number of TFBS of the regulatory code TFs exceeds a certain threshold [Markstein et al. (2002), Berman et al. (2002), Rajewsky et al. (2002), Lifanov et al. (2003), Schroeder et al. (2004)]. The regulatory code is obtained from biological knowledge. Currently only three specific regulatory codes are known for gene expression in the embryonic blastoderm, mesoderm and neuroectoderm. Thus the computational studies are limited to only few specific types of CRMs. Moreover, currently the binding motifs (PWMs) are accurately known only for a few TFs. Thus the scope of computational CRM prediction is presently very limited.

I-4 Research Objectives

The objective of the present research is to utilize the information of order and distance preferences of protein-DNA binding sites in each of the three problem domains, *viz.* DNA motif detection, general promoter and TSS prediction, and modeling and detection of CRMs, and create pragmatic computational models/strategies which improve the prediction performance. This section briefly summarizes the specific research problems that have been addressed in this research.

I-4.1 Detection of Localized Motifs

The present research especially addresses localized motif discovery in long regulatory sequences. Currently there is a need for analyzing motifs in long sequences in ChIP experiments, vertebrate promoters, etc. Recent studies [Keich and Pevzner (2002a,b), Buhler and Tompa (2002), Chin et al. (2004)] have shown that in long sequences random patterns become at least as prominent as the real motif, therefore any motif finding algorithm will report a number of spurious motifs that overshadow the real motif. In addition, for most motif finding algorithms the time and memory requirements increase greatly for an increase in sequence length. This forms the motivation for pursuing a specialized approach for motif detection in long regulatory sequences.

It is recognized in the literature that binding sites usually occur within the regulatory sequences in a position-specific manner relative to a biological landmark. For example many TFBS are appropriately located relative to the TSS to allow TFs to anchor at specific positions with respect to each other and the TSS [Smale and Kadonaga (2003), Roepcke et al. (2006)]. Several other examples are reported in this dissertation. In such situations, it is possible to detect the motif by searching for it in an appropriate local sequence interval after aligning the sequences relative to an anchor point. Localization removes the sequence regions that do not contain the motif, thus increasing the strength of the motif relative to noisy random patterns. For instance, in Figure I-8(a) a random pattern appears as most repeated and conserved within the complete sequence length, whereas in Figure I-8(b), if only a short local interval relative to an anchor point is analyzed, the real biologically relevant motif is discovered. [Ohler et al. (2002)] analyzed motifs in 1941 Drosophila regulatory sequences of length 300bp each aligned (-250,+50) relative to the TSS. The analysis of complete 300bp sequence did not reveal many of the core promoter motifs. However, in a separate analysis of the local region (-60,+40), most core promoter motifs were discovered. Similarly [Molina and Grotewold (2005)] analyzed the (-50,-1) and (+1,+50) regions of Arabidopsis Thaliana promoters separately in order to discover the core promoter motifs.

An apparent solution is to subdivide the long sequences (aligned relative to an anchor point) into short overlapping intervals of equal length and analyze each interval with a motif finding algorithm. However there are inherent problems in this approach. Firstly, apart from specific situations such as the analysis of core promoters, the region of localization of the motifs is not known *a priori*. When a general motif finding tool is used to search for motifs in an arbitrary sequence interval, it reports a number of random motifs that are locally over-represented but not globally conserved. The difference between a locally over-represented random motif and a globally conserved “localized motif” is illustrated in Figure I-9. The localized motif has a specific confinement within a sequence interval when observed at a global level, while a random motif has no such confinement. The scoring function of a general motif finding algorithm assigns high

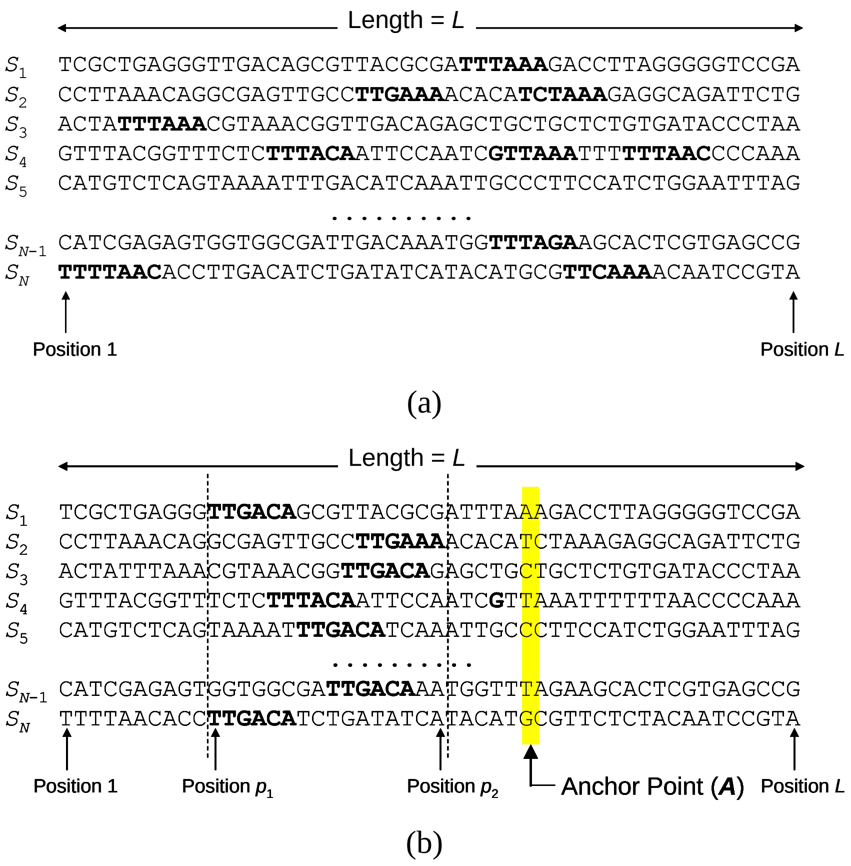

Length = L
S1 TCGCTGAGGGTTGACAGCGTTACGCGATTTAAAGACCTTAGGGGGTCCGA
S2 CCTTAAACAGGCGAGTTGCCTTGAAAACACATCTAAAGAGGCAGATTCTG
S3 ACTATTTAAACGTAAACGGTTGACAGAGCTGCTGCTCTGTGATACCCTAA
S4 GTTTACGGTTTCTCTTTACAATTCCAATCGTTAAATTTTTTAACCCCAAA
S5 CATGTCTCAGTAAAATTTGACATCAAATTGCCCTTCCATCTGGAATTTAG
SN-1 CATCGAGAGTGGTGGCGATTGACAAATGGTTTAGAAGCACTCGTGAGCCG
SN TTTTAACACCTTGACATCTGATATCATACATGCGTTCAAAACAATCCGTA
Position 1
Position L
(a)
Length = L
S1 TCGCTGAGGGTTGACAGCGTTACGCGATTTAAAGACCTTAGGGGGTCCGA
S2 CCTTAAACAGGCGAGTTGCCTTGAAAACACATCTAAAGAGGCAGATTCTG
S3 ACTATTTAAACGTAAACGGTTGACAGAGCTGCTGCTCTGTGATACCCTAA
S4 GTTTACGGTTTCTCTTTACAATTCCAATCGTTAAATTTTTTAACCCCAAA
S5 CATGTCTCAGTAAAATTGACATCAAATTGCCCTTCCATCTGGAATTTAG
SN-1 CATCGAGAGTGGTGGCGATTGACAAATGGTTTAGAAGCACTCGTGAGCCG
SN TTTTAACACCTTGACATCTGATATCATACATGCGTTCTCTACAATCCGTA
Position 1
Position p1
Position p2
Anchor Point (A)
Position L
(b)

Figure I-8. Discovering (6,1) motifs within a set of N sequences $S_1, S_2, \ldots, S_N$ of length L. In (a) the random pattern TTTAAA is seen to eclipse the real motif TTGACA when the complete sequence is analyzed, but in (b) the real motif TTGACA becomes dominant when only the local interval (p_1,p_2) is considered.

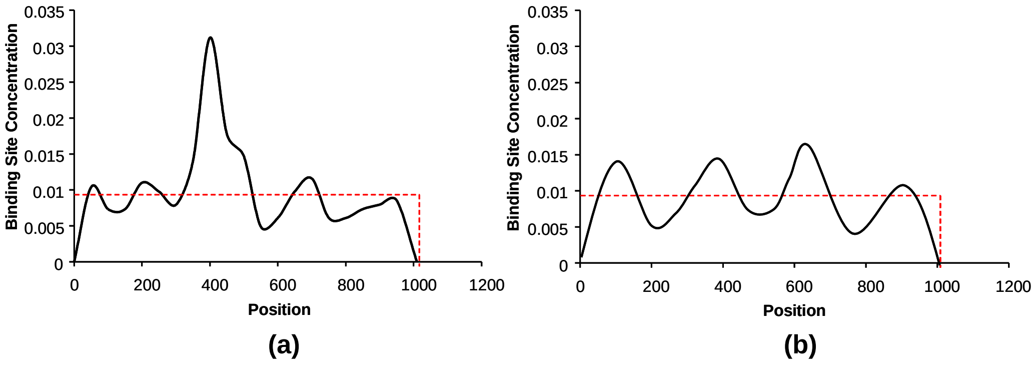


Figure I-9. Difference between the distribution of binding sites of (a) a localized motif, and (b) a spurious motif. While both may appear over-represented in a local sequence interval, localized motifs have a prominent region of confinement within the entire sequence length.

scores to the random motifs, making it difficult to differentiate them from localized motifs. Moreover, among a large number of motifs reported over all intervals, it is not easy to identify motifs that are most relevant over the entire sequence length. Secondly, the interval length must be chosen carefully as if it is too short compared to the localization region then the motif may not appear prominently in any of the intervals, and if it is too long then the motif may again remain obscured. This is illustrated by an example in Figure I-10, where the detection of multiple motifs spread over regions of different length requires selection of different interval lengths. In practice, even a 100 bp difference in the interval length yields entirely different results. Thirdly, the manual task of fragmenting the sequences and combining together the results for different intervals is laborious, time consuming and prone to error. It would be useful to have an automated, efficient algorithm which can accurately demarcate the region of localization of the motifs and detect them.

[Tharakaraman et al. (2005)] incorporated positional preference in their motif finding algorithm GLAM by performing gapless local alignment over windowed

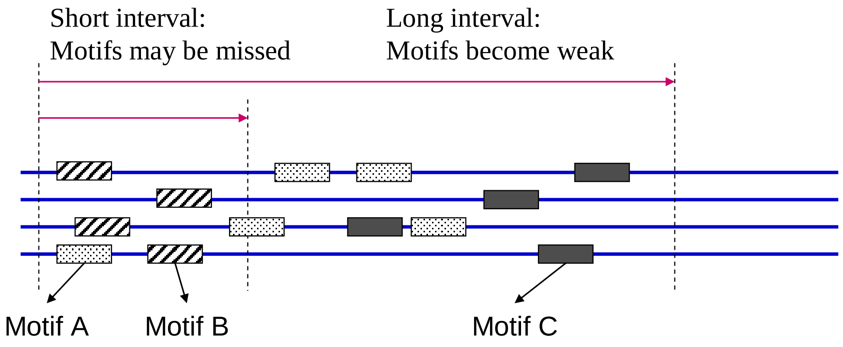


Figure I-10. An illustration of the difficulties in analyzing sub-intervals of long regulatory sequences – for short intervals, motifs A and C are missed, and for long intervals the motifs may become weak.

subsequences of the original sequence set (aligned relative to the TSS) instead of the complete length. However the algorithm, being slow and computationally expensive, is practical only for the analysis of short sequences. This dissertation introduces the concept of localized motif finding and presents an algorithm called *LocalMotif* [Narang et al. (2006)] for detecting localized motifs in long regulatory sequences.

I-4.2 Bayesian Network Model for General Promoter Prediction

Computational algorithms for general promoter prediction currently have a sensitivity of about 35% and produce a large number of false positives. Different algorithms make different simplifying assumptions about the context of TFBS in a promoter. The accuracy achievable by a promoter model greatly depends upon how well it emulates the real biological context. For example, a simple model of promoter as a region with a high TFBS density [Prestridge (1995)] had only 13% sensitivity and two false predictions per true prediction [Fickett and Hatzigeorgiu (1997)], while a more refined model considering positions of the TFBS relative to the TSS [Down and Hubbard (2002)] improved the accuracy to 29% sensitivity and 0.5 false predictions per true prediction. Further modeling refinements have been proposed in the literature [Werner (2003)], such accounting for synergistic or antagonistic coordination among binding sites, modeling the positions and order of binding sites relative to each other, modeling physical properties of sequences around the TSS such as DNA bendability, stability, curvature, chromatin structure etc. These aspects have not yet been implemented.

A parallel approach in computational promoter prediction is using artificial intelligence (AI) based systems. These algorithms do not directly search for known motif signals, but rather perform unsupervised learning of string features (motifs) that are unique to promoter sequences. The features may be identified through discrimination between a set of training examples of promoter and non-promoter sequences using machine learning and statistical techniques [Hutchinson (1996), Chen et al. (1997), Scherf et al. (2000), Bajic et al. (2003)]. The context of these features within the promoter is also learnt from training examples in an unsupervised manner using AI modeling techniques such as artificial neural networks. Increasing the modeling complexity and carefully tuning the training process allows high accuracy to be achieved with the unsupervised learning approach [Scherf et al. (2000), Bajic et al. (2003)]. The advantage of this approach lies in the ability to recognize using machine intelligence compositional aspects of promoter sequences that are not so far physically understood.

The present research combines known biological concept of modeling positions and order of TFBS relative to the TSS, with the AI approach of performing de-novo learning of promoter features from sequence, in a computational promoter prediction model of improved performance called *BayesProm* [Narang et al. (2005)].

I-4.3 Cis-Regulatory Module Prediction in the Drosophila Genome

The current computational approach for CRM prediction characterizes the CRMs as short (~1 kb) genomic segments containing high density of binding sites for a set of co-acting TFs [Frech et al. (1998); Wasserman and Fickett (1998); Frith et al. (2001); Berman et al. (2002); Markstein et al. (2002); Rajewsky et al. (2002); Bailey and Noble (2003); Sinha et al. (2003); Berman et al. (2004); Markstein et al. (2004); Schroeder et al. (2004)]. The set of cooperating TFs is called the regulatory code. The binding sites of the TFs are recognized with the help of positional weight matrices (PWMs) for the TFs [Stormo (2000)].

The main challenge for this approach is that the compositions of different CRMs regulating gene expression in different developmental stages and tissues are exceedingly varied. Each specific expression profile is governed by a specific regulatory code. Presently the regulatory codes are extracted based on tedious wet-lab experiments and biological knowledge. Only three such codes are currently known to our best knowledge. Thus the applicability of the approach is quite limited. Another limitation is that good quality PWMs, which are required to predict the TFBS clusters, are available only for a few TFs. Currently the PWMs have been computed from a small number of experimental TFBS sequences determined by DNAse footprinting. Most of these PWMs lack sufficient sensitivity and specificity [Narang et al. (2006)].

On the other hand, available experimental data on CRMs has expanded in the recent years, but has not been utilized so far towards computational modeling and prediction of CRMs. Sequence based modeling of CRMs such as using oligonucleotide frequencies has been recently attempted and has shown some degree of success in modeling blastoderm CRMs [Chan and Kibler (2005)]. However, the performance diminishes considerably on various other CRM types [Li et al. (2007)]. The main reason, as shown in the present research, is that oligonucleotide motifs produce a large number of false matches in the non-TFBS segments of a CRM. These non-TFBS segments are not conserved across CRMs. Therefore the model is inaccurate.

The present research develops a computational CRM modeling and prediction approach called *Modulexplorer* [Narang et al. (2008)] to perform de-novo learning of regulatory codes for Drosophila CRMs from CRMs of unknown types. Modulexplorer inputs a database of known CRMs and a set of non-CRM background sequences and characterizes the TFBSs within the CRMs de-novo. It then uses a probabilistic Bayesian network model to learn the TFBS interactions in CRMs. These interactions describe the regulatory codes. The trained model is used to discover novel CRMs.

I-5 Organization of the Thesis

Three specific research problems which form the subject of this thesis were briefly introduced in Section I-4. Each of these is presented in a separate chapter – the localized motif finding problem is addressed in Chapter 4, followed by the general promoter prediction problem in Chapter 5 and finally cis-regulatory module prediction in Chapter 6. Each chapter is self-contained with the problem statement, methods and results. A review of the current literature within the scope of this research is given in Chapter 2. Some common mathematical preliminaries are provided in Chapter 3. The main conclusions of this research and future work are finally discussed in Chapter 7.

CHAPTER - II

LITERATURE REVIEW

Several computational approaches have appeared in the literature addressing each of the three problems subject of the present research. An exhaustive review of these would make this dissertation voluminous. Thus only a summary of the relevant existing approaches is presented below, highlighting their similarities and differences with the present research.

II-1 Detection of DNA Motifs

Numerous computational methods and tools have been reported over the past fifteen years or so for discovering motifs in regulatory regions of genes. Recent reviews on the subject can be found in [Tompa et al. (2005), D'haeseleer (2006a,b), Wasserman and Krivan (2003)]. The different approaches differ in terms of the motif model, scoring function and algorithm.

A number of different representations of a motif are available in the literature. Most algorithms model the motif as either consensus sequence, or consensus sequence with possible gaps, or as positional weight matrix (PWM) (refer Chapter 3). Other probabilistic model based representations of a motif such as hidden Markov model [Durbin et al. (1998), Xing et al. (2004)], Bayesian network model [Barash et al. (2003)], variable order Bayesian network [Ben-Gal et al. (2005)], etc. are also found in the literature. The present research uses a particular consensus based representation called (l,d) motif [Waterman et al. (1984), and Pevzner and Sze (2000)], where the motif is a nucleotide pattern of fixed length l such that any observed binding site has a maximum of d point mutations from this pattern. Though the PWM representation is preferred for modeling motifs with experimentally known binding preferences, the consensus or (*l*,*d*) motif representation is found equally or more effective in *ab-initio* motif detection [Pavesi et al. (2001), Tompa et al. (2005)].

A recent review of the various scoring functions used for motif detection can be obtained in [Li and Tompa (2006)], while an assessment of various scoring functions was performed in [Friberg et al. (2005)]. The simplest scoring functions for consensus based motif representation are the *total distance score* and *sum of pairs score* [Pevzner and Sze (2000)], which measure the degree of conservation of a (*l*,*d*) candidate motif within the set of input sequences. However, these scoring measures do not capture the complexity of DNA sequences in terms of their non-uniform oligonucleotide content. Thus several motif finding tools score the *statistical over-representation* of a motif in the given set of sequences, for example oligo-analysis [van-Helden et al. (1998)], MobyDick [Bussemaker et al. (2000)], YMF [Sinha and Tompa (2000)], Projection [Buhler and Tompa (2002)], etc. The over-representation is measured relative to the general nucleotide content of the given set of sequences, known as *genomic background*. The background is usually modeled as a stationary stochastic process with a Markov model (see Section III-1). While earlier tools used a zero order Markov model to represent the genomic background, it has been realized recently that higher order Markov models produce better efficiency of motif detection [Thijs et al. (2001), Marchal et al. (2003), Pavesi et al. (2001)]. In addition to over-representation, there are other important measures of goodness of a motif such as *relative entropy* [Stormo (2000)] which measures the amount of surprise in observing the motif pattern under the background model, *sequence specific score* [Pavesi et al. (2001)] which measures whether the motif appears in a sufficiently large percentage of the input sequences, etc.

The algorithms used to search for the best candidate motif usually belong to one of the two categories: word enumeration and probabilistic optimization. The choice of the algorithm partially depends upon the chosen motif representation. Word enumeration based algorithms employ a consensus sequence representation of the motif with or without gaps. An *exhaustive enumeration* approach [Waterman et al. (1984)] involves considering all possible 4^l candidate (l,d) motifs and scoring them. Though an exact algorithm, it has high time complexity. Thus Pevzner and Sze (2000), and Eskin and Pevzner (2002) introduced three *heuristic search* algorithms, SP-STAR, WINNOWER and MITRA. SP-STAR first considers candidate patterns that have an exact match in any of the sequences, and then heuristically extends the search to include more candidate patterns which are similar to the best scoring patterns. WINNOWER and MITRA translate the motif finding problem to an equivalent problem of finding cliques in a graph, and find a quick heuristic solution by pruning inessential edges in the graph. A faster implementation of exhaustive enumeration is possible using suffix tree [Ukkonen (1995)], which can enumerate all valid occurrences of a candidate pattern in all the sequences in O(1) time. Taking advantage of this approach, fast algorithms such as SMILE [Marsan and Sagot (2000)] and Weeder [Pavesi et al. (2001)] have appeared. Recently the motif finding problem has also been formulated as a search for the maximum density subgraph of a graph whose nodes are the words in the input sequences, and whose edges connect similar words [Fratkin et al. (2006)]. The resulting optimization can be performed in polynomial time. Some word enumeration algorithms consider only the exact matches of a length *l* pattern in the sequences in order to detect the motif [Staden (1989), van Halden et al. (1998), Tompa (1999)]. All word enumeration algorithms algorithms may be extended to detect gapped motifs by considering only the significant positions in the alignment.

Probabilistic motif finding algorithms represent a motif as a positional weight matrix (PWM). The PWM which has the lowest probability of occurring by chance (or highest score) describes the most novel pattern, which is presumably the motif being sought. Considering that one binding site for the motif is present in each of the *N* sequences of length *L*, $\left(L-l+1\right)^N$ different PWMs can be possibly formed, making an exhaustive search algorithm impractical. Therefore different algorithms have been devised to efficiently search for the optimal PWM. An approximate heuristic method was used in CONSENSUS [Hertz et al. (1990)]. A systematic optimization approach later appeared as the MEME algorithm [Bailey and Elkan (1994)]. MEME fits a statistical model to the given set of sequences, consisting of the motif model (i.e., the PWM), the background model described as a zero order Markov model, and a weight parameter representing the mixing frequency of the motif and the background models. The accuracy of the overall statistical model is measured by a likelihood function, which is optimized iteratively using the expectation maximization (EM) algorithm to find the best motif and background models. Being an EM based solution, MEME finds the local rather than the global optimum. A related optimization approach called Gibbs sampling, which is a stochastic equivalent of the EM, has been implemented in several other tools such as GibbsDNA [Lawrence et al. (1993)], AlignACE [Roth et al. (1998)], MotifSampler [Thijs et al. (2002)], BioProspector [Liu et al. (2002)], ANN-spec [Workman and Stormo (2000)] etc.

Latest advances in motif detection try to make use of information other than just the regulatory sequence to improve the prospects of detecting the motif. For instance, in regulatory sequences that have been identified using ChIP-chip analysis, ChIP enrichment information may be used to enhance motif detection [Liu et al. (2002), Ettwiller et al. (2007)]. From the knowledge of the nature of interaction between nucleotides and amino acids in DNA-binding domains of a set of transcription factors, binding sites for other related transcription factors may be possible to derive [Mandel-Gutfreund et al. (2001), Kaplan et al. (2005)]. Specialized algorithms are being developed to discover composite motifs, which are spaced dyads or ordered sets of motifs with strong distance constraints [van Helden et al. (2000), Eskin and Pevzner (2002), Wijaya et al. (2007)].

The literature in the area of motif finding is indeed vast, and to maintain the brevity of this review, aspects and references of lesser relevance to the present context have been intentionally left out.

II-2 General Promoter Modeling and Transcription Start Site Prediction

A number of tools for the detection of general promoters and TSS are reported in the literature. There are two categories of modeling approaches or tools for promoter prediction. The first category of tools [Kondrakhin et al. (1995), Prestridge (1995), Down and Hubbard (2002)] utilize positional weight matrices (PWM) derived from experimental data [Bucher (1990)] for detecting putative TFBS and identify sequence regions with a high density of binding sites as possible promoters. The state of the art in this category, Eponine [Down and Hubbard (2002)], improves the model quality and prediction accuracy by associating with each PWM the probability distribution of its position relative to the TSS. The second category of tools [Hutchinson (1996), Chen et al. (1997), Scherf et al. (2000), Bajic et al. (2003)] recognize promoters based on their sequence composition. Characteristic features of promoter sequences are learnt automatically from a set of training examples using machine learning or statistical techniques. An unknown sequence is then classified as promoter or non-promoter based on its feature content. Most tools use oligonucleotides of fixed length as features [Hutchinson (1996), Chen et al. (1997), Bajic et al. (2003)] and select the best features based on occurrence frequencies of oligonucleotides in promoter *versus* non-promoter training datasets. PromoterInspector [Scherf et al. (2000)] uses IUPAC groups, which are oligonucleotides permuted with wildcards, as features.

About 6-10 years ago, the first generation of tools could predict less than 30% of the actual TSS, while reporting one false positive every 1000 bp [Fickett and Hatzigeorgiou (1997)]. Recent research has focused on achieving improved TSS prediction performance through better tuning and increased modeling complexity. The resultant 2nd generation tools [Werner (2003)], such as PromoterInspector, Eponine, Dragon Promoter Finder, *etc.* have accuracy which is suitable for whole genome scale prediction. However, the increase in sensitivity has been much less compared to the improvement in the specificity of these tools. More recently, biologically motivated approaches such as CpG+ [Hannenhalli and Levy (2001)] and gene start finding tools such as First Exon Finder [Davuluri et al. (2001)] and Dragon Gene Start Finder [Bajic and Seah (2003)] have exploited features such as CpG islands and first splice donor sites to improve the accuracy of TSS prediction. Approaches utilizing physico-chemical properties [Uren et al. (2006)] and structural properties [Abeel et al. (2008)] of DNA have been proposed recently, however they have lower accuracy than the sequence based methods.

II-3 Modeling and Detection of Cis-Regulatory Modules

The literature on the computational modeling and detection of cis-regulatory modules is fairly recent and limited. Modeling and prediction techniques have developed independently in two different areas *viz.* vertebrate CRMs and CRMs in Drosophila melanogaster. Few vertebrate CRM models have appeared in the literature such as FASTM [Klingenhoff et al. (1999)], logistic regression analysis [Wasserman and Fickett (1998)] and Modulesearcher [Aerts et al. (2003)]. These models study specific CRMs present close to the TSS, which are involved in tissue specific gene expression such as in muscle tissues. They are not discussed in detail here since the focus of this dissertation is on Drosophila CRMs, and models for Drosophila CRMs are characteristically different from vertebrate CRM models.

The simplest computational model for Drosophila CRMs was as a cluster of TFBS. Markstein et al. (2002) modeled a *homotypic* CRM in Drosophila (Figure II-1a) as a cluster of TFBS for a single TF. They considered a cluster of three or more binding sites of the TF named *Dorsal* in a 400 bp window as a CRM. Similarly [Berman et al. (2002)] modeled a *heterotypic* CRM as a cluster of TFBS for multiple TFs (Figure II-1b). A minimum of fifteen TFBS for a set of five TFs – two maternal TFs (bicoid and caudal) and three gap TFs (hunchback, Kruppel and knirps) – were requisite in a 700 bp window to classify it as a CRM. Prediction quality was improved by [Rajewsky et al. (2002)] and

[Schroeder et al. (2004)] by allowing overlapping and weak binding sites into the computational model, considering a statistically optimal combination of TFBS over a sequence window rather than single matches. The cluster model is specified based on biological knowledge of mutually interacting TFs in a CRM, and it requires high quality PWMs as input. Thus it has limited application.

Another way of modeling a CRM as a cluster of TFBS is using hidden Markov models (HMM) [Frith et al. (2001)]. PWMs for a set of related TFs are supplied by the user based on prior biological knowledge, and the HMM uses these to discover clusters of TFBS (CRMs) in a given genomic sequence. The HMM (Figure II-1c) has three types of states: inter-cluster background, intra-cluster (i.e. between the TFBS) background, and motif states. There is a separate motif state for each TF, with its emission probabilities defined using the PWM. The model scans the regulatory sequence base by base to compute the probability that the query sequence contains a cluster of TFBS.

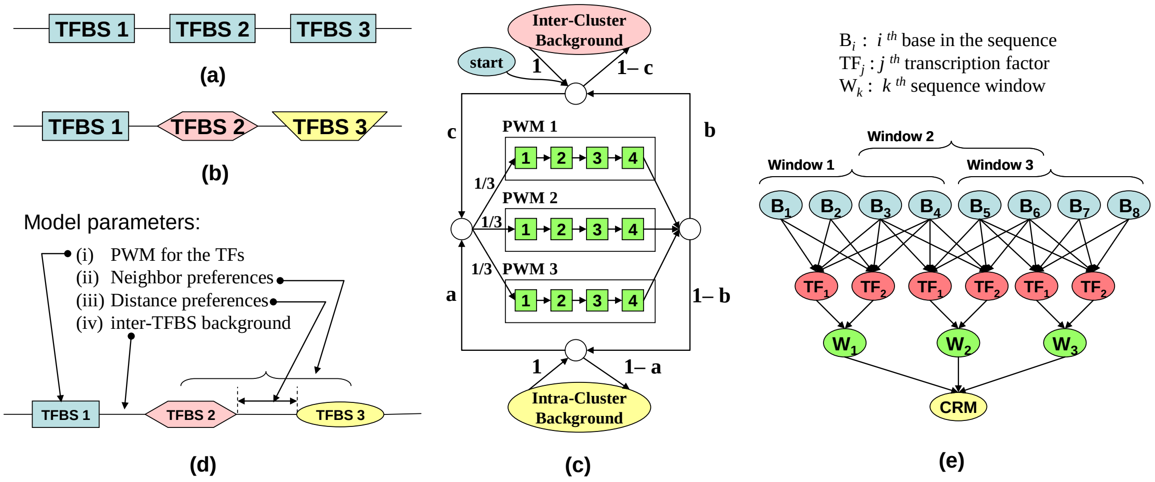


Figure II-1. Computational models for cis-regulatory modules: (a) homotypic cluster of TFBS [Markstein et al. (2002)], (b) heterotypic cluster of TFBS [Berman et al. (2002)], (c) hidden Markov model [Frith et al. (2001)], (d) statistical

model of Gupta and Liu (2005), (e) discriminatory Bayesian network model of Segal and Sharan (2005).

The HMM approach was further developed by Sinha et al. (2003) to include the information of order in which TFBS are organized in a CRM (i.e. when one TFBS consistently follows another). The modified model produced superior results compared to the basic model where binding sites are expected to occur in any random order. In another study, Bailey and Noble (2003) incorporated a penalty for inter-cluster and intra-cluster distances within the HMM, and again observed an improvement in the quality of predictions. These studies point out two important factors in CRM modeling, i.e., the information of gap and order among the TFBS.

Two algorithms have appeared recently for learning a new CRM model de-novo from sequence data of co-regulated genes. Given as input a set of sequences putatively containing CRM of the same type, they attempt to discover multiple coexisting motifs and learn their PWMs. However the requirement is that all given sequences contain the same CRM type restricts their applicability to very specific datasets. Gupta and Liu (2005) propose a statistical model (Figure II-1d) with the following unknown parameters: (i) the number of TFs, (ii) PWMs for the TFs, (iii) neighbor preferences of each TFBS in the form of a transition matrix, (iv) distance preferences between neighboring TFBS modeled as a truncated geometric distribution, (v) inter-TFBS background modeled by a Markov chain. The model parameters are learnt from sequence data using Bayesian inference with Markov chain Monte Carlo and Gibbs sampling algorithms. Segal and Sharan (2005) use a discriminative Bayesian network (Figure II-1e) to learn a fixed number of PWMs that best discriminate between two given sequence sets – sequences that contain an unknown CRM, and background sequences. The model defines a CRM as a cluster of binding sites for the PWMs in a fixed length sequence window, with the discriminative Bayesian network model measuring odds that (i) a particular sequence contains a CRM window, (ii) a window is a CRM , and (iii) a window contains binding site for a given PWM. Maximum likelihood estimation of the model parameters using the expectation maximization (EM) algorithm can thus obtain the unknown PWMs. In both studies, initialization of parameters is the most crucial aspect. *Ab-initio* motif finding with human intervention is used to produce intelligent initial guesses for the PWMs in both studies.

The present research concerns development of a CRM model which learns regulatory codes de-novo. None of the computational studies (including in vertebrates) have so far addressed this problem. The present research is unique in this aspect.

CHAPTER - III

PRELIMINARIES

This chapter discusses the fundamental computational modeling concepts used throughout this dissertation, including Markov modeling of genomic background (Section III-1), consensus sequence and PWM based representations of DNA motifs (Section III-2), fundamentals of Bayesian network modeling (Section III-3), and measures of prediction accuracy (Section III-4). Each section is self-contained and is referred to in the later chapters wherever required. A reader familiar with the problem domain may skip this chapter and refer back to the relevant sections if required.

III-1 Stochastic Model of the Genome

III-1.1The Background Model

In the human genome, which contains approximately 3 billion bases, only 1% of the sequence is exons that code for proteins, 24% is introns, 22% is intergenic DNA and the rest 53% is repetitive DNA. Apart from the existence of short regulatory signals such as transcription factor binding sites proximal or distal to the TSS, no specific function is known for most DNA. Mathematical model for this ‘background’ DNA is required in order to be able to distinguish it from functional elements. Simple frequencies of individual bases have been used in several computational studies [Lawrence et al. (1993), Bailey and Elkan (1994), van Halden et al. (1998), Tompa (1999)]. However, the genomic background is not as simple. For instance, a dinucleotide feature such as CpG island cannot be identified using individual base frequencies. As mentioned in Chapter 2, a more complex representation of the background in the form of a higher order Markov model has been found useful in improving the efficiency of the computational method.

The Markov model (or Markov chain) is a stochastic process which has the Markov property that the current state of the process depends only on the recent past. Let X_t be a discrete time stochastic process observed at $t = 1, 2, 3, \ldots$, and let the state space of X_t be also discrete. The Markov property for a Markov model of order *q* states that at any time instant *i*, the distribution of the observation X_i is conditionally dependent only on the previous *q* observations, i.e.,

$$\Pr\left(X_i \middle| X_{i-1}, X_{i-2}, \ldots, X_0\right) = \Pr\left(X_i \middle| X_{i-1}, X_{i-2}, \ldots, X_{i-q}\right) \qquad (3.1)$$

In the Markov model representation of the genome background, a sequence of nucleotides $S_1 S_2 \ldots S_N$, where $S_i \in \{A, C, G, T\}$, is treated as an instance or realization of the Markov process X_t with the random variable X_i having discrete states $\{A, C, G, T\}$. This physically implies that the nucleotide S_i at any position *i* depends only on the previous *q*-mer of nucleotides $S_{i-q} \ldots S_{i-2} S_{i-1}$.

A simple way of visualising a Markov chain is through a finite state machine. Consider a Markov model of order 1, i.e., $\Pr\left(X_i \middle| X_{i-1}, X_{i-2}, \ldots, X_0\right) = \Pr\left(X_i \middle| X_{i-1}\right)$, which implies that the next observed nucleotide in the sequence depends only on current nucleotide. The finite state representation of this model is shown in Figure III-1. The probability distribution $\Pr\left(X_i \middle| X_{i-1}\right)$ is characterized by the set of probabilities $p_{B_1 B_2}$, where $B_1, B_2 \in \{A, C, G, T\}$, and $p_{B_1 B_2}$ is the probability of observing the base B_2 following the base B_1 in the sequence. These probabilities are independent of the position index *i* of the base in the sequence. The 4×4 matrix of these probabilities,

$P = \left[p_{B_1 B_2} \right]_{B_1 B_2 \in \{A,C,G,T\}}$, is called the *transition matrix*. Let the probability distribution for the first base, $\Pr\left(X_1\right)$, be the vector $I = \left[p_A \quad p_C \quad p_G \quad p_T \right]$. The probability distribution for the second base is given as $\Pr\left(X_2\right) = I.P$, and similarly for the i^{th} base as $\Pr\left(X_i\right) = I.P^{i-1}$. Thus the complete Markov chain, in this first order case, can be characterized by the initial vector *I* and the transition matrix *P*.

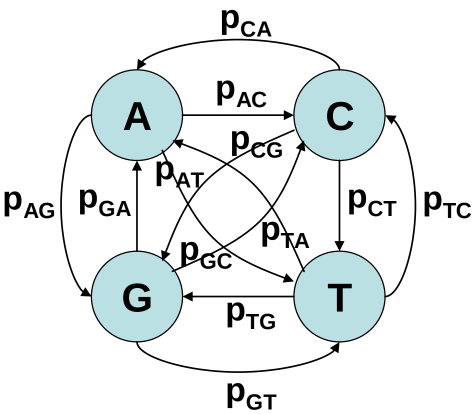


Figure III-1. Finite state machine visualization of a first order Markov model for sequence background.

The above concepts may be generalized to an order *q* Markov model. The state transition in this case is from a *q*-mer to a single nucleotide. The transition probabilities are thus the probabilities $\Pr\left(B_1 B_2 \ldots B_q \rightarrow B_{q+1}\right)$, and the transition matrix *P* is a $4^q \times 4$ matrix. The initial state is again a *q*-mer, and thus the initial probability vector, *I*, contains 4^q entries.

The background model parameters *P* and *I* are usually estimated from a set of given background genomic sequences. Such sequences are collected at random from intergenic regions which are not suspected to contain any functional elements. Let

$f\left(B_1B_2\ldots B_l\right)$ denote the number of occurrences of the *l*-mer $B_1B_2\ldots B_l$ in these sequences. Then the initial and transition probabilities are estimated as

$$\Pr\left(B_1B_2\ldots B_q\right)=1/f\left(B_1B_2\ldots B_q\right), \tag{3.2}$$

and

$$\Pr\left(B_1B_2\ldots B_q\rightarrow B_{q+1}\right)=\frac{f\left(B_1B_2\ldots B_{q+1}\right)}{f\left(B_1B_2\ldots B_q\right)} \tag{3.3}$$

III-2 Computational Modeling of Protein-DNA Binding Sites (Motifs)

Due to their degenerate nature, the binding sites are not fixed strings but are represented as a model called motif. Two frequently used ways of representing a motif *viz.* consensus sequence and positional weight matrix are described below.

III-2.1Consensus sequence

The consensus sequence is a string or regular expression that matches all the binding site examples closely, but not necessarily exactly. For instance, the consensus CONS1 in Figure III-2 is determined by choosing the base with the highest occurrence frequency at each position of the binding site. The consensus CONS2 uses IUPAC nomenclature (Figure III-3) of single letter codes to represent ambiguity in the base at any particular position. A base is considered significant at a position if occurring in more than 25% of the binding sites.

There is a tradeoff between sensitivity and specificity in choosing the consensus representation and the number of allowed mismatches with the consensus. Sensitivity refers to the percentage of binding sites that can be identified using the chosen consensus and the maximum mismatch value. For example, the consensus CONS1 above has a sensitivity of 40% with 2 allowed mismatches and 90% with 3 mismatches. Specificity refers to how frequently a match with the consensus would be found. The probability of a random match with consensus CONS1 is 1 in 1.1 million for up to 2 mismatches and 1 in 81,000 for up to 3 mismatches. The CONS2 consensus has a sensitivity of 50% with 1 mismatch and 90% with 2 mismatches, and specificity of 1 in 1.6 million with 1 mismatch and 1 in 41,000 with 2 mismatches.

	Position →														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Seq 1	C	T	T	G	G	C	C	A	A	T	C	A	G	A	A
Seq 2	T	T	C	A	G	C	C	A	A	T	C	G	G	A	G
Seq 3	C	G	C	G	G	C	C	A	A	T	C	A	G	C	G
Seq 4	T	T	T	A	G	C	C	A	A	T	C	A	G	C	T
Seq 5	C	C	T	G	G	C	C	A	A	T	C	A	G	C	G
Seq 6	C	C	C	G	G	C	C	A	A	T	C	A	G	C	G
Seq 7	G	T	T	A	G	C	C	A	A	T	C	A	G	C	A
Seq 8	A	T	C	A	G	C	C	A	A	T	G	A	G	C	T
Seq 9	C	C	C	A	G	C	C	A	A	T	C	A	G	A	G
Seq 10	C	T	C	A	G	C	C	A	A	T	G	G	G	C	G
CONS1	C	T	C	A	G	C	C	A	A	T	C	A	G	C	G
CONS2	C	Y	Y	R	G	C	C	A	A	T	C	A	G	M	G

Figure III-2. A small sample of binding sites for the transcription factor NF-Y.

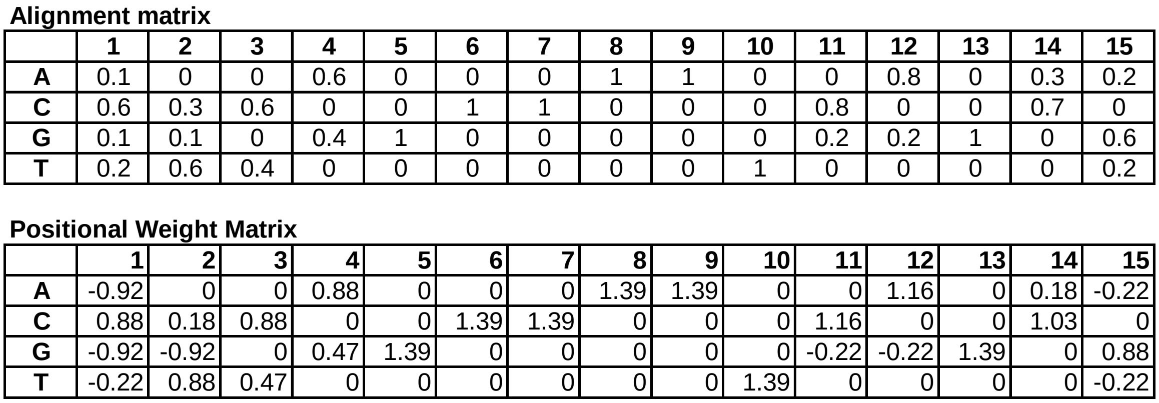
Symbol	A	C	G	T	R	Y	M	K
Meaning	A	C	G	T	A/G	C/T	A/C	G/T
Symbol	S	W	H	B	V	D	N	
Meaning	G/C	A/T	A/C/T	G/C/T	A/C/G	A/G/T	A/C/G/T	

Figure III-3. Single-letter IUPAC codes for representing degeneracy of nucleotides.

The consensus representation is thus not unique and the optimal consensus depends upon the application in question [Day and McMorris (1992)]. For representing protein-DNA binding sites, the CONS1 type of representation, which is basically the (*l*,*d*) motif, is most often used.

III-2.2Positional Weight Matrix

A more informative representation for the binding site of a protein is in the form of a positional weight matrix (PWM) [Stormo et al. (1982), Stormo (2000)]. The PWM records the base conservation at each binding site position. First an alignment matrix is formed whose entries are the frequencies, $f_{b,j}$, of the nucleotides, $b \in \{A, C, G, T\}$, in the positions, $j \in \{1, 2, \ldots, l\}$, among known binding site sequences. For example Figure III-4 shows the alignment matrix for the binding site data shown in Figure III-2. The base conservation is measured by a weight $w(b, j) = \ln\left(\frac{f_{b,j}}{p_b}\right)$, where p_b is the background frequency of the base *b*. The weight $w(b, j)$ is positive when the proportion of base *b* at the position *j* in the alignment is greater than its proportion in general (according to background). It measures the amount of surprise in the observed conservation of the base. The $4 \times l$ matrix of the weights $w(b, j)$ is called the positional weight matrix as shown in Figure III-4.



Alignment matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0.1	0	0	0.6	0	0	0	1	1	0	0	0.8	0	0.3	0.2
C	0.6	0.3	0.6	0	0	1	1	0	0	0	0.8	0	0	0.7	0
G	0.1	0.1	0	0.4	1	0	0	0	0	0	0.2	0.2	1	0	0.6
T	0.2	0.6	0.4	0	0	0	0	0	0	1	0	0	0	0	0.2

Positional Weight Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	-0.92	0	0	0.88	0	0	0	1.39	1.39	0	0	1.16	0	0.18	-0.22
C	0.88	0.18	0.88	0	0	1.39	1.39	0	0	0	1.16	0	0	1.03	0
G	-0.92	-0.92	0	0.47	1.39	0	0	0	0	0	-0.22	-0.22	1.39	0	0.88
T	-0.22	0.88	0.47	0	0	0	0	0	0	1.39	0	0	0	0	-0.22

Figure III-4. Positional weight matrix developed from the collection of NF-Y TFBS in Figure III-2.

Another interpretation of the weight is that it represents the information content, which is a measure of discrimination between the binding of a functional DNA sequence and an arbitrary DNA sequence. The information is measured by the formula

$$I_{total} = \sum_{j=1}^{l} I_j \quad \text{with} \quad I_j = \sum_{b=\{A,C,G,T\}} f_{b,j} \ln\left(\frac{f_{b,j}}{p_b}\right) = \sum_{b=\{A,C,G,T\}} f_{b,j} w(b,j), \qquad (3.4)$$

where I_j is the information in the base conservation at position *j*. Note that $\sum_b f_{b,j} = 1$. I_{total} is also known as the *relative entropy* or the Kullback-Liebler distance between the background and the motif.

A strong correlation has been observed between the information represented by the PWM and the affinity of the protein's binding with a sequence [Stormo (2000)]. Consider a sequence $S = S_1 S_2 \ldots S_l$ with $S_i \in \{A, C, G, T\}$. The binding energy of the protein's interaction with the sequence *S* has been observed as directly correlated with the measure $\Delta G(S) = \sum_{j=1}^{l} w(S_i, j)$. This is the sum of the values that each base of the sequence *S* has in the weight matrix. The implication of this formula is that each weight estimates the binding energy at that position in the binding site and each position contributes independently to the total binding energy.

The PWM can thus be used to search for potential binding sites in an uncharacterized sequence $S = S_1 S_2 \ldots S_L$. At each position, $p \in \{1, 2, \ldots, L - l + 1\}$, of the uncharacterized sequence *S*, a window $\tilde{S}_p = S_p S_{p+1} \ldots S_{p+l-1}$ of length *l* is selected. Using the PWM, the "matrix score" for this window is calculated by the formula [Bucher (1990)]:

$$\text{Matrix score for } \tilde{S}_p = \frac{\sum_{j=1}^{l}\left[w\left(S_{p+j-1}, j\right) - w\left(\min_j, j\right)\right]}{\sum_{j=1}^{l}\left[w\left(\max_j, j\right) - w\left(\min_j, j\right)\right]}, \tag{3.5}$$

where $\max_j$ and $\min_j$ represent the rows for which $w\left(b, j\right)$ is maximum and minimum respectively in the column *j*. The matrix score is a real number within the range [0, 1]. If the matrix score for the window $\tilde{S}_p$ exceeds a chosen threshold value, it is marked as a potential binding site. Typically the score threshold is selected based on the scores of known binding sites. Unfortunately, however, PWM based binding site detection is not fully reliable and can produce a large numbers of false positives [Stormo (2000)].

III-3 Bayesian networks

Bayesian networks offer several advantages as a modeling tool within the context of bioinformatics applications. This section briefly discusses the most fundamental Bayesian network modeling concepts.

A Bayesian network is a formalism to represent and reason about probabilistic cause-and-effect relationships among a set of entities or events in an intuitive manner. It has two components – (i) a graphical map of the cause-and-effect relationships among the entities or events in the domain, and (ii) a numerical measurement of the extent of this dependence.

In the graph, each entity or event is represented as a node, and the cause-effect relationships among the nodes are shown by directed edges linking causes to effects. This is technically called a directed acyclic graph (DAG). For example, consider a Bayesian network model of the causes of heart disease as shown in Figure III-5. The different events or causes associated with heart disease such as diet, obesity, blood pressure, smoking etc. are shown in oval shaped nodes. The causal relationships are shown as edges, e.g. since exercise directly affects obesity, blood pressure and arteriosclerosis, it is the parent of these three nodes.

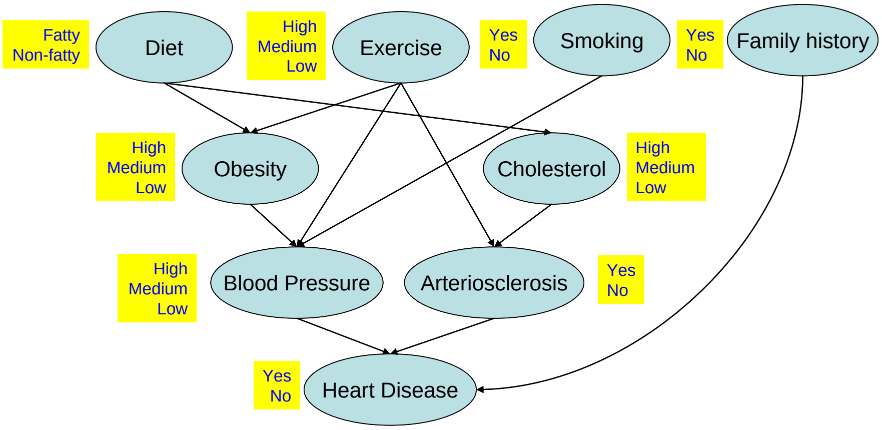


Figure III-5. A Bayesian network for modeling the causes of heart disease.

In the numerical representation, each entity or event (from now onwards referred to as a node in the Bayesian network) is represented by a variable which can take a set of possible values or states for the event. For example, the set of possible states of each node are shown alongside the nodes in Figure III-5 in rectangular captions. A conditional probability table (CPT) is associated with each variable to quantify the extent to which the variable is likely to be affected by other variables. For example the CPT of obesity is illustrated in Figure 6, showing the probabilistic dependence of an individual’s obesity on his diet and exercise habits. Each row of the CPT shows how obesity is affected by a particular combination of its parents, diet and exercise. E.g. a fatty diet with low exercise is likely to produce obesity in 35% of the cases (row 3 of the CPT). Note that the sum of probabilities in each row of the CPT is always 1.

Diet	**Exercise**	**Obesity=High**	**Obesity=Medium**	**Obesity=Low**
Fatty	High	0.2	0.4	0.4
Fatty	Medium	0.25	0.5	0.25
Fatty	Low	0.35	0.5	0.15
Non-fatty	High	0.1	0.2	0.7
Non-fatty	Medium	0.15	0.25	0.6
Non-fatty	Low	0.2	0.3	0.5

Figure III-6. Conditional probability table (CPT) for the node “obesity” in the Bayesian network of Figure III-5.

In the Bayesian network structure as shown in Figure III-5, a node from which there is an edge to another node is called a parent of that child node, e.g. the node “diet” is a parent of the node “obesity”. Similarly there is an ancestor-dependent relationship between nodes that are linked in a chain, e.g. “diet” is an ancestor of “blood pressure”. These relationships describe how one variable influences the state of another variable. The parent nodes directly influence the child node, while the ancestor nodes have an indirect influence upon their descendants. There exists a conditional independence relationship in the network, which is stated as follows: a node is independent of its ancestors given its parents. E.g. since diet affects blood pressure not directly but through obesity, once the information of a person’s obesity is available, knowledge of his diet does not give any additional information about his blood pressure.

Defining a Bayesian network model for a given problem involves specifying (a) the variables or nodes in the graph, (b) the set of possible states for each node, (c) the edges connecting the nodes in the graph, (d) the probability distributions or CPTs associated with each node. The former three, i.e. the nodes, states and the edges, comprise the Bayesian network *structure*, and the latter comprises the *parameters*. The structure represents modeler’s understanding or beliefs about the problem domain, and there is a fair bit of flexibility possible in choosing the structure.

In mathematical terms, the concept of conditional independence is explained as follows. Each node in the Bayesian network is a random variable. The complete *joint distribution* of this set of *N* random variables $X_1, X_2, \ldots, X_N$ is given by the *chain rule* as

$$\Pr\left(X_1, X_2, \ldots, X_N\right) = \prod_{i=1}^{N} \Pr\left(X_i \middle| X_{i-1}, X_{i-2}, \ldots, X_1\right). \tag{3.6}$$

Note that the variable X_i is conditioned on the variables $X_{i-1}, X_{i-2}, \ldots, X_1$ which precede it in the topological ordering. The conditional independence between the variables allows this joint distribution to be simplified. Instead of being conditioned on all its predecessors, the node X_i is conditioned only on its parents. Thus in the simplified expression,

$$\Pr\left(X_1, X_2, \ldots, X_N\right) = \prod_{i=1}^{N} \Pr\left(X_i \middle| \mathrm{Pa}\left(X_i\right)\right), \tag{3.7}$$

where $\mathrm{Pa}\left(X_i\right)$ denotes the set of parents of the node X_i in the Bayesian network. If each variable X_i has *m* possible states, the full joint distribution would require O(m^N) parameters. Whereas the factored form of the Bayesian network would require only O(Nm^k) parameters, where *k* is the maximum number of parents for any node. Thus the Bayesian network formalism makes mathematical modeling much simpler.

The purpose of a Bayesian network is to estimate certainties of events that are not directly observable. For example, whether or not a patient has heart disease cannot be directly known, however a doctor can *infer* about it using knowledge of associated symptoms. As information regarding the symptoms accumulates, the doctor’s belief about the existence of heart disease changes accordingly. For example, if the doctor comes to know that the patient smokes, his belief about the patient's chances of having heart disease increases. The Bayesian network can be used to make intelligent inferences similar to the medical expert. After representing the problem domain in terms of a Bayesian network, one can use it to reason how information about states of certain nodes in the network changes the belief about states of other nodes. This is called *inference* using a Bayesian network.

An interesting aspect of Bayesian network modeling is that both the network's structure and parameters (CPTs) can be determined from a known set of data automatically using algorithms such as Expectation-Maximization (EM). Estimation of the parameters is called *parameter learning* and estimation of the structure is called *structure learning*, and the complete process of learning from given data is called *training* of the Bayesian network. How parameter learning and inference are performed in a Bayesian network is explained below.

As a machine learning tool, a Bayesian network can learn from examples to simulate the real world phenomenon. Learning, in this context, refers to the procedure of updating the parameter values (CPTs) of the Bayesian network model to make it representative of the known examples. The known examples are referred to as training data. The measure of how well the model fits the training data is provided by the likelihood function, which indicates how likely the Bayesian network is to produce this data, i.e.

$$\text{Likelihood function} = \Pr(\text{ Data} \mid \text{Model }). \tag{3.8}$$

The learning algorithm updates the model parameters so as to maximize the value of the likelihood function, and the parameter values thus obtained are known as maximum likelihood estimates. The basic idea therefore is to find a model configuration which is more likely than any other to produce the given data.

The *expectation-maximization (EM) algorithm* is a general method which can be used to obtain maximum likelihood estimates of the parameters of a Bayesian network for a given training dataset. The EM algorithm works even when the dataset is incomplete or has missing values. Missing values are encountered not only in problems where there are limitations in the data gathering process, but rather they occur more frequently in situations where there are hidden or unobserved variables in the system.

The EM is an iterative algorithm with two steps – Expectation step (E-step) and Maximization step (M-step). In the E-step, the current parameters are used to estimate the missing data using the inference procedure as was described above. In the M-step, the filled-in data is used to perform maximum likelihood estimation of the parameters. The updated parameter values obtained in the M-step are again used in the next E-step to make a new (improved) estimate of the missing data. An M-step again follows to update the parameter values. In this way the EM steps are repeated iteratively until convergence.

Bayesian networks are a powerful formalism for mathematical modeling of real world phenomena. The above short description gave an overview of the essential concepts including model building, inference and parameter estimation. For detailed mathematical treatment, the reader may refer to [Jensen (2001)] and [Narang et al. (2006)].

III-4 Measures of Accuracy

Since a major theme of the present work is detection of functional elements in the genome, evaluation of prediction accuracy is frequently required. As functional elements are specific discrete signals on the genome, the result of a prediction is binary, i.e., either the prediction is true, or the prediction is false. The comparison between the set of predictions and the set of existing functional elements can be summarized in terms of four cases, *viz.* true positive, false positive, true negative, and false negative:

	The algorithm reported a prediction	**The algorithm did not report a prediction**
A functional element exists	True Positive (TP)	False Negative (FN)
No functional element exists	False Positive (FP)	True Negative (TN)

Any measure of prediction performance is derived fundamentally in terms of the number of TP, TN, FP and FN. Important measures are as follows:

Performance measure	**Definition**	**Physical interpretation**
Sensitivity (*Se*)	$Se = \frac{TP}{TP + FN}$	probability that a prediction is reported given the functional element is present
Specificity (*Sp*)	$Sp = \frac{TN}{TN + FP}$	probability that a prediction is not reported given the functional element is absent
Positive Predictive Value (*Ppv*)	$Ppv = \frac{TP}{TP + FP}$	probability that a functional element exists given that a prediction is reported
Negative Predictive Value (*Npv*)	$Npv = \frac{TN}{TN + FN}$	probability that a functional element does not exist given that no prediction is reported
Correlation coefficient (CC)	$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$	strength of the relationship between predictions and actual occurrences

All of the above measures range between 0 to 1 and their high values are desirable.

Apart from the above quantitative measures, an important measure of accuracy of a prediction algorithm is graphically represented in terms of the Receiver Operating Characteristics Curve (ROC Curve). While the above quantitative measures reflect only the current prediction accuracy of the predictor, the ROC curve shows a complete rigorous picture of the goodness of the prediction model. Although both high sensitivity and specificity are desirable, unless the predictor is perfect there is always a tradeoff between them. This is because the predictor may be either too liberal or too conservative in reporting positive predictions depending upon the threshold value it uses. The ROC curve shows this tradeoff under varying threshold values. It is the plot of sensitivity vs. (1–specificity) as the predictor threshold varies.

As shown in Figure III-7, the perfect predictor yields a point in the upper left corner (coordinate (0,1)) of the ROC space. Whereas the ROC curve of a completely random predictor is the 45º diagonal line. For a mundane predictor, the ROC curve lies somewhere above the 45º diagonal, and the further away this curve is from the diagonal the better the predictor’s performance.

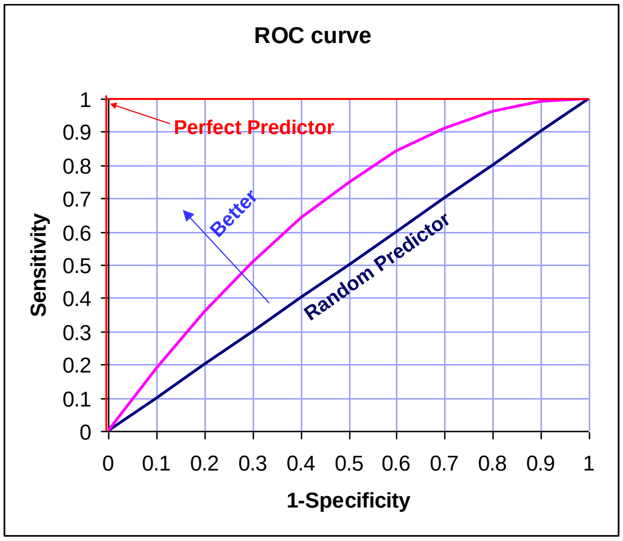


Figure III-7. The Receiver Operating Characteristics (ROC) curve.

CHAPTER - IV

DETECTION OF LOCALIZED MOTIFS

This chapter discusses the problem of detecting localized DNA motifs. The use of positional information yields significant advantage in this application. Motivation for the problem was given in Chapter 1. The problem is relevant towards motif discovery in long regulatory sequences that have been aligned relative to an anchor point, especially for genomes of higher eukaryotes (metazoans). The localized motif finding problem is defined in Section IV-1. A new scoring measure called *spatial confinement score* is introduced in Section IV-2, which allows assessment of whether or not a motif has localized occurrence within the sequences and an accurate demarcation of the interval of localization. The spatial confinement score is combined with existing scoring measures including motif over-representation and relative entropy in Section IV-3 to give an overall account of the goodness of a motif. The existing scoring measures have been reformulated in a form that the different scores can be easily combined into a single score and compared across motifs of different lengths and mutations. This allows selection of the most relevant motifs among candidates of different lengths, mutations and in different sequence intervals, and removal of redundant motifs. A time and memory efficient algorithm is developed in Section IV-4 to utilize the scoring function to detect motifs in long regulatory sequences. Experiments on simulated and real datasets reported in Section IV-6 show that LocalMotif can automatically detect localized motifs and accurately identify their position interval of localization in long sequence datasets. Such motifs can be detected by other motif finding algorithms only when the search is restricted to the relevant interval. The localization interval information provided by

LocalMotif is useful for the biological identification of motifs and for studying the composition of gene regulatory sequences.

IV-1 Problem Definition

The motif finding problem is well-defined in the literature. In the definition by Pevzner and Sze (2000), a set of N DNA sequences $\mathbf{S} = \{S_1, S_2, ..., S_N\}$ is given in which instances of an unknown pattern M of length l appear at different unknown positions. The instances of M in the sequences are not exact but mutated, with up to a maximum of d point substitutions. The problem is to discover the pattern M given l and d. The pattern M is called the *motif* and each of its instances in the sequences is called a *binding site*.

Note that the above definition uses consensus (*l*,*d*) representation of a motif. Though the PWM representation is usually preferred for motifs with known binding preferences, for *ab-initio* motif finding the consensus representation is found effective in detecting motifs, especially ones that do not have an exact occurrence in the sequence [Keich and Pevzner (2002a), Pevzner and Sze (2000)]. Recent benchmark assessment of different motif finding algorithms [Tompa et al. (2005)] confirms competent performance of consensus based algorithms such as Weeder [Pavesi et al. (2001)] and YMF [Sinha and Tompa (2003)].

The above definition considers that instances of the motif may be present anywhere across the complete sequences length, which is true for most short sequence datasets. Localized motif finding however considers that in long sequences, a significant proportion of the motif instances are found confined within a local sequence interval relative to an anchor point. The localized motif finding problem is thus stated as a variation of the above definition:

Given is a set of N input DNA sequences $\mathbf{S} = \{S_1, S_2, ..., S_N\}$ of length L each, aligned relative to an anchor point A as shown in Figure IV-1. The instances of an unknown pattern M of length l, mutated up to a maximum of d point substitutions, occur confined within an unknown interval (p_1, p_2) of the sequences. The aim now is to discover both M and (p_1, p_2) given $\mathbf{S}$, l and d.

The following sections present an algorithm called *LocalMotif* as a solution to the localized motif finding problem. Sections IV-2 and IV-3 describe the LocalMotif scoring function, while Section IV-4 describes the algorithm.

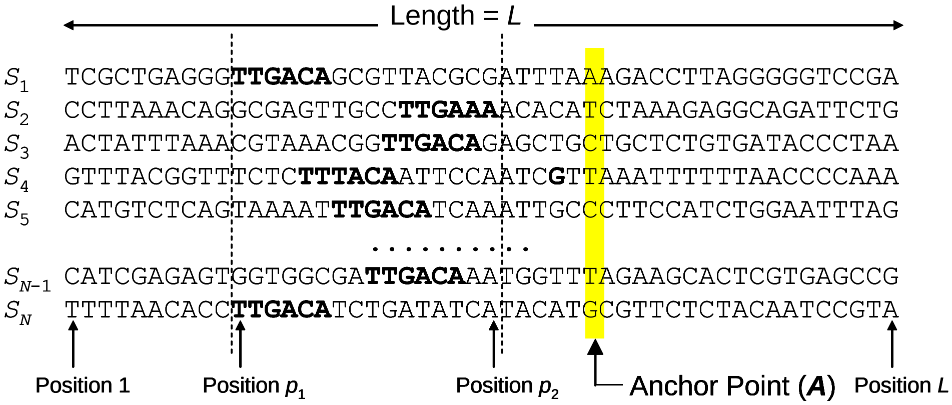

Length = L
S1 TCGCTGAGGGTTGACAGCGTTACGCGATTTAAAGACCTTAGGGGGTCCGA
S2 CCTTAAACAGGCGAGTTGCCTTGAAAACACATCTAAAGAGGCAGATTCTG
S3 ACTATTTAAACGTAAACGGTTGACAGAGCTGCTGCTCTGTGATACCCTAA
S4 GTTTACGGTTTCTCTTTACAATTCCAATCGTTAAATTTTTTAACCCCAAA
S5 CATGTCTCAGTAAAATTTGACATCAAATTGCCCTTCCATCTGGAATTTAG
. . . .
SN−1 CATCGAGAGTGGTGGCGATTGACAAATGGTTTAGAAGCACTCGTGAGCCG
SN TTTTAACACCTTGACATCTGATATCATACATGCGTTCTCTACAATCCGTA
Position 1
Position p1
Position p2
Anchor Point (A)
Position L

Figure IV-1. Discovering (6,1) motifs within a set of N sequences $S_1, S_2, \ldots, S_N$ each of length L. The random pattern TTTAAA is seen to eclipse the real motif TTGACA.

IV-2 Scoring Function

The LocalMotif scoring function includes three different independent measures of the goodness of a motif, *viz.* relative entropy score (RES), over-representation score (ORS) and spatial confinement score (SCS). While the former two scoring measures exist in the literature, spatial confinement score has been introduced in LocalMotif to aid the detection of localized motifs. All scoring measures are brought to a consistent and normalized form as entropy measured relative to a suitable basis, so that they may be combined together and are comparable across motifs with different $\left(l,d\right)$. Detailed derivations of the formulae are provided in the Appendices.

IV-2.1 Relative entropy score

The general nucleotide composition of the regulatory sequences is called background. The TFBS are expected to be distinct from the background since the TF can distinguish them from surrounding nucleotide patterns. Relative entropy score (RES) [Hertz and Stormo (1999), Stormo (2000), Thijs et al. (2002)] measures the difference between the motif model *M* and background model *B*. Let all observed TFBS of the motif be aligned vertically, and the average frequency of occurrence of each nucleotide $b \in \left\{A,C,G,T\right\}$ at each position $i = 1, 2, \ldots, l$ be $f_{b,i}$. The entropy of the motif *M* relative to the background model *B* is usually measured as the Kullback-Leibler divergence $D\left(M \,||\, B\right)$:

$$\text{Relative entropy score (RES)} = D\left(M \,||\, B\right) = \sum_{i=1}^{L}\sum_{b} f_{b,i} \ln\left(\frac{f_{b,i}}{p_b}\right) \tag{4.1}$$

where p_b are the *a priori* frequencies of the nucleotides according to the background model. The background model in LocalMotif is a Markov model of user-defined order *q* [Thijs et al. (2002)]. The expression for RES is normalized as described in the Appendix B. The normalized score is given by

$$D_{norm}\left(M \parallel B\right) = \frac{1}{l\ln 4}\sum_{i=1}^{l}\sum_{b} f_{b,i}\ln\left(f_{b,i}\right) - \frac{1}{\ln 4}\sum_{b}\overline{f}_b \ln\left(p_b\right), \tag{4.2}$$

where $\overline{f}_b = \frac{1}{l}\sum_{i=1}^{l} f_{b,i}$. The normalized RES usually lies in the range (0,1) and is independent of the motif length *l*.

IV-2.2 Over-representation score

Since the motif is enriched in the input sequences, its number of instances in the sequences must be significantly greater than that expected by chance (according to the background). The over-representation score is a statistical measure of the deviation between number of observed and chance occurrences. In random sequences that have been sampled from a Markov background model, the number of chance occurrences of a motif approximately follows the Gaussian distribution. The Z-score can thus be used to measure the statistical difference between the observed and expected number of instances of a motif [Tompa (1999)]. It is given by the formula:

$$Z - score = \frac{\left(n/(NL)\right) - e}{\sigma}, \tag{4.3}$$

where n is the number of observed instances, N is the total number of input sequences, L is the average length of an input sequence, e is the probability of generating an instance of the motif according to the background model, and σ is the standard deviation for the sampling distribution of e given as $\sigma = \sqrt{e(1-e)/(NL)}$. However the Z-score is not comparable across motifs with different $\left(l,d\right)$. An entropy measure for over-representation is thus derived here. The Gaussian distribution is a large sample approximation of the original binomial distribution according to the central limit theorem. The normalized entropy measure is derived beginning with the binomial distribution.

Consider the experiment of finding all the instances (TFBS) of a (l,d) motif *M* in a set of DNA sequences. Among all nucleotide patterns of length *l*, let the proportion of TFBS patterns be *e*, while the proportion of non-TFBS patterns be $(1-e)$. Among *n* observed patterns, the probability of observing *k* TFBS is given by the binomial distribution $P(k,n|e) = {}^{n}C_{k}(e)^{k}(1-e)^{n-k}$. Now let the estimated proportion of TFBS of a motif according to the background be e_0, and let the actual observed proportion be e_1. The over-representation is measured as the Kullback-Leibler divergence between the binomial distributions $P(k,n|e_0)$ and $P(k,n|e_1)$. The expression is explained in detail in Appendix B.

$$\text{Over-representation score}(\text{ORS}) = D(E_0 \,||\, E_1) = N(l,d)\left[e_0 \ln\left(\frac{e_0}{e_1}\right) + (1-e_0)\ln\left(\frac{1-e_0}{1-e_1}\right)\right], \qquad (4.4)$$

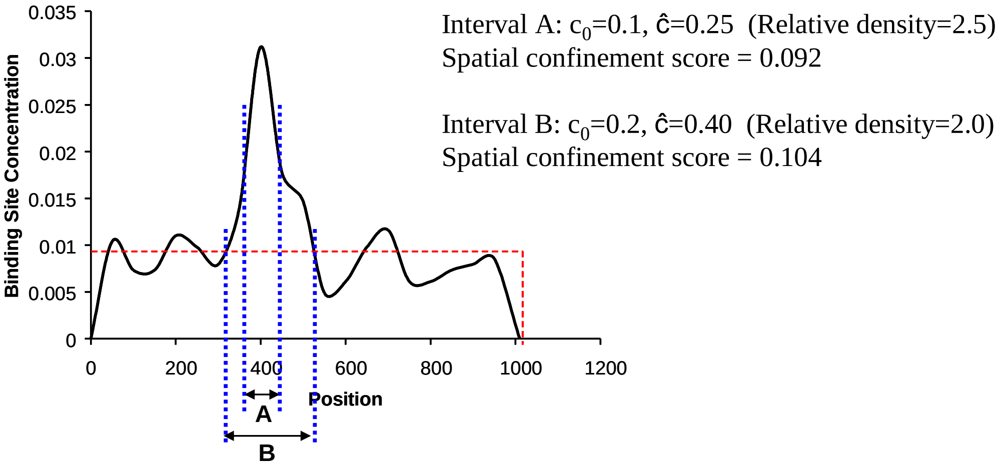
where $N(l,d)$ is a normalization factor described in Appendix B.

IV-2.3 Spatial confinement score

In computational motif finding algorithms, usually the TFBS are considered as randomly distributed across the entire sequence length. LocalMotif however considers that the distribution of TFBS may be non-uniform and localized in a certain interval (p_1, p_2) of the sequences which have been aligned relative to an anchor point. Let *c* denote the proportion of TFBS that fall within a sequence interval (p_1, p_2), i.e., if *n* is the total number of TFBS across entire sequence length *L*, and n_1 is the number of TFBS in the interval (p_1, p_2), then $c = n_1/n$. If the TFBS are uniformly distributed across the entire sequence length *L*, then it is expected that the proportion of TFBS falling within any interval (p_1, p_2) will be $c = c_0 = |p_2 - p_1|/L$. For example, in any interval of length $L/2$ one would expect to find 50% of the TFBS. However if the TFBS distribution is non-uniform, the proportion would be higher in some intervals and lower in others. LocalMotif intends to discover the shortest interval that encompasses the maximum proportion of TFBS. It thus compares the proportion of TFBS that lies within the interval and the proportion that lies outside it. The interval which maximally separates the two has the highest *spatial confinement score*. Let $\hat{c}$ be the observed proportion of TFBS that lie within an interval (p_1, p_2) and $(1-\hat{c})$ that lie outside it. Let the corresponding proportions according to uniform distribution be c_0 and $(1-c_0)$. The spatial confinement score for the interval is given by the entropy difference (KL-divergence) between the observed and uniform proportions. Its mathematical definition and derivation is presented in Appendices A and B.

$$\text{Spatial confinement score (SCS)} = D(\hat{c} \,||\, c_0) = \hat{c}\ln\left(\frac{\hat{c}}{c_0}\right) + (1-\hat{c})\ln\left(\frac{1-\hat{c}}{1-c_0}\right). \quad (4.5)$$

Note that a short interval with high density of TFBS may not have a spatial confinement score as high as a longer interval with slightly lesser density of TFBS if the longer interval encompasses a large proportion of the TFBS compared to its surroundings. For example, in Figure IV-2, the score for interval B is higher than for interval A.



Interval A: c_0=0.1, $\hat{c}$=0.25 (Relative density=2.5)
Spatial confinement score = 0.092

Interval B: c_0=0.2, $\hat{c}$=0.40 (Relative density=2.0)
Spatial confinement score = 0.104

Figure IV-2. Illustration of how spatial confinement score finds the shortest interval encompassing the maximum proportion of TFBS – though interval A has higher density of TFBS, its score is lower since a large proportion of TFBS still lie outside it.

IV-3 Combined score

The three scoring measures mentioned above, *viz.* relative entropy score (RES), over-representation score (ORS) and the spatial confinement score (SCS) measure three completely independent characteristics of a motif. All of them have been expressed in the form of an entropy measure based on KL divergence between an observed and a reference probability distribution. The score of a motif is thus independent of situational parameters such as motif length *l*, number of allowed substitutions *d*, sequence length *L*, interval length $\left(p_1, p_2\right)$, and so forth. Being in a normalized form, the scores usually range between (0,1) and have consistent values barring extreme situations such as erroneous measurement of the background distribution. The combined score may be computed by the Hamming measure, which is simply a sum of the three different scores, or Euclidean measure, which is the root mean square of the three scores. In addition the individual scores give a meaningful description of what characteristic of a particular motif makes it more favored. An example in this relation is presented in Section IV-6.1 below.

Combined score:

$$\text{Hamming measure} = RES + ORS + SCS \tag{4.6}$$

$$\text{Euclidean measure} = \sqrt{RES^2 + ORS^2 + SCS^2}$$

IV-4 Algorithm

The LocalMotif algorithm must score candidate motifs in different sequence intervals and report the best scoring ones. An exhaustive enumeration strategy would require scoring all possible 4^l candidate patterns in all possible sequence intervals, leading to a complexity of $O\left(4^l.l^2\right)$. One of the objectives of LocalMotif is fast processing of long sequence datasets. The algorithm therefore includes several optimizations which are briefly explained below. The algorithm pseudocode is presented in .

IV-4.1 Creating a positional dictionary

Positional dictionary optimizes computation of the number of instances of a candidate pattern in a given sequence interval. All unique length *l* sub-strings (*l*-mers) found in the input sequences form the different entries of the dictionary. The position of every single occurrence of each *l*-mer is recorded in this dictionary. Occurrences of overlapping identical patterns are excluded, e.g., if the string “TATATATA” occurs in an input sequence, and 4-mer patterns are of interest, then the dictionary entry “TATA” will record only two occurrences instead of three. The dictionary is cross-referenced so that entries whose *l*-mer patterns have a Hamming distance of *d* or less from each other are interlinked. Interlinking facilitates quick enumeration of all binding site occurrences for every *l*-mer candidate.

```
INITIALIZATION
    Build a dictionary of all $l$-mers found within the sequences and their occurrence
    positions, and link $l$-mers having a Hamming distance ≤ $d$ from each other.

FIRST PASS
    FOR M=all $l$-mers in the dictionary:
      FOR p1 = 0 to $L$ with step $s$:
        Compute the number of binding sites of M in the interval (p1, p1+$s$).

SECOND PASS
    FOR M=all $l$-mers in the dictionary:
      FOR p1 = 0 to ($L$ – $s$) with step $s$:
        FOR p2 = (p1+minsize) to (p1+maxsize) with step $s$:

          Using the values in intervals (p1,p2–$s$ ) and (p2–$s$ , p2 ), compute for the
          interval (p1,p2 ) the variables $n_1$, $n_0$, e, $\sigma$, $c_0$, $\hat{c}$ . Thus compute score for the
          interval (p1,p2).

          DISCARD SIMILAR PATTERNS
              FOR all stored motifs M' in the list:
                IF M is similar to M' AND (p1,p2) overlaps (p1',p2') :
                  IF score of M < score of M' THEN discard M and retain M' ELSE
                  retain M and discard M'

EXTEND MOTIF SEARCH
    Perform clustering and majority pattern generation.
    Add majority pattern to the dictionary and score it in all intervals as per above
    steps.
    Repeat the extension till the average score stops increasing.

OUTPUT THE TOP SCORING MOTIFS AND THEIR POSITION INTERVALS.
```

Figure IV-3. The LocalMotif algorithm.

IV-4.2 Speed-up for score computation

Scoring each candidate *l*-mer in all possible position intervals $(p_1, p_2): 0 \le p_1 < p_2 \le L$, would be formidable. Only the intervals $(p_1, p_2):\ p_1 < p_2;\ p_1, p_2 \in \{0, s, 2s, 3s, \ldots, L\}$ are considered, where *s*, called step size, is a small integer value set to 5 in the current implementation. Interestingly the scoring function need not be determined individually for each position interval. The score for a longer interval can be computed directly from the scores for shorter constituent intervals. The relations are derived in Appendix C. Computations are thus performed over two passes – scores for all length *s* intervals are computed in the first pass, and scores for longer intervals are calculated directly from the scores for shorter constituent intervals in the second pass. The bottleneck in score computation is the first pass, so direct computation in second pass reduces the time complexity in sequence length.

IV-4.3 Early discarding of similar patterns

While the candidate *l*-mers are being scored over various intervals, a list of scores is maintained sorted in a descending order. Lower scoring *l*-mers having similar pattern and overlapping position intervals with higher scoring *l*-mers are deleted from the list of possible motifs to maintain only the top η motifs, where η is a user-defined percentage of the total number of candidate motifs. This limits the memory requirements of the algorithm. Similarity between two *l*-mers is evaluated using the Needleman-Wunsch global alignment algorithm (with possible gaps). The alignment score threshold, A_{thresh}, for measuring the similarity is a function of *l*.

IV-4.4 Extending the motif search

The LocalMotif algorithm does not perform an exhaustive search over all possible 4^l *l*-mer patterns to seek the best motifs. Initially only the *l*-mers occurring directly within the input sequences are considered as candidate motifs. It is possible that *l*-mers not occurring directly within the input sequences may be the best motifs. A heuristic algorithm similar to SP-STAR [Pevzner and Sze (2000)] extends the search over other probable patterns. The η best scoring *l*-mers are clustered according to the goodness of their alignment, so that each cluster $M_{clus} = \{M_1, M_2, \ldots, M_m\}$ contains similar patterns of length *l*. A majority pattern is computed for each cluster, whose i^{th} letter is the most frequent i^{th} letter in M_{clus}, with ties broken arbitrarily [Pevzner and Sze (2000)]. The majority pattern of each cluster is a new candidate motif. The new generation of candidate motifs is added to the cross-referenced positional dictionary and scored in all sequence intervals. Best scoring η candidate motifs are again selected and the clustering and majority pattern procedure is repeated until scores of a new generation do not show any improvement over previous generations.

IV-4.5 Combining motif candidates with different (l,d) combinations:

In each run, the LocalMotif algorithm finds motifs for a fixed value of *l* and *d*. To combine the results of separate runs of LocalMotif with varying (l,d), a post-processing algorithm has been written. Since the LocalMotif scoring function does not depend upon *l* and *d*, motifs with different *l* and *d* can be directly compared in their scores. Motifs with similar pattern are again identified by alignment using the Needleman-Wunsch algorithm and among a pair of motifs with greater than 65% similarity (measured relative to the shorter motif), the one with lower score is discarded. If two motifs have high (>90%) similarity and overlapping intervals of localization, they are combined into a single motif taking union of their intervals. The automated algorithm performs runs for the (l,d) combinations (6,1), (7,1), (8,1), (9,2), (10,2), (11,2) and (12,3), and combines their results. The particular (l,d) combinations have chosen in accordance with the recommendation of [Buhler and Tompa (2002)], with maximum possible *d* for a given *l* that avoids random motifs.

IV-5 Implementation

The basic LocalMotif algorithm has been implemented in C++, and is supplemented by a user-friendly interface and post-processor written in Python. The complete source code and compiled binaries for both Unix and Windows platforms are available freely at the website http://www.comp.nus.edu.sg/~bioinfo/LocalMotif. Following parameters can be controlled by the user to suit the requirements of the particular dataset and the available computing resources:

- *Specification of the background model*: the user can choose both the order of the background Markov model and the way of specifying its parameters. The background model parameters can either be directly specified or the user can provide a set of sequences from which the program automatically learns the parameters.
- *Number of motifs to be retained in memory*: this is the parameter η described in Section IV-4.3. Larger value of η is better for extension of the motif search, but the tradeoff is RAM requirements,
- *Maximum interval length*: in the analysis of long sequence datasets, setting a maximum interval length (such as 1000 bp) makes the analysis not only faster but also more accurate since the motifs will become subtle for longer interval lengths,
- *Number of best motifs to output*, and
- Choice of *single or double strand analysis*.

The program outputs the discovered motifs, the interval of localization of each motif, the three individual scores (RES, ORS and SCS), and combined score of each motif. The individual scores reveal the prominent characteristics of a motif and may be used to reject outliers (e.g. a motif reported with large RES and SCS but small ORS is probably a noisy pattern). In addition, details of the intermediate processing such as scores of patterns similar to each selected motif (Section IV-4.3) and motif extensions (Section IV-4.4) are written to a separate file for reference of the specialist.

IV-6 Results

The scoring function of LocalMotif is demonstrated first in Section IV-6.1. Then the performance of LocalMotif is reported over sequences of different lengths in both synthetic datasets (Section IV-6.2) and real datasets (Section IV-6.3). Comparison is made with two other freely available motif finding tools: MEME [Bailey and Elkan (1994)] and Weeder [Pavesi et al. (2001)]. MEME is one of the most commonly used motif finding tools due to its robustness and simplicity, while Weeder has been reported as one of the best motif finding tools by [Tompa et al. (2005)]. Also, while Weeder uses the (l,d) motif model, MEME is based on the positional weight matrix (PWM) model [Stormo (2000)].

IV-6.1 Analysis of the scoring function

The LocalMotif scoring function is illustrated through a planted (l,d) motif problem. A dataset consisting of 50 sequences, each of length 3000 bp, was generated using a zero-order uniform Markov background model. Instances of a length 7 pattern, ATGCATG, mutated with two base substitutions each were randomly implanted in 75% the sequences as a (7,1) motif. The sequences were analyzed with LocalMotif for (7,1) motifs. The motif instances were confined to lie within the position interval (2000, 2500). Note that as per the analysis of [Buhler and Tompa (2002), Keich and Pevzner (2002a)], the (7,1) motif is a subtle motif impossible to detect within the 3000 bp length sequence as there are at least 6600 competing random motifs. However it is possible to discover in the localized 500 bp region as within this region it is not subtle with no competing random motifs. Five top scoring motifs reported by LocalMotif and their scores are shown in Table IV-1. The planted (7,1) pattern was correctly identified as the top motif and its interval of localization was accurately determined. Although the localized motif has a low ORS compared to several competing random patterns, it has a substantially higher spatial confinement score (SCS) of 0.485 as compared to the spurious motifs whose SCS is less than 0.3. LocalMotif assigned a higher total total score to the localized motif due to its high spatial confinement. Over-represented random motifs are not expected to be spatially confined.

Table IV-1. Results of using LocalMotif to analyze simulated sequences of length 3000 bp containing a planted (7,1) motif ATGCATG – five top scoring motifs and their predicted localization intervals are reported.

Motif pattern	**Motif interval**	**Motif score**	**Score components**		
			RES	**ORS**	**SCS**
ATGCATG	(2060,2445)	1.308	0.497	0.326	0.485
GGACGCT	(15,115)	1.216	0.481	0.500	0.235
AGCGCCG	(455,575)	1.209	0.481	0.439	0.289
GTCCGAT	(85,200)	1.173	0.482	0.408	0.282
TCCCTGC	(2340,2450)	1.167	0.481	0.411	0.275

A contour plot of the over-representation score (ORS), localization score (SCS) and the total score (SCO) for the (7,1) motif in various position intervals is shown in . Note that the relative entropy score of the motif does not depend upon the interval being analyzed and is therefore not shown. The ORS contours show that this score is large wherever there is a local concentration of the binding sites. Thus several short sub-intervals within the region (2000,2500) have a large ORS. Whereas the SCS contrours show that SCS is large only in the actual interval of localization of the motif. The variation in ORS values is lower compared to the variation in SCS values. Thus the total score (SCO) contours, which is a sum of ORS, SCS and RES, is biased towards SCS variations and is thus maximum at the actual interval of localization, i.e., (2000,2500). The spatial confinement score thus plays an important role in the detection of localized motifs and their accurate intervals of localization.

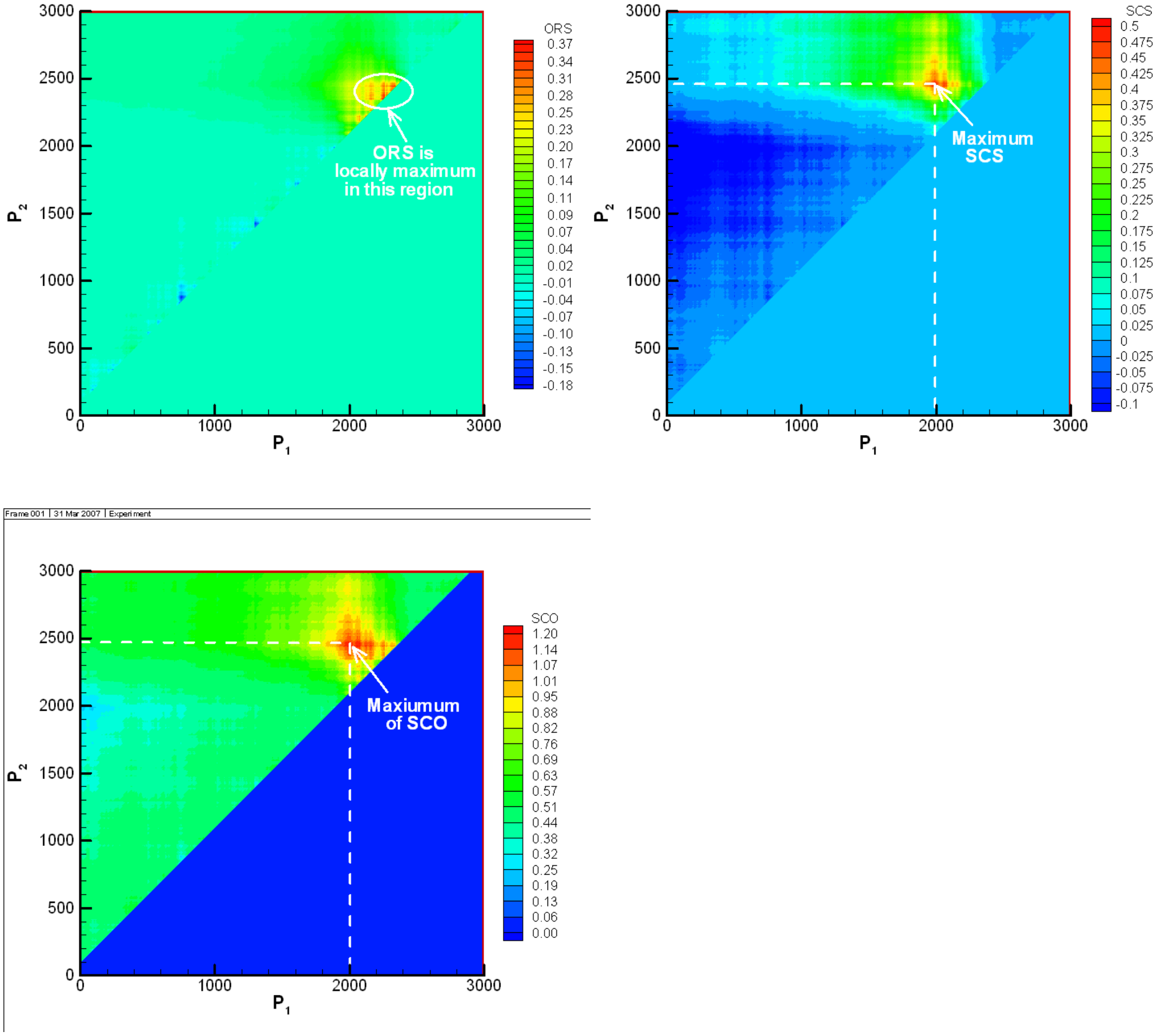


Figure IV-4. Contours showing (a) the total score, (b) over-representation score, and (c) spatial confinement score of the motif ATGCATG in different position intervals (p_1,p_2) of the planted motif sequences.

IV-6.2 Performance on Simulated datasets

IV.6.2.1 Short sequence datasets

The test on simulated short sequence datasets evaluates the accuracy and robustness of motif detection as well as the accuracy of localization interval predictions made by Localmotif. Each dataset consists of N nucleotide sequences, each of length $L < 1000$, generated from a background Markov model of order q. Some of the sequences are implanted with an instance of a (l, d) motif M within a local position interval, $I = (p_1, p_2)$.

A total of 100 such datasets were generated while randomly varying the following parameters: (i) sequences length *L*, (ii) percentage of sequences, *k*, that contain an instance the motif (iii) distinctness of the motif from sequence background (i.e. relative entropy) (iv) ratio of interval length (in which the motif is confined) to sequence length, $\bar{p} = |I|/L$. Note that $|I|$ denotes length of the interval *I*, which equals $(p_2 - p_1)$. All these parameters, together with *l*, *d*, motif pattern *M*, and the background model, were varied randomly to simulate a fair variety of test conditions. The ranges of parameter values studied is given in Table IV-2.

Table IV-2. Ranges of parameters studied in simulated short sequence datasets.

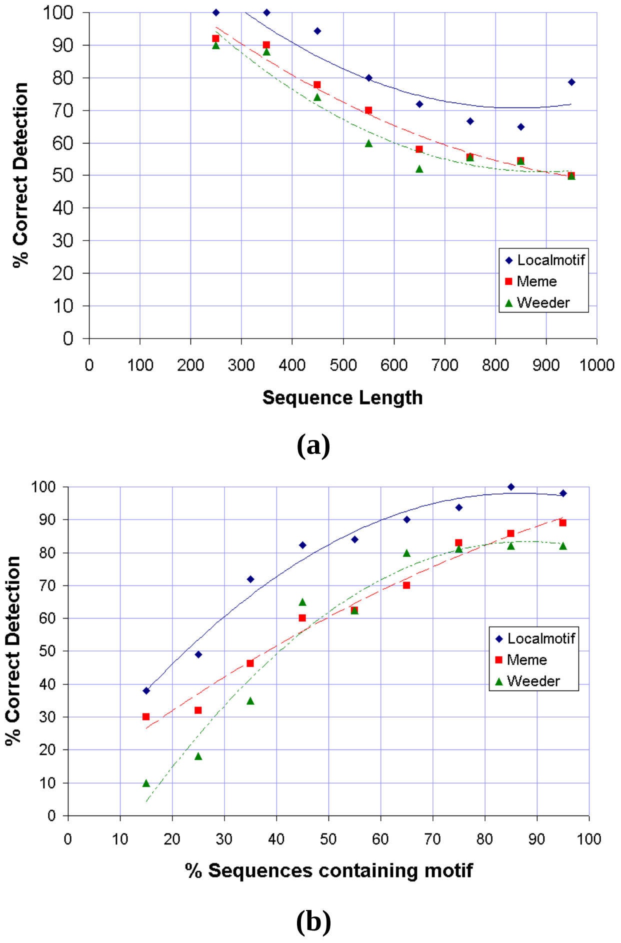
Parameter	*N*	*L*	*q*	*k*	(*l*,*d*)	$\bar{p}$
Range	50-100	200-1000	0-2	20-100	(6,1)-(10,3)	10-100

Figure IV-5 shows the performance of motif detection with (a) varying sequence length, *L*, and (b) varying percentage, *k*, of sequences that contain a motif instance. The motif becomes increasingly subtle with increasing *L* or decreasing *k* since the number of competing random patterns increases. Figure IV-5 indeed shows a diminishing performance of motif detection with increasing *L* or decreasing *k* for all the tested motif finding tools. However the accuracy is observed to be consistently higher for LocalMotif as compared with MEME and Weeder. This is because LocalMotif's performance is dependent on the localization interval length rather than the total sequence length. The localized search reduces the number of competing random patterns and increases the comparative motif signal strength. Thus LocalMotif has greater accuracy for long sequences that contain a localized motif. However for datasets where motifs are not localized, the comparatively higher accuracy of LocalMotif may not hold.

The length and position of the interval within which the motif is localized has been varied randomly in the simulated datasets to test whether LocalMotif can correctly ascertain this interval. The accuracy of predictions has been measured in terms of the percentage of overlap between the actual interval, I_a , and predicted interval, I_p . Precisely,

$$\text{overlap percentage} = \frac{|\, I_a \cap I_p \,|}{max(|\, I_a \,|,|\, I_p \,|)}. \tag{4.7}$$

The mismatch in predicted and actual interval lengths is also penalized in this formula by taking the ratio with respect to the larger interval. As seen in Figure IV-6, LocalMotif determined the position interval very accurately (overlap ≥ 0.8) in more than 60% of the cases. This shows the effectiveness of the spatial confinement scoring function used to in LocalMotif to determine the localization interval.



(a)

(b)

Figure IV-5. Performance of MEME, Weeder and Localmotif in simulated short sequence datasets with (a) varying sequence length, L , (b) varying percentage, k , of sequences containing motif instances.

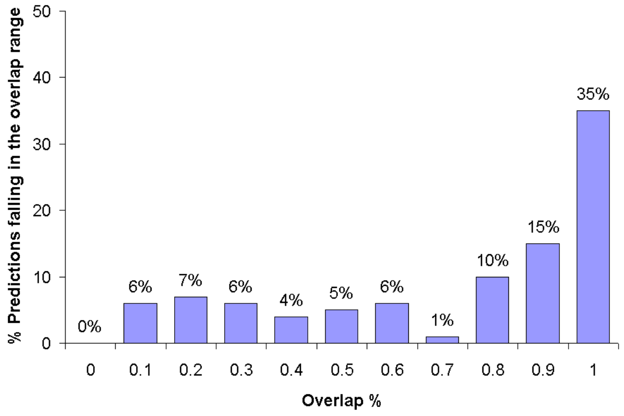


Figure IV-6. Accuracy of LocalMotif's interval predictions.

IV.6.2.2 Long sequence datasets

Each long sequence synthetic dataset consisted of 50 sequences of length 1000-5000 bp each, implanted with instances of one to five (*l*,*d*) motifs in position intervals of width 200-600 bp. Ten such datasets were generated with randomly chosen number of motifs, motif patterns, sequence length, localization interval and background model.

For prediction of localized motifs using MEME and Weeder, each dataset was split into smaller fragments. The fragments were of the same length and overlapped each other by 50% to ensure that the interval of localization is not missed due to improper positioning of the fragment boundaries. Also three different fragment lengths, 200, 400 and 600, were tried. Each fragment was individually analyzed using MEME and Weeder. Results obtained on the individual fragments were pooled together and sorted according to the score value in case of Weeder and expect value (E-value) in case of MEME.

LocalMotif, on the other hand, was run directly on the long sequence datasets with a maximum interval length prescribed as 1000. Sequences fragmentation was not required since LocalMotif automatically determines the motif's interval of localization.

For each program, the top ten reported motifs were retrieved for each dataset. The accuracy of motif detection was measured as sensitivity (Se), i.e., percentage of actual motifs successfully detected. Specificity or positive predictive value could not be measured since each program only reported ten best candidates and thus it is hard to give a definition of false positive. Table IV-3 shows that MEME, Weeder and LocalMotif could determine 56%, 50% and 81% of the localized motifs respectively in intervals with non-zero overlap percentages, and thus LocalMotif was found to be most accurate.

Table IV-3. Accuracy of motif detection in synthetic long sequence datasets.

Program	Planted motifs	Correct predictions	Sensitivity
MEME	32	18	56%
Weeder	32	16	50%
LocalMotif	32	26	81%

IV-6.3 Performance on Real datasets

IV.6.3.1 Short promoter sequences surrounding the TSS

Metazoan promoter sequences immediately surrounding the TSS usually contain a few highly conserved core promoter motifs. An example of computational motif discovery in such sequences is in the set of 1941 Drosophila promoter sequences of length 300 bp each (aligned -250 to +50 relative to the TSS) compiled by [Ohler et al. (2002)]. Ohler et al. (2002) had used MEME to determine the core promoter motif content of these sequences. They performed two separate runs of MEME, one over full length (300bp) sequences, and the other over a sub-interval -60 to +40 relative to the TSS.

The full length (300bp) sequences have been examined with LocalMotif. Background Markov model of order 2 was learnt from a set of 361 Drosophila intron sequences. Weeder could not process this dataset due to its large size. Results compiled in Figure IV-7 show that MEME discovered prominent core promoter motifs only when analyzing the -60 to +40 sub-region, whereas LocalMotif could detect them given the full 300 bp region. LocalMotif additionally reported accurate localization intervals of the motifs, which is useful in their identification, e.g. the downstream promoter element (DPE) is confirmed as it is found in the +25 to +45 interval. Moreover, it is observed in Figure IV-7 that all biologically meaningful motifs reported by LocalMotif have a higher

SCS of 0.14 or more as compared to the two spurious motifs (at positions 8 and 9) whose SCS is less than 0.06. Thus spatial confinement score additionally allows discarding of spurious motifs.

LocalMotif Results (-250 to +50)						
Rank	***Motif***	***Score***	***RES***	***SCS***	***ORS***	***Position***
1	TCAGTC	1.920	0.420	0.500	1.000	[-5,+15] → Initiator
2	GTCACACT	1.382	0.430	0.233	0.719	[-10,+20] → new motif
3	CTATAAAA	1.275	0.350	0.153	0.772	[-35,-15] → TATA box
4	CAGTTG	1.266	0.423	0.172	0.671	[-5,+15] → Initiator
5	CGGACGTG	1.121	0.444	0.374	0.303	[+25,+45] → DPE
6	CTATCGAT	1.119	0.402	0.145	0.572	[-75,+0] → DRE
7	TCCGTT	0.934	0.411	0.146	0.377	[-5,+15] → Initiator
8	ATATATAT	0.895	0.324	0.026	0.544	[-205,-90]
9	CTCTCTCT	0.869	0.392	0.054	0.424	[-120,-70]
10	GCGTTCGG	0.866	0.424	0.153	0.289	[+10,+40] → DPE

MEME Results (-250 to +50)			**MEME Results (-60 to +40)**		
Rank	***Motif***	***Score***	***Rank***	***Motif***	***Score***
1	GGTCACACT	5.0e-369 → new motif	1	GGTCACACT	5.1e-415 → new motif
2	CTCTCTC	1.7e-203	2	TATCGATA	1.7e-183 → DRE
3	CGCCGCC	1.1e-151	3	TATAAA	2.1e-138 → TATA box
4	TTTTTTT	1.5e-155	4	TCAGTT	3.4e-117 → Initiator
5	TATCGATA	4.4e-78 → DRE	5	CAGCTG	2.9e-93
6	CAGCCTG	1.5e-80	6	GTATTTT	1.9e-62
7	GGCAACGC	1.4e-55	7	CATCTCT	1.9e-63
8	GTGTGTGT	6.4e-96	8	GGCAACGC	5.1e-29
9	TGCTTTTG	1.2e-39	9	GCGTGCGG	1.9e-12 → DPE
10	GCGCTTTAC	9.5e-24	10	CGAACGGAACG	8.3e-9

Figure IV-7. Motifs discovered by MEME and LocalMotif in Drosophila promoters.

IV.6.3.2 Short regulatory regions upstream of the TSS

LocalMotif was further tested for the detection of conserved motifs in sets of orthologous regulatory sequences upstream of the TSS for a single gene in several species. Motif detection in such datasets is known as phylogenetic footprinting. Standard methods for phylogenetic footprinting include (i) identification of conserved regions in a global multiple alignment of the sequences using a tool such as CLUSTALW [Thompson et al. (1994)], (ii) using existing motif finding programs such as MEME to detect conserved patterns, or (iii) using algorithms tailor-made for phylogenetic footprinting such as Footprinter [Blanchette and Tompa (2002)]. Using LocalMotif over such datasets is in one sense similar to using a motif finding program such as MEME since LocalMotif does not exploit the phylogenetic relationships between the sequence. However an important difference is that LocalMotif searches for conserved patterns in an aligned sub-interval of the sequences, which is meaningful due to the structural similarity among the orthologous sequences.

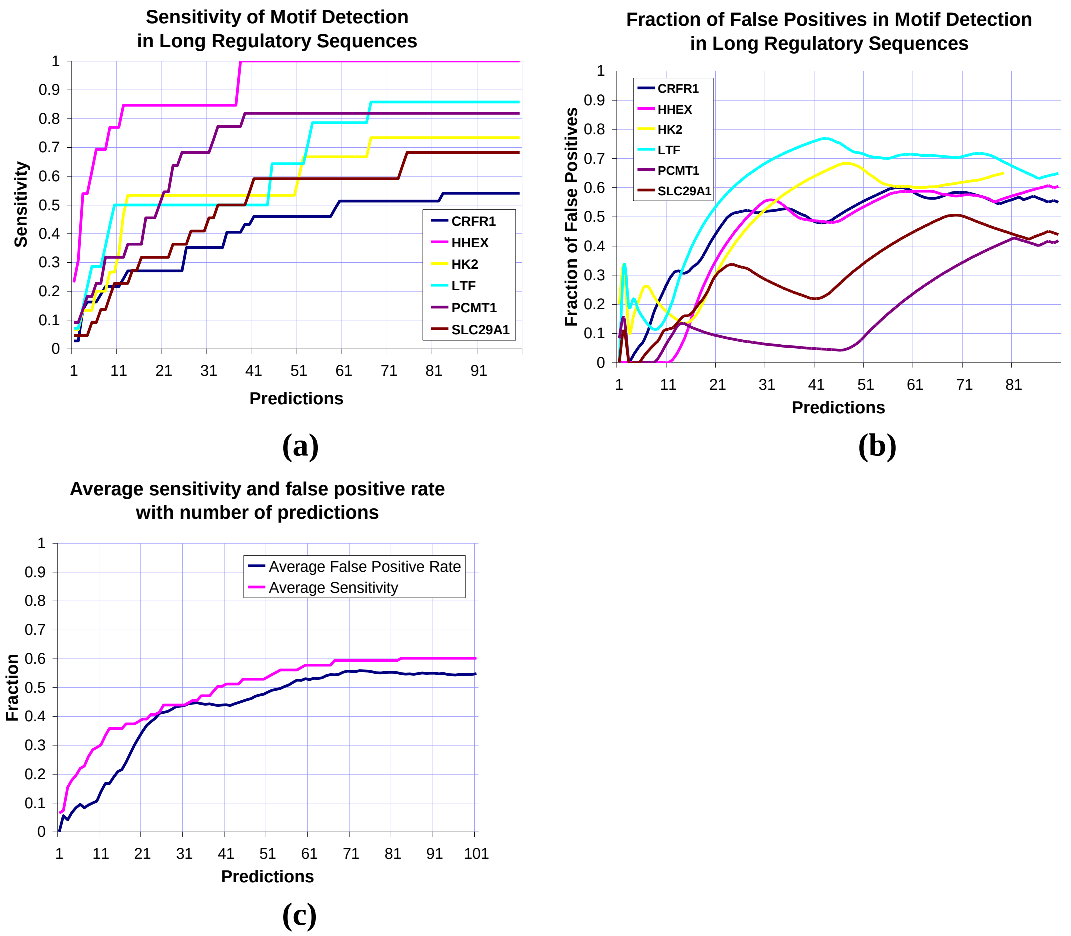
The test datasets in this study were derived from [Blanchette and Tompa (2002)] as both experimentally verified and computationally predicted conserved regulatory elements are available for reliable comparison of the results. There are 7 datasets, each containing 400-1000 bp long orthologous upstream regions (5' of the translation start site) of a single gene in the genomes of 5-20 different metazoan species. LocalMotif was used to analyze these datsets considering uniform Markov nucleotide background. The detailed results are shown in Supplementary Figure 1 at the end of this dissertation. In summary, LocalMotif discovered 46 out of the 49 motifs listed by [Blanchette and Tompa (2002)] accurately with their respective intervals of localization.

IV.6.3.3 Long regulatory segments upstream of the TSS

Datasets of experimentally characterized long regulatory sequences are scarce in the literature. However TFBS annotations in segments of ~1 kb length upstream of the TSS are available in the literature for several vertebrate genes. Some TFBS are experimentally validated while the rest are predicted in-silico using tools such as TRANSFAC [Matys et al. (2003)]. The annotations can be considered as high quality due to manual curation by field experts. Six datasets are compiled in the present study, each containing 3000 bp upstream sequences of a single gene in different vertebrate genomes, where either the human or mouse ortholog is characterized in the literature. The sequences are aligned relative to either the TSS or the translation initiation site, whichever is more reliably known. Comparison of motifs discovered by LocalMotif with the published TFBS annotations is summarized in Figure IV-8 and the details are provided in Supplementary Figure 2 at the end of this dissertation. Figure IV-8 shows how the sensitivity and false positive rate of TFBS detection varies as the number of top motifs reported by LocalMotif is increased. Here sensitivity is defined as the fraction of total known TFBS that could be predicted (i.e. True Positives / Total Positives), and false positive rate is defined as the fraction of reported motifs that are incorrect, i.e. which do not overlap any known TFBS. A sensitivity of 50% with a false positive rate of 44% is reached within the first 40 predictions, after which the sensitivity does not improve significantly. Among 122 annotated TFBS in the literature within the six datasets, one or more predicted motifs occurred within 87 (71%) TFBS. The localization intervals of the motifs as predicted by LocalMotif matched very well with the annotated TFBS. Considering the long length of the sequences being analyzed, the localization information was very useful for accurately locating the binding sites and led to a significant reduction in the number of false positives. Thus LocalMotif is promising for the identification of conserved motifs in long upstream regulatory regions of genes.

IV.6.3.4 Sequences flanking a known TFBS

The ERE dataset is an example of vertebrate sequences with wide spacing among regulatory elements and high degree of mutation in the binding sites. It contains 57 estrogen receptor (ER) target sequences from human chromosomes 21 and 22 discovered



(a)

(b)

(c)

Figure IV-8. Variation of sensitivity and false positive rate of Localmotif's predictions in long regulatory sequences upstream of the TSS as the number of predicted motifs is increased.

by ChIP analysis of in-vivo ER-chromatin complexes [Carroll et al. (2005)]. Almost all sequences lie distal from the TSS beyond the promoter region and have lengths ranging from 0.2 to 2.5kbp. About 34 ER full binding sites (length 15 bp, consensus AGGTCACCNTGACCT) have been mapped in this sequence set. Experimental studies have revealed binding sites for an associated factor called Forkhead (consensus TTGTTTNCTT) proximal to the ER binding sites [Carroll et al. (2005)].

To verify whether Forkhead binding adjacent to ER sites can be discovered in-silico, a new set of 34 sequences was prepared with one known ER full site in each sequence. The ER site acts as the anchor point. The positions of Forkhead binding sites relative to ER binding sites are shown in Figure IV-9. Results of motif finding (processing both strands) with MEME, Weeder and LocalMotif on this dataset are reported in Figure IV-10. Weeder was used with its default human background model, whereas human chromosome 21 and 22 intergenic sequences were used to prepare a zero order background for LocalMotif. LocalMotif reliably discovered the Forkhead motif with consensus TTTTTTTCTT, with about 60% of the true Forkhead sites found within the list of reported binding sites (refer Supplementary Figure 3 at the end of the dissertation). Thus, LocalMotif was found useful for discovering correlated motifs in vertebrate regulatory sequences.

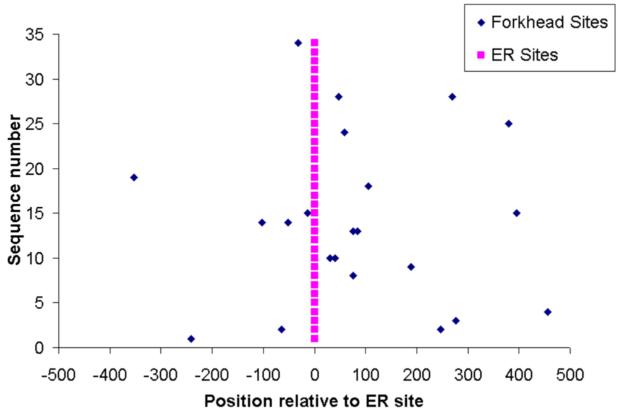


Figure IV-9. Distribution of forkhead binding sites relative to ER binding sites.

Software	***Motifs Predicted***
MEME	TC**AAGGTCA**G/C**TGACCT**TGA →**ER** A**GAGGGAAGA**/T**CTTCCCTC**T →**new**
Weeder	GT**TGACTTT**G/CA**AAGTCA**AC →**ER**
LocalMotif	**GGTCA**CCCTG/CAGGG**TGACC** [-20,+30] →**ER** **AAGAAAAAAA**/**TTTTTTTCTT** [-100,+300]→**FH** GG**GAGGGAAG**/**CTTCCCTC**CC [-190,+190]→**new**

Figure IV-10. Motifs discovered by MEME, Weeder and LocalMotif in ERE dataset.

IV-7 Conclusions

This chapter introduced a new algorithm called LocalMotif to detect motifs in localized intervals of long sequences (such as vertebrate regulatory sequences) aligned relative to a common anchor point. The algorithm uses a novel statistical scoring function to determine the interval of localization of the motif. It is optimized for fast processing of long sequence datasets. Test results on simulated and real datasets show that LocalMotif offers advantage over existing motif finding algorithms in accurately detecting localized motifs in long sequences.

CHAPTER - V

GENERAL PROMOTER PREDICTION

This chapter develops a novel statistical model for promoters and a technique for detecting promoter regions (TSS) in genomic sequences. A number of existing techniques analyze the occurrence frequencies of oligonucleotides in promoter sequences as compared to other genomic regions. In contrast, the present work studies the *positional densities* of oligonucleotides in promoter sequences. Modeling based on positional densities eliminates the need of any non-promoter sequence dataset or any model of the background oligonucleotide content of the genome. Instead, using only the positive dataset of promoter sequences, the statistical model automatically recognizes a number of TFBS along with their occurrence positions relative to the TSS.

The concept of positional density is introduced in Section V-3. Based on this model, a continuous naïve Bayes classifier is developed in Sections V-4 and V-5 for the detection of promoters and TSS in genomic sequences. The model is trained specifically on the dataset of human promoter sequences, and therefore a brief overview of the composition of human promoters has been presented in Section V-2. Results of promoter prediction on a number of datasets derived from the human genome and performance comparison with existing 2nd generation promoter prediction tools are described in Section V-7.

V-1 Introduction

The advantages of an unsupervised feature extraction and AI modeling approach in the computational modeling and detection of promoter sequences were described in Chapter 1, while the development of 2nd generation promoter prediction tools was described in Chapter 2. Among the 2nd generation tools, an important but relatively less explored approach is using probability models. An early attempt in this direction by Audic and Claverie (1997) used simple Markov chains of order four to six to model promoter sequences. However, the authors reported low performance of the model due to its simplicity and its overfitting of training promoter sequences. Ohler *et al.* (1999) used interpolated Markov chains, which is a generalization that combines several simple Markov chains of different orders. It takes into account statistics of higher orders without overfitting the model to training data. Ohler *et al.* (1999) initially reported performance equivalent to first generation promoter prediction tools. However, improved results have been reported recently upon retraining the model on a larger dataset of Drosophila core promoters [Ohler et al. (2002)].

In a slightly different context of locating regulatory regions in genomic sequences with promoters as a subset, a hidden Markov model was developed by Crowley *et al.* (1997). The model assumed DNA sequences as a hidden Markov process and detected change-points between non-regulatory and regulatory segments based on the appearance of clusters of binding sites in a local region. Although regulatory features such as enhancer, locus control regions and promoters could thus be identified, no attempt was made to accurately predict the promoter region and TSS. No other significant research in this direction has been published to the best knowledge of the author.

The present research extends the application of a purely probability model based approach in eukaryotic promoter prediction. Specifically, it attempts a Bayesian network model of general eukaryotic promoters [Narang et al. (2005)]. The promoter sequence is modeled in probabilistic framework (using a continuous naïve Bayes representation) as a set of TFBS occurring with varying probabilities in different regions of the sequence. This is commensurate with the current biological understanding of promoters as a combination of different regions, *viz.* core promoter, proximal promoter and distal promoter, with the TFBS in different regions having different degrees of mobility. The position of each TFBS is expressed probabilistically in the form of a statistical distribution. The nature of the positional distribution defines the location relative to the TSS as well as the degree of mobility of the binding site. For instance, the positional distribution of TATA box and CAAT box are shown in Figure V-1. The close location relative to the TSS as well as the low mobility of the TATA box is clearly described by the peak of the positional distribution in the –30 to –40 region. Similarly the location of the CAAT box in the proximal promoter region as well as its higher mobility as compared to the TATA box is observed in its positional distribution curve which has a greater spread in the –140 to –80 region. Thus in sharp contrast to previous works, the present research models the positional densities of oligonucleotides instead of their occurrence frequencies. This has several advantages as described later in Section V-3.

V-2 Structure of Human Promoters

The structure and functioning of eukaryotic promoters has been discussed in several reviews, e.g., [Werner (1999), Pederson et al. (1999), Zhang (2002)]. In general, the promoter is understood as a combination of different regions with different functions. The sub portion of the promoter surrounding the TSS is called *core promoter*. It interacts with RNA polymerase II and basal transcription factors, and is the minimal sequence that is required for initiating transcription. Gene-specific regulatory elements present up to

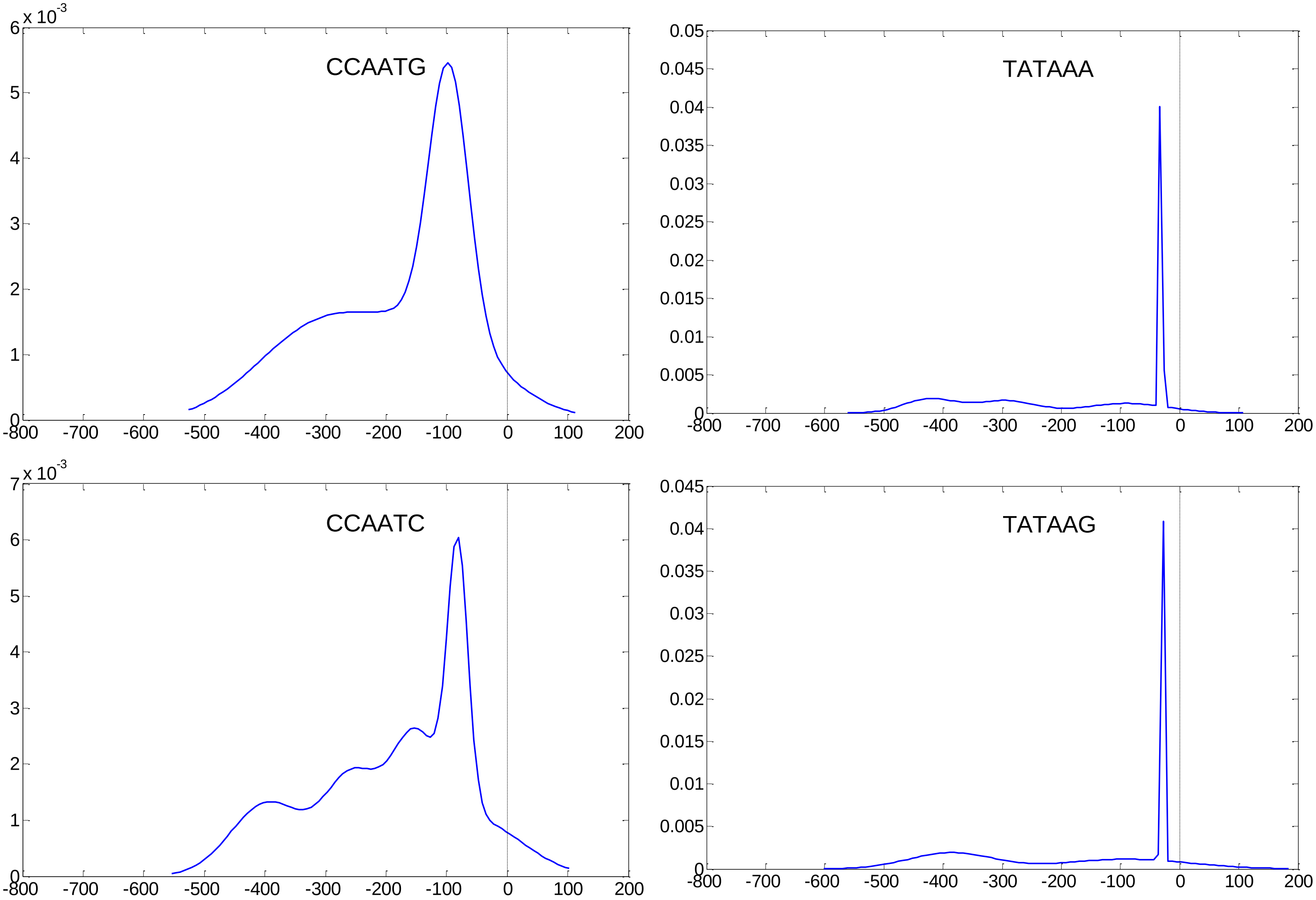


Figure V-1. Positional densities of the TATA box and CAAT box binding sites in a set of 1796 promoter sequences obtained from the eukaryotic promoter database.

few hundred base pairs upstream of the core promoter are commonly referred to as the *proximal promoter region*. These are recognized by TFs called activators and determine the efficiency and specificity of promoter activity. Further, there are enhancers and distal promoter elements which may be located far distant from the TSS, but can considerably affect the rate of transcription. Although well-organized, eukaryotic promoters are very varied in their structure. Therefore only human promoters have been used in this work to simplify the study.

Multiple studies have been reported recently on the composition of core promoters of human genes [Bajic et al. (2004), Smale and Kadonaga (2003)]. Well defined transcription factor binding motifs exist within the core promoter region, which determine the location of the start site and the direction of transcription. It is indicated that roughly 30% or less of human core promoters have a TATA box at -25 to -30 position with consensus TATAAA [Suzuki et al. (2001)]. The TATA box tends to be surrounded by GC rich sequences, including the TFIIB recognition element, BRE, lying as an upstream extension (consensus SSRCGCC). Upto 80% human promoters (both TATA and TATA-less) have an initiator element (Inr) located at the transcription start site [Suzuki et al. (2001)]. It has a consensus sequence YCAYYYYY, with the base ‘A’ lying at the position of TSS. In promoters that are TATA-less but have an Inr, a downstream promoter element (DPE) is usually found at +28 to +32 positions [Smale and Kadonaga (2003)]. About half of the human promoters are associated with CpG islands [Suzuki et al. (2001)], and the functional regulatory elements in these sequences have been difficult to identify [Smale and Kadonaga (2003)]. Some of these promoters contain none of the common core promoter elements discussed above. Exact compositional characterization of known human core promoter sequences is found in [Bajic et al. (2004)].

A larger variation is observed in the composition of the proximal promoter region. Also, the location and orientation with respect to the TSS of transcription factor binding motifs in proximal promoter region is more flexible than that in the core promoter. CAAT box, GC box, E box, GATA box, octamer *etc.* are some of the frequently encountered proximal promoter elements. Some of these elements (such as GC and CAAT boxes) can be present in either orientation.

The context in which a binding site is present within a promoter sequence plays an important role. For example, two interacting transcription factors bound to closely situated sites may lead to non-additively high or low levels of transcriptional activity. Such effects have been compiled in the COMPEL database [Kel-Margoulis et al. (2002)].

There are several other factors involved in transcriptional regulation, such as enhancers/silencers, insulators, chromatin structure, locus control regions and so forth. However these are beyond the scope of this research. The present work utilizes only core promoter and proximal promoter regions for the detection of promoter regions. This invariably limits to some degree the performance of the computational model.

V-3 Oligonucleotide Positional Density

Transcription factor binding motifs in promoter sequences are frequently identified by analyzing the occurrence frequencies of oligonucleotides [Hutchinson (1996), Chen et al. (1997), Scherf et al. (2000), Bajic et al. (2003), van Helden et al. (1998), Bajic et al. (2004)]. Oligonucleotides that are statistically over-represented in promoter sequences as compared to non-promoter sequences usually correspond to the consensus sequences of transcription factor binding motifs. The comparative analysis of oligonucleotide occurrence frequencies in promoter *versus* non-promoter sequence datasets is, however, difficult in practice due to several reasons, such as [Bajic et al. (2004)]:

(i) Oligonucleotide frequency distribution varies significantly across different samples of promoter and non-promoter sequence data. Thus the quality of results is significantly affected by the quality of both promoter and non-promoter sequence data.

(ii) Results also depend to a great degree upon the statistical measure and threshold settings used in the analysis.

(iii) When the training set of promoters is biased, it is difficult to identify important but less represented motifs.

In the present research, the *positional densities* of oligonucleotides are studied. Positional density of an oligonucleotide measures the probability of its occurrence at various positions relative to the TSS within promoter sequences (Figure V-2). The density function only represents the preference of the oligonucleotide to occur at various positions around the TSS, and is independent of its total frequency of occurrence in the promoter sequences. The density is expected to be non-uniform for an oligonucleotide that corresponds to the consensus sequence of a motif. This is because several motifs in core promoter and proximal promoter regions occur within a preferred range of positions relative to the TSS. For example, the TATA box usually lies in the position window -30 to -25 within vertebrate promoter sequences. Consequently, the hexamer TATAAA occurs with much higher probability in this range of positions within promoter sequences. Indeed, its positional density is heavily skewed as observed in Figure V-2.

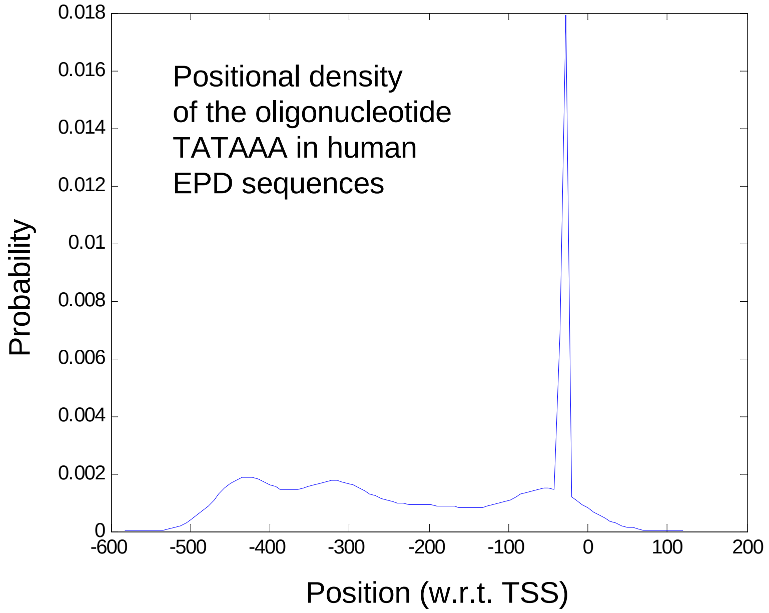


Figure V-2. An illustration of the positional density of the oligonucleotide TATAAA, obtained using 1796 human promoter sequences in EPD. The TSS is located at position 0. The curve indicates the probability of observing the oligonucleotide TATAAA at various positions upstream and downstream of the TSS.

Thus, the information of transcription factor binding motifs and their preferred position in promoters is encoded in the shapes of oligonucleotide positional density curves. The positional density analysis presented in this paper exploits this information. The technique is robust since it does not involve extraneous factors such as tuning of various parameters or determination of background frequencies of oligonucleotides from non-promoter data. Furthermore, it can identify some less frequent but important motifs, since the skew in the positional density is independent of the actual occurrence frequency of the oligonucleotide. The efficiency of this method is demonstrated in Section V-7.1 by its ability to learn most well-known motifs from a set of example promoter sequences.

V-4 Bayesian Network Model for General Promoter Prediction

This section describes the concept and implementation of the novel Bayesian network based statistical technique for general promoter prediction. The method pivots around the positional densities of oligonucleotides of a fixed length, *k*, within the promoter sequences. In general, there are 4^k possible oligonucleotides of length *k* since there are only four DNA nucleotides, A, C, G and T. The oligonucleotides are represented by the symbol, K_i, and indexed in alphabetical order, where the index, i, spans over the range $1, 2, \ldots, 4^k$. In a set of training promoter sequences, the occurrence positions of each oligonucleotide are observed relative to the TSS; taken as negative upstream (*i.e.*, towards 5') of the TSS, positive downstream (*i.e.*, towards 3') of the TSS, and +1 at the TSS.

V-4.1 The Promoter Model

The promoter prediction technique defines two different statistical models – a *promoter model*, π, and a *non-promoter model*, $\bar{\pi}$. The statistical promoter model measures for each oligonucleotide, K_i, its *positional density*, $f_i\left(p \middle| \pi\right)$, in promoter sequences. The positional density gives the preference of K_i to occur at various positions around the TSS in the promoter sequences. It is a probability density function such that

$$\Pr\left(p_1 < P_i < p_2 \middle| \pi\right) = \int_{p_1}^{p_2} f_i\left(p \middle| \pi\right) dp, \tag{5.1}$$

where P_i is the random variable representing the position of occurrence of K_i relative to the TSS. The support $\left[a,b\right]$ of the density functions $f_i\left(p|\pi\right)$ depends upon the length of the training promoter sequences available as shown in Figure V-3a. Thus, in equation (5.1), $p_1, p_2 \in \left[a,b\right]$. The total frequency of occurrence of K_i within the promoter sequences is irrelevant, and hence the total area under the positional density curve is unity for any K_i. As an example, the positional density function for the hexamer TATAAA in human promoters is shown in Figure V-2. The positional density of TATAAA has a sharp peak in the position range -30 to -25, indicating its high preference to occur in these positions in the promoter sequences.

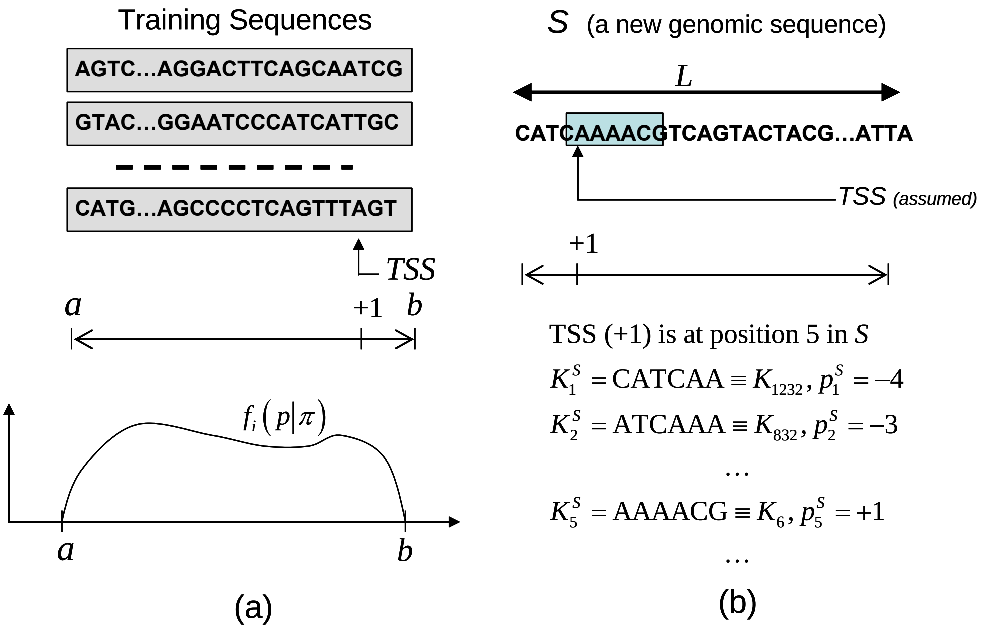

Training Sequences
AGTC...AGGACTTCAGCAATCG
GTAC...GGAATCCCATCATTGC
CATG...AGCCCCTCAGTTTAGT
TSS
a
+1
b
$f_i\left(p|\pi\right)$
a
b
(a)
S (a new genomic sequence)
L
CATCAAAACGTCAGTACTACG...ATTA
TSS (assumed)
+1
TSS (+1) is at position 5 in S
$K_1^S = \text{CATCAA} \equiv K_{1232}, p_1^S = -4$
$K_2^S = \text{ATCAAA} \equiv K_{832}, p_2^S = -3$
…
$K_5^S = \text{AAAACG} \equiv K_6, p_5^S = +1$
…
(b)

Figure V-3. (a) Relationship between positional density definition and training promoter sequences, (b) modeling a nucleotide sequence, *S*, for promoter inference (Equation 5.4).

The non-promoter model, on the other hand, is defined simply as a uniform density function for all K_i, *i.e.*, for all $i \in 1, 2, \ldots, 4^k$,

$$f_i\left(p \middle| \bar{\pi}\right) \equiv u\left(p\right) = \begin{cases} 1/\left|b-a\right| & \text{for } p \in \left[a,b\right] \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

This is due to the assumption that oligonucleotides do not show any particular positional preference around any position anchor in the case of non-promoter sequences. This assumption can be easily verified on any non-promoter dataset.

The statistical promoter prediction technique considers any nucleotide sequence, *S*, of length, *L*, as a combination of $\left(L-k+1\right)$ length-*k* oligonucleotides, $K_1^S, K_2^S, \ldots, K_{L-k+1}^S$, occurring at various positions, $p_1^S, p_2^S, \ldots, p_{L-k+1}^S$ around the assumed TSS position, *T* as shown in Figure 2(b). The superscript *S* is introduced in the notation to avoid confusing with the symbols K_i and P_i used above for defining the positional densities of oligonucleotides in the promoter model. The observed sequence *S* is likened to an experiment of drawing one ball each independently from $\left(L-k+1\right)$ different urns, K_x^S, where the probability of drawing the ball of type p_x^S from the urn K_x^S is given as $\Pr\left(p_x^S \middle| K_x^S\right)$. Thus, the probability of observing the sequence *S* is given as

$$\Pr\left(S\right) = \prod_{x=1}^{L-k+1} \Pr\left(p_x^S \middle| K_x^S\right) \Pr\left(K_x^S\right) \tag{5.3}$$

Now, if the sequence *S* is hypothesized as a promoter sequence, the probabilities $\Pr\left(p_x^S \middle| K_x^S\right)$ are the positional densities $f_i\left(p_x^S \middle| \pi\right)$ in the promoter model, where the oligonucleotide K_x^S found at position *x* in the sequence *S* is actually the oligonucleotide

K_i. Substituting the corresponding $f_i\left(p_x^S \middle| \pi\right)$ in equation (5.1) would then yield the probability $\Pr\left(S \middle| \pi\right)$. However, when *S* is not a promoter sequence, the probability, $\Pr\left(S \middle| \bar{\pi}\right)$, of observing the sequence *S* is found in a similar fashion, but using the non-promoter model, $f_i\left(p_x^S \middle| \bar{\pi}\right)$.

Finally, the probability that the observed sequence *S* is actually a promoter sequence is obtained using the Bayesian formula,

$$\Pr\left(\pi \middle| S\right) = \frac{\Pr\left(S \middle| \pi\right)\Pr\left(\pi\right)}{\Pr\left(S \middle| \pi\right)\Pr\left(\pi\right) + \Pr\left(S \middle| \bar{\pi}\right)\Pr\left(\bar{\pi}\right)} \qquad (5.4)$$

V-4.2 Naïve Bayes Classifier Representation

It is interesting to note that the promoter prediction technique described in section V-4.1 can be neatly expressed in terms of a continuous naïve Bayes model. The *naïve Bayes* model is the simplest case of a Bayesian network and is frequently used for classification [Jensen (2001), Friedman et al. (1997)]. Although independence is generally a poor assumption, in practice naive Bayes often competes well with more sophisticated classifiers [Domingos and Pazzani (1996)]. A generative classifier for the present problem is shown in Figure V-4. The oligonucleotide position random variables, P_i (refer equation (5.1)), and a class variable, *C*, are the nodes in the graph. *C* is a binary variable representing promoter (π) and non-promoter ($\bar{\pi}$) classes. Thus, the variables P_i form continuous nodes, while *C* forms a discrete node. The naïve Bayes graph implies that the position random variables, P_i, are assumed to be independent of each other given the class *C*. The probability model is described by the distributions $\Pr\left(P_i \middle| \pi\right)$

and $\Pr\left(P_i|\bar{\pi}\right)$. These are nothing but the positional densities, $f_i\left(p|\pi\right)$ and $f_i\left(p|\bar{\pi}\right)$ respectively.

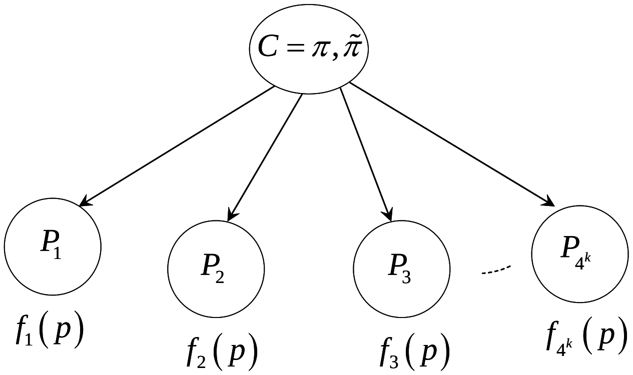


Figure V-4. The naïve Bayes classifier for promoter prediction.

V-4.3 Modeling and Estimation of Positional Densities

Now the mathematical modeling and the estimation of the positional densities $f_i\left(p|\pi\right)$ from a training dataset of promoter sequences is described. Although the position variable, *p*, is discrete, for the purpose of convenience of modeling and estimation, it is treated as a continuous variable over the range $\left[a,b\right]$. The positional density is approximated as a finite mixture of Gaussians,

$$f_i\left(p\,|\,G_i,\theta_i,\pi\right)=\sum_{s_i=1}^{G_i}\alpha_{s_i}\phi\left(p\,|\,\mu_{s_i},\sigma_{s_i}^2\right), \tag{5.5}$$

$$\text{with } \alpha_{s_i}\geq 0 \text{ and } \sum_{s_i}\alpha_{s_i}=1,$$

where G_i is the number of components in the mixture; $\phi\left(p\,|\,\mu_{s_i},\sigma_{s_i}^2\right)$ is a Gaussian distribution with parameters mean, μ_{s_i} , and variance, $\sigma_{s_i}^2$, *i.e.*,

$\phi\left(p \mid \mu_{s_i}, \sigma_{s_i}^2\right) = (2\pi)^{-1/2} \exp\left(-\left(p-\mu_{s_i}\right)^2 \big/ 2\sigma_{s_i}^2\right)$; α_{s_i} are the mixing proportions; and $\theta_i = \left\{\alpha_{s_i}, \mu_{s_i}, \sigma_{s_i} \middle| s_i = 1, 2, \ldots, G_i\right\}$ is the set of all model parameters.

The model is learnt from a training dataset of n promoter sequences, $\{S_1, S_2, \ldots, S_n\}$, aligned with respect to the TSS and spanning over the position range $[a,b]$. For each oligonucleotide, K_i , the set of all observations, $\underset{\sim}{p_i} = \begin{bmatrix} p_i^1 & p_i^2 & \ldots & p_i^{N_i} \end{bmatrix}$, is obtained where p_i^j is the position of the jth occurrence of the oligonucleotide K_i with respect to the TSS in training promoter sequences, and N_i is the total number of occurrences of K_i in all these sequences.

The maximum likelihood estimate of the model parameters, θ_i, is obtained from observations $\underset{\sim}{p_i}$. Given that these are all statistically independent observations from the mixture density, the log-likelihood function is written as

$$L_i\left(\theta_i \mid G_i\right) = \sum_{j=1}^{N_i} \log\left(f_i\left(p_i^j \middle| G_i, \theta_i, \pi\right)\right). \tag{5.6}$$

The maximum likelihood estimate of θ_i , denoted by θ_i^* , is defined as $\theta_i^* = \arg\max_{\theta_i}\left(L_i\left(\theta_i\right)\right)$.

Obtaining the maximum likelihood estimate, θ_i^*, requires taking the derivative of the likelihood function in equation (5.6) and equating it to zero. The resulting equations, however, are nonlinear and there is no closed form solution. Therefore the well known expectation maximization (EM) algorithm is used to obtain the parameter estimates. The EM algorithm assumes initial arbitrary values of the parameters and then iteratively updates them to converge at a local maximum of the log likelihood function. Detailed EM equations used iteratively for updating the estimates of Gaussian mixture model parameters $\theta_i = \left\{ \alpha_{s_i}, \mu_{s_i}, \sigma_{s_i} \middle| s_i = 1, 2, \dots, G_i \right\}$ with respect to the dataset $\underset{\sim}{p_i} = \begin{bmatrix} p_i^1 & p_i^2 & \dots & p_i^{N_i} \end{bmatrix}$ are written as follows [Carlin and Louis (2000)]:

for each observation p_i^j with $j = 1, 2, \dots, N_i$,

$$\Pr\left(s_i \mid p_i^j \right) = \frac{\alpha_{s_i} \phi\left(p_i^j \mid \mu_{s_i}, \sigma_{s_i}^2 \right)}{\sum_{s_i=1}^{G_i} \alpha_{s_i} \phi\left(p_i^j \mid \mu_{s_i}, \sigma_{s_i}^2 \right)}, \tag{5.7}$$

$$\alpha_{s_i}^{new} = \left(1/N_i \right) \sum_{j=1}^{N_i} \Pr\left(s_i \mid p_i^j \right), \tag{5.8}$$

$$\mu_{s_i}^{new} = \left(1/\alpha_{s_i} N_i \right) \sum_{j=1}^{N_i} \Pr\left(s_i \mid p_i^j \right) p_i^j , \tag{5.9}$$

and

$$\sigma_{s_i}^{2\ \ new} = \left(1/\alpha_{s_i} N_i \right) \sum_{j=1}^{N_i} \Pr\left(s_i \mid p_i^j \right) \left(p_i^j - \mu_{s_i} \right)^2 . \tag{5.10}$$

Equations (5.7)-(5.10) are applied iteratively over the complete dataset $\underset{\sim}{p_i}$ for all the mixture components, $s_i = 1, 2, \dots, G_i$, until convergence is obtained. A suitable convergence criterion is that the maximum change in the updated value of any of the parameters between two successive iterations is less than some value δ, where δ can be set at 10^{-4}.

Since the EM algorithm converges to some local maxima (or sometimes saddle points) of the likelihood function, usually the results are highly dependent upon the initial parameter values chosen. Thus it requires several re-runs with different random initializations of the parameters to arrive at a satisfactory solution. To overcome such problems, the current implementation is based on the greedy learning algorithm described in [Verbeek et al. (2003)]. In this implementation, instead of starting with a random initialization of all components and improving upon these components with EM, the mixture is built component-wise. In the beginning, there is only one component, i.e., $G_i = 1$. For this mixture, the parameters are computed trivially as the sample mean and variance. Then a new component is inserted, i.e., $G_i^{new} := G_i + 1$ and

$$f_i\left(p \mid G_i^{new}, \theta_i, \pi\right) = (1-\beta)\, f_i\left(p \mid G_i, \theta_i, \pi\right) + \beta\phi\left(p \mid \mu_{G_i+1}, \sigma_{G_i+1}^2\right) \qquad (5.11)$$

As discussed in [Verbeek et al. (2003)], the newly inserted component with parameters $\left(\mu_{G_i+1}^{*}, \sigma_{G_i+1}^{2\,*}, \beta^{*}\right)$ is optimal in the sense that its insertion maximizes the likelihood function over the set of all possible insertions. The complete set of parameters for this new mixture are then updated using EM (equations (5.7)-(5.10)) until convergence.

The optimum number of components, $G_{i,opt}$, in the mixture density is obtained using Akaike Information Criterion (AIC). AIC is expressed as

$$AIC\left(G_i\right) = -2L\left(\theta_i^{*}\right) + 2n\left(\theta_i\right), \qquad (5.12)$$

where $n\left(\theta_i\right)$ is the number of free parameters in the set θ_i, which in this case is $3G_i - 1$. The optimum number of components, $G_{i,opt}$, is the value of G_i that minimizes AIC.

V-5 Inference Over Long Genomic Sequences

In sections V-4.1 and V-4.2, the method of classifying a given nucleotide sequence S of length, $L = \left|b - a\right|$, as promoter or non-promoter was discussed. Now the technique is extended to detect the transcription start sites in a given long genomic sequence.

If the classifier has been trained using example promoter sequences of length $L = |b - a|$ and with TSS location defined as the origin, the same configuration is used during inference. A window of size *L* is selected from the given genomic sequence. The naïve Bayes classifier infers the probability that this sequence window belongs to the promoter class. The window is moved across the whole sequence as shown in Figure V-5, and all regions with high probability of being a promoter are identified. This sliding window approach has been used earlier in [Scherf et al. (2000)]. The predicted TSS location is obtained from the window that has maximum probability of being a promoter in a local region.

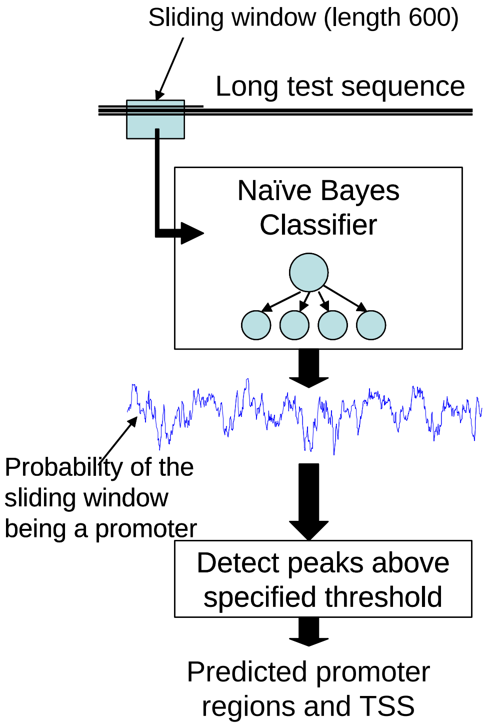

Sliding window (length 600)
Long test sequence
Naïve Bayes Classifier
Probability of the sliding window being a promoter
Detect peaks above specified threshold
Predicted promoter regions and TSS

Figure V-5. Using naïve Bayes classifier to detect promoter region and TSS in long genomic sequences.

V-6 Implementation

The continuous naïve Bayes classifier has been implemented as a software called BayesProm in Microsoft Visual C++®, and the binary executable is available freely at the website http://www.comp.nus.edu.sg/~bioinfo/BayesProm. It was trained using a set of 1796 human promoter sequences obtained from the Eukaryotic Promoter Database Version 74 [Schmid et al. (2004)]. These sequences were of length 600; -499 to +100 relative to the TSS. Thus, the window size for the classifier was fixed as $a = -499,\ b = +100$.

Classifier parameters that require tuning included (i) the length of oligonucleotides, *k*, and (ii) the probability threshold, ϕ, above which a sequence region can be classified as promoter. Testing was performed by varying oligonucleotide lengths from $4 \le k \le 10$. The value $k = 6$ yielded much superior results as compared to any other length. Hence, BayesProm uses only hexamers. The threshold value, ϕ, is left free for being set by the users (within reasonable limits) depending upon their requirements of sensitivity *vs.* specificity of the predictions [Bajic et al. (2003)].

For training the BayesProm model, 80% of the 1796 EPD sequences were used as training set, while the rest were used as validation set. Partitioning of the sequences into training and validation sets was performed randomly. Five such uncorrelated cross-validation sets were generated. The training and cross-validation results are reported in Table V-1. Accuracy was tested on both training and validation sets, and simultaneously on a negative set of sequences consisting of human exon and 3' UTR sequences derived from Genbank. Note that the negative sequence set was in no way used for training.

Sensitivity on the positive set was consistently between 75% to 85%, while false positive rate on the negative set was less than 1%.

Table V-1. Results of cross-validation studies in the training of BayesProm. The complete dataset of 1796 human promoter sequences was randomly divided into 1436 training sequences (80%) and 360 validation sequences (20%). Five such uncorrelated cross-validation sets were generated. A negative set of 5000 human exon and 3' UTR sequences obtained from Genbank was used simultaneously for testing.

Set no.	# TP in training (out of 1436)	# TP in validation (out of 360)	#FP over the negative set (out of 5000)
1	1221 (85%)	306 (85%)	32 (0.7%)
2	1188 (82%)	279 (78%)	55 (1.1%)
3	1197 (83%)	267 (74%)	46 (0.9%
4	1067 (74%)	285 (79%)	42 (0.9%)
5	1203 (84%)	294 (82%)	37 (0.8%)

V-7 Results

The performance of the novel statistical approach has been evaluated in two aspects – (i) the performance of TSS predictions, and (ii) the ability of the model to accurately learn various transcription factor binding motifs and their locations around the TSS from training promoter sequences.

V-7.1 Prominent Features Correspond to Well-Known Transcription Factor Binding Motifs

An advantage of statistical models as in the present work is that the physical features learnt by the model can be directly evaluated. The significant features learnt by the promoter model in the form of oligonucleotide positional densities are reported here.

As described above, the training data from which these features were learnt is the set of human promoters obtained from EPD over the range -499 to +100 relative to the TSS.

It is implied by equation (5.1) that the probability of an oligonucleotide, K_i, occuring within a position window $\left(p_1, p_2\right)$ relative to the TSS is given by the area under its positional density curve within this position window, *i.e.*, $\int_{p_1}^{p_2} f_i\left(p\right)dp$. Using this formula, occurrence probabilities of all oligonucleotides within several narrow position windows were computed. Subsequently, oligonucleotides having a high occurrence probability within the same position window were grouped together. In each such group, similar oligonucleotides were clustered and used to construct a consensus sequence. It was found that several such consensus sequences correspond to those of well-known transcription factor binding sites, such as the TATA box, initiator and so on. Figure V-6 illustrates some of the results of this analysis.

The results indicate that the features learnt by the statistical model corresponds to actual biological information contained within promoter sequences. It is plausible that the modeling technique presented in this work may be useful in computationally deriving new biological conclusions out of a dataset of promoter sequences. Work is in progress in this direction.

V-7.2 Results of TSS Prediction

The transcription start site prediction accuracy of continuous naïve Bayes model BayesProm is tested on three real human promoter datasets – a relatively short sequence dataset of Genbank sequences, a long genomic contig and human chromosome 22. A thorough performance analysis is reported in this section in terms of complete ROC

Consensus	Preferred position	Corresponding oligonucleotides	Window Position	Probability
TATAAA (TATA box)	-35 to -25	TATAAA	-40 to -20	0.564
		TATAAC	-40 to -20	0.25
		TATAAG	-40 to -20	0.473
		TATATA	-40 to -20	0.365
		TAAAAG	-40 to -20	0.364
		TAAAGG	-40 to -20	0.299
		TAAATA	-40 to -20	0.275
		TGTATA	-40 to -20	0.307
		ATAAAA	-40 to -20	0.299
		ATAAAG	-40 to -20	0.348
		ATAAAT	-40 to -20	0.285
		ATATAA	-40 to -20	0.394
		CCTATA	-40 to -20	0.437
		CTATAA	-40 to -20	0.597
		CTATAT	-40 to -20	0.413
		GCTATA	-40 to -20	0.543
		GTATAA	-40 to -20	0.568
		GTATAT	-40 to -20	0.331
CCAAT (CCAAT box)	-165 to -40 (-90 mean position)	ACCAAT	-140 to -80	0.259
		CAATGG	-140 to -80	0.201
		CCAATC	-140 to -80	0.201
		CCAATG	-140 to -80	0.279
		GACCAA	-140 to -80	0.209
		GCCAAT	-140 to -80	0.232
GGGCGG (GC box)	-164 to +1	GGCGGG	-140 to -80	0.203
		GGGCGG	-140 to -80	0.208
		GGGGCG	-140 to -80	0.218
		CGGCGG	-80 to -20	0.201
		CGGGGC	-80 to -20	0.256
		GCGCCG	-80 to -20	0.203
		GCGGCG	-80 to -20	0.201
		GCGGGC	-80 to -20	0.211
		GCGGGG	-80 to -20	0.253
		GGCGGG	-80 to -20	0.275
		GGGGCG	-80 to -20	0.266
		CGGCGG	-20 to +40	0.249
		GCGGCG	-20 to +40	0.251
		GGCGGC	-20 to +40	0.254
YCAYYYY (Y=C/T) (Initiator)	A at +1	TCAGTC	-20 to +20	0.221
		CAGTCG	-20 to +20	0.269
		AGTCGT	-20 to +20	0.254
		GTCGTT	-20 to +20	0.273
		TCATAC	-20 to +20	0.211
		TCATTC	-20 to +20	0.248
		CAGTTC	-20 to +20	0.211
		CATTCT	-20 to +20	0.222
		CTCATT	-20 to +20	0.227
		ATCATC	-20 to +20	0.203
		CATACT	-20 to +20	0.200
		GATCAC	-20 to +20	0.209
		TAGCCG	-20 to +20	0.214
		… and others		

Figure V-6. Important consensus sequences recognized by the naïve Bayes model.

characteristics, showing the sensitivity *versus* positive predictive value (ppv) of predictions. As described in Section III-4, sensitivity is defined as the ratio, $Se = TP/(TP + FN)$, where true positives (TP) is the number of TSS that could be correctly predicted, while false negative (FN) is the number of TSS that could not be predicted. Thus it is the percentage of actual TSS that could be successfully predicted. On the other hand, ppv is defined as, $ppv = TP/(TP + FP)$, where false positives (FP) is the number of incorrect predictions reported by the software. Thus, ppv is a measure of the credibility of predictions. Although both high ppv and high sensitivity are desirable, in practice as ppv is increased, the sensitivity of the software goes down.

Another measure of performance concerns the distance of predicted TSS locations from the annotated TSS. Fickett and Hatzigeorgiou (1997) assumed a TSS prediction as correct if it lies 200 bp upstream or 100 bp downstream of the annotated TSS. However, for TSS prediction on the genomic scale such as the full chromosome 22 sequence, a less strict criterion of 2000 bp upstream and 500 bp downstream was chosen by Scherf. *et al.* (2001). In this paper, the former criterion is used for short sequence Genbank dataset, while the latter is used in the case of chromosome 22 sequence. In addition, a comprehensive picture of the prediction accuracy is reported in the form of a histogram showing the number of accurate predictions that lie within a given distance from the annotated TSS.

V.7.2.1 Results on Genbank Dataset

The dataset was prepared from all Genbank (Release 142.0) [Benson et al. (2002)] flat files having a “promoter” feature key annotation. Among these, sequences of length less than 1000 were discarded as these were too short for evaluation. The remaining sequences were compared with the EPD promoter sequences using BLAST [Altschul et al. (1990)]. Sequences that had similarity with any of the EPD sequences with an expect (E) value of less than 1.0E–10 (*i.e.*, greater than 80% similarity) were discarded. Finally a set of 646 human genomic sequences containing a total of 1100 annotated TSS was obtained. TSS prediction accuracy of the present software, BayesProm, on this dataset is compared with a well known 2nd generation promoter prediction tool, Eponine [Down and Hubbard (2002)]. Other programs could not be compared due to unavailability of a batch processing interface.

The predicted TSS locations were compared with annotated TSS, and a ROC curve showing the sensitivity *versus* positive prediction rate is shown in Figure V-7. In one analysis (Case B), a prediction is considered correct if the predicted TSS lies within ±1000 base pairs of the annotated TSS. In a stricter evaluation (Case A), the allowed deviation is limited to ±200 base pairs. In both analyses, BayesProm reports high sensitivity, while Eponine reports high specificity.

A graphical evaluation of the prediction accuracy of BayesProm is illustrated in , where a histogram of the prediction error is plotted for all true predictions. For most of the predictions, the prediction error is almost zero, as shown by the high peaks around distance zero. The number of predictions with high error is relatively less, as is indicated by the trailing of the histogram with increasing distance. Comparison with Eponine in Figure V-8 reveals that Eponine is highly specific, with very few predictions at distances larger than 200. However, BayesProm has good sensitivity as is indicated by the large number of total predictions within any given distance range.

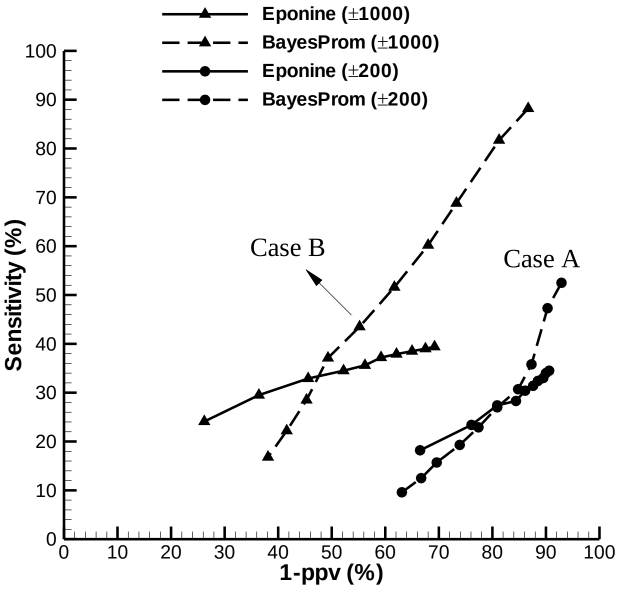


Figure V-7. ROC curve showing the TSS prediction performance of BayesProm and Eponine on Genbank dataset. In case A, TSS predictions within ±200 bp of the annotated TSS were considered correct, while in case B, this range was extended to ±1000 bp. Eponine is seen to be highly specific, while BayesProm has high sensitivity.

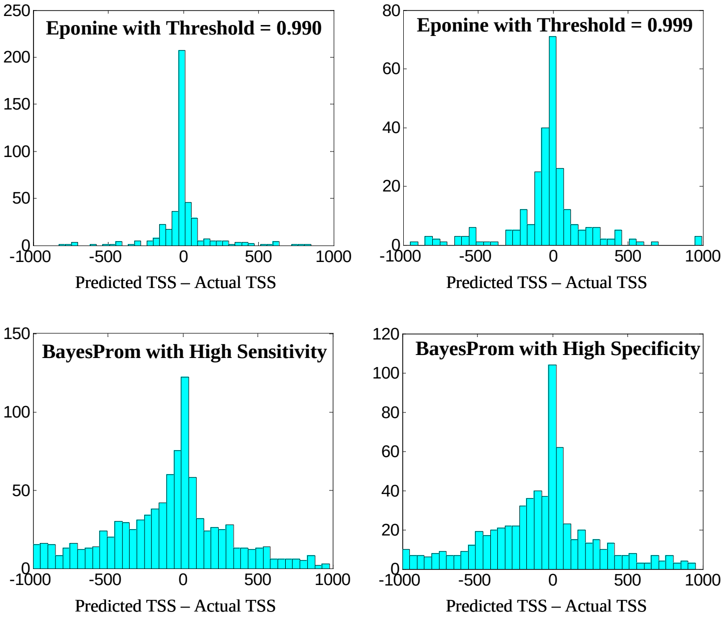


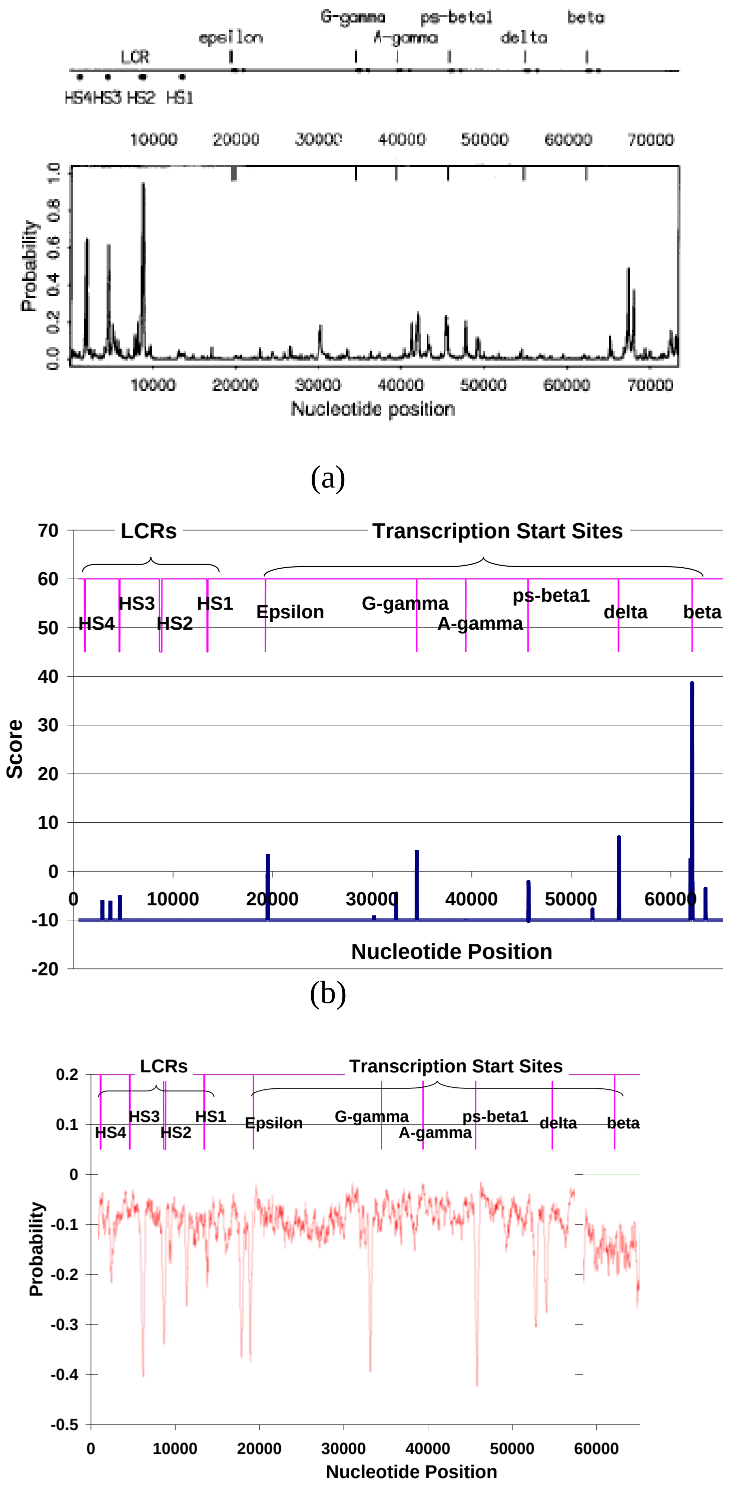
Figure V-8. Density of true predictions relative to the annotated TSS on Genbank dataset. Both Eponine and BayesProm report a histogram peak at zero distance, indicating the accuracy of these softwares. Eponine is seen to be highly specific but less sensitive, while BayesProm is moderately specific but highly sensitive.

V.7.2.2 Comparison with other statistical promoter prediction tools

The accuracy of BayesProm has been compared with two other statistical promoter prediction tools – (i) a hidden Markov model by Crowley et al. (1997), and (ii) interpolated Markov chain model (McPromoter) by Ohler et al. (1999). Since the software developed by Crowley et al. is not readily available, results of BayesProm and McPromoter were compared with their published results. The experiment involved prediction of the regulatory regions in the human β globin locus on chromosome 11 (GenBank accession no. U01317). The sequence contains four locus control regions, *viz.* HS1, HS2, HS3 and HS4; and six transcription start sites, *viz.* beta, delta, epsilon, ps-beta1, A-gamma and G-gamma. The probability of the presence of a regulatory region at various positions within the sequence as predicted by each of the assessed tools is shown in Figure V-9. The peaks of the probability curves indicate predicted regulatory regions. As observed in Figure V-9(a), the HMM model of Crowley *et al.* predicted accurately three out of four locus control regions. BayesProm (Figure V-9(b)), on the other hand, could predict five out of six transcription start sites accurately with very low false positive rate, thus affirming its TSS prediction capability.

V.7.2.3 Results on Human Chromosome 22

Chromosome 22 is a relatively short and better annotated portion of the complete human genome. The annotation of this 33.6Mb sequence, provided by Collins et al. (2003), gives TSS locations of 393 protein coding genes based on experimentally determined full length cDNA transcripts. The availability of a large number of experimentally annotated TSS makes chromosome 22 a good benchmark test dataset.



(a)

(b)

(c)

Figure V-9. Predictions of regulatory regions in the human β globin locus on chromosome 11 (Genbank accession no. U01317) using (a) Hidden Markov Model by Crowley et al. (1997), (b) BayesProm, showing only predictions above threshold of –10, and (c) Interpolated Markov Chain model by Ohler et al. (1999). It is observed that the HMM in (a) can only predict the locus control regions, while BayesProm accurately predicts five of the six transcription start sites with very few false positives.

Therefore several promoter prediction tools including PromoterInspector, Eponine, Dragon Promoter Finder, First Exon Finder, Dragon Gene Start Finder, *etc.* have been tested on the chromosome 22 dataset.

Figure V-10 shows the ROC curve of chromosome 22 prediction results obtained from BayesProm over several different sensitivity settings. The performance of some of the best 2nd generation promoter prediction tools available today is also shown. The evaluation of Eponine was carried out first hand, whereas the results of other software were referenced from published literature, including [Scherf et al. (2001)] for PromoterInspector, [Bajic et al. (2003a)] for Dragon Promoter Finder, and [Bajic et al. (2003b)] for First Exon Finder and Dragon Gene Start Finder.

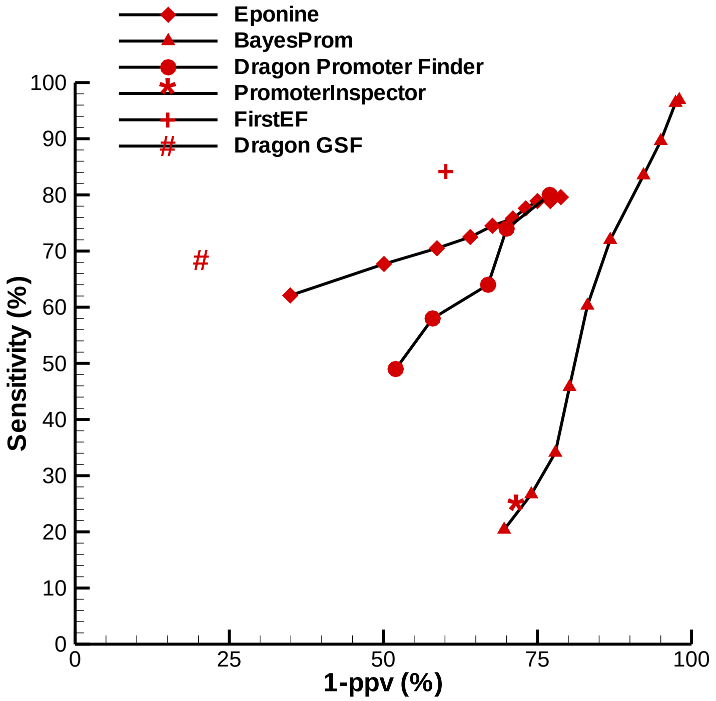


Figure V-10. ROC curve showing the evaluation of BayesProm and several 2nd generation promoter prediction tools on chromosome 22 dataset. The test criterion was same as that used by Scherf et al. (2001).

The ROC curve for BayesProm shows its ability to achieve a ppv of up to 25% for a moderate sensitivity of 30%. This is superior to any of the 1st generation promoter prediction tools, which usually have a ppv less than 10% over all sensitivity ranges [Fickett and Hatzigeorgiou (1997)]. Also note that most of the 2nd generation tools shown in Figure V-10 are fine-tuned for superior performance by carefully selection of model parameters and training dataset. Dragon Gene Start Finder and First Exon Finder are further optimized using additional biological knowledge. Thus the performance of BayesProm as a purely statistical model trained on raw dataset of all EPD human sequences is encouraging. Especially at low ppv values, BayesProm exhibits greater sensitivity than even the best 2nd generation tools.

V-8 Conclusions

The present work extends the scope of statistical models in computational promoter prediction. In contrast to other computational tools that use PWM or oligonucleotide occurrence frequencies, the present work utilized oligonucleotide positional distributions. The technique is free from the practical difficulties that are usually encountered in the analysis of oligonucleotide occurrence frequencies. The purely statistical model has a sound biological basis, and upon training with a dataset of known human promoter sequences, it could automatically learn the transcription factor binding motifs and their occurrence positions relative to the TSS. It could predict human TSS with accuracy competent with some of the 2nd generation promoter prediction tools.

The present work introduced a new modeling framework. However, there are several possible directions in which the present promoter prediction tool can be improved and fine tuned for superior performance. These include careful selection of the training sequence data, feature selection to remove unprofitable oligonucleotides from the model, separate modeling of CpG island and non-CpG island related promoters and incorporating biological knowledge as in [Hannenhalli and Levy (2001)]. Introducing dependencies among the nodes in the Bayesian network model could also improve the model. In addition, the ideas presented in the paper can easily be extended to various other problems in bioinformatics that require analysis of DNA sequence content, especially motif finding.

The present study extends the scope of statistical models in general promoter modeling and prediction. Promoter sequence features learnt by the model correlate well with known biological facts. Results of human transcription start site prediction compare favorably with existing 2nd generation promoter prediction tools.

CHAPTER - VI

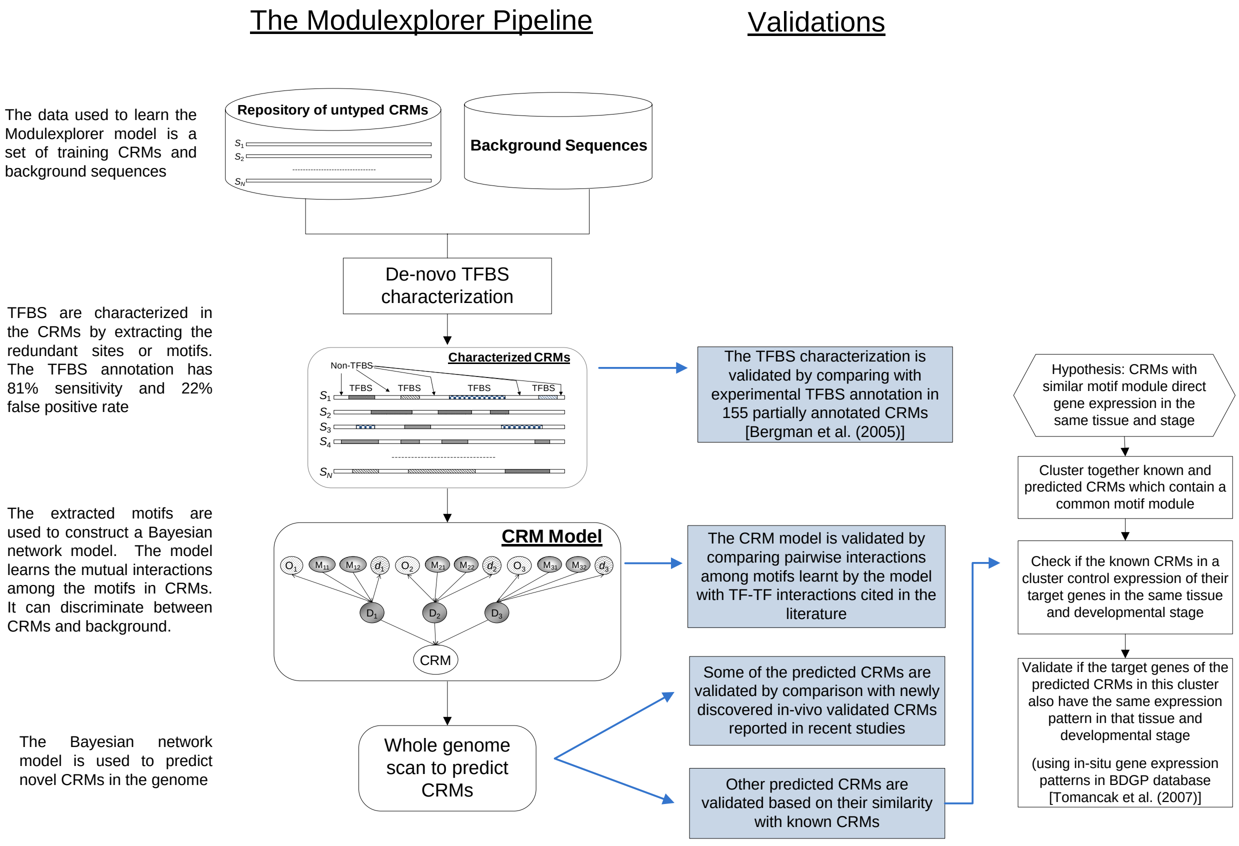
CIS-REGULATORY MODULE PREDICTION

This chapter describes computational modeling of cis-regulatory modules (CRMs) in the genome of Drosophila melanogaster. A CRM is a short DNA sequence that activates or represses the expression of a gene in a particular tissue at a particular development stage. A CRM is usually described to contain a cluster (or module) of motifs for the binding of co-acting transcription factors. CRMs with similar motif module are hypothesized to control the same gene expression pattern. A motif module which governs a specific gene expression pattern is called a regulatory code. So far few regulatory codes are known which have been determined based on wet lab experiments. The research described in this chapter presents the first computational approach to learn regulatory codes de-novo from a repository of CRMs.

A probabilistic graphical model called *Modulexplorer* [Narang et al. (2008)] is developed in this chapter to derive the regulatory codes and to predict novel CRMs. An overview of the Modulexplorer model is given in Section VI-1. The data and methods used to train the Modulexplorer model are described in Sections VI-2 and VI-3 respectively. Training and test performance of the model is evaluated in Section VI-4. Validation of the model is described in Section VI-5. Using the model, 813 novel CRMs were recovered from the Drosophila melanogaster genome regulating gene expression in different tissues at various stages of development. These novel CRMs are described in Section VI-6. Then the recovery of specific regulatory codes for CRMs controlling gene expression in the drosophila embryonic mesoderm, the ventral nerve cord, the eye-antennal disc and the larval wing imaginal disc is described in Section VI-7. The target genes of CRMs following a specific regulatory code have been validated to express in the corresponding tissue at the corresponding development stage. Also 31 genes have been newly implicated in the development of these tissues. The implications of the study are discussed in Section VI-8.

VI-1 Modulexplorer CRM Model

The Modulexplorer pipeline is shown in Figure VI-1. The input to Modulexplorer is a database of known CRMs and a set of non-CRM background sequences. Modulexplorer first characterizes the TFBSs within the CRMs de-novo. It represents



The data used to learn the Modulexplorer model is a set of training CRMs and background sequences

TFBS are characterized in the CRMs by extracting the redundant sites or motifs. The TFBS annotation has 81% sensitivity and 22% false positive rate

The extracted motifs are used to construct a Bayesian network model. The model learns the mutual interactions among the motifs in CRMs. It can discriminate between CRMs and background.

The Bayesian network model is used to predict novel CRMs in the genome

The TFBS characterization is validated by comparing with experimental TFBS annotation in 155 partially annotated CRMs [Bergman et al. (2005)]

The CRM model is validated by comparing pairwise interactions among motifs learnt by the model with TF-TF interactions cited in the literature

Some of the predicted CRMs are validated by comparison with newly discovered in-vivo validated CRMs reported in recent studies

Other predicted CRMs are validated based on their similarity with known CRMs

Hypothesis: CRMs with similar motif module direct gene expression in the same tissue and stage

Cluster together known and predicted CRMs which contain a common motif module

Check if the known CRMs in a cluster control expression of their target genes in the same tissue and developmental stage

Validate if the target genes of the predicted CRMs in this cluster also have the same expression pattern in that tissue and developmental stage

(using in-situ gene expression patterns in BDGP database [Tomancak et al. (2007)]

Figure VI-1. The Modulexplorer pipeline to learn a CRM model from a repository of uncharacterized CRMs and background sequences, and to use the model for predicting novel CRMs is shown in (a). Also shown are the validations that have been conducted in this study to verify the model and the novel CRMs predicted by the model.

these TFBSs with dyad motifs [van Helden et al. (2000); Eskin and Pevzner (2002); Rombauts et al. (2003); Favorov et al. (2005)] in degenerate IUPAC alphabet to achieve high specificity. Then using a probabilistic Bayesian network model, it learns the TFBS interactions which are over-represented in CRMs while under-represented in non-CRMs. The TFBS interactions describe the regulatory codes. The trained model is then used to discover novel CRMs.

The Modulexplorer Bayesian network model for a CRM is shown in Figure VI-2. In the Bayesian network model, the TFBSs are the causal elements or parent nodes while

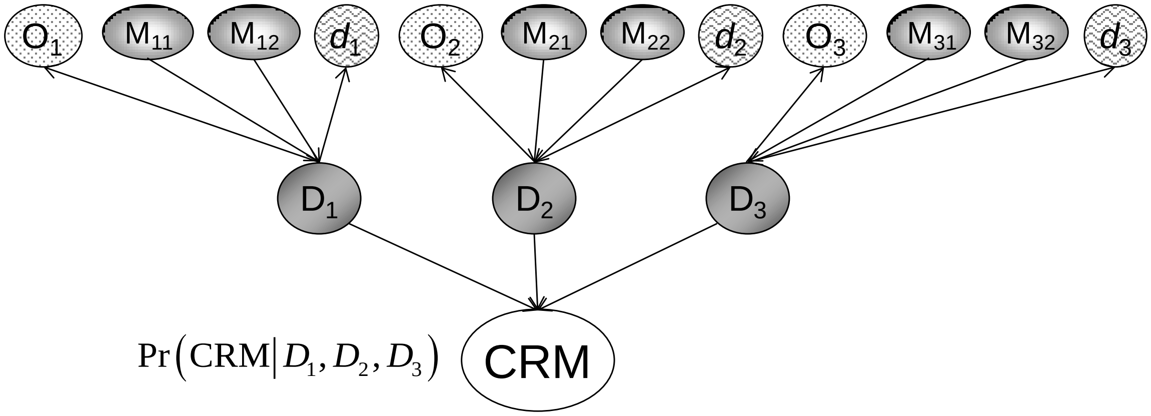


Figure VI-2. The Modulexplorer Bayesian network model. The model describes a CRM as a cluster of multiple interacting TFBS with distance and order constraints. The nodes D_i are the dyad motifs representing the TFBSs. They have states 0 or 1 according to whether the motif is absent or present in the CRM. The CRM is their common effect or hypothesis, represented as the child node. Each dyad motif D_i has two monad components $\left(M_{i1}, M_{i2}\right)$ with a spacer of 0 to 15 bp. These monads are represented by individual nodes M_{i1}, M_{i2} having states 0 or 1, *i.e.* present or absent, and are related to the dyad node D_i by a noisy-AND relationship. The spacer length (or distance), discretized as low or high, is modeled by the node d_i. Furthermore each D_i is associated with an *order* either left or right according to whether M_{i1} appears to the left or to the right of M_{i2} in the CRM.

the CRM is their common effect or child node. The basic idea is to consider a CRM as a cluster of TFBSs for cooperating TFs with certain distance and order constraints. The TFBS interactions are encoded in the probabilities at the edges of the Bayesian network. The interaction probabilities are learnt de-novo by the Bayesian network from training CRM and non-CRM sequences with unsupervised learning. Distance and order constraints are considered between pairs of closely interacting motifs. After training, the Bayesian network model functions as a classifier to discriminate between CRM and non-CRM sequences. During inference, the Bayesian network assigns high probability of CRM to a sequence which contains a combination of closely interacting TFBSs.

VI-2 Data

Experimentally validated CRMs for this study were derived from version 1 of the REDfly database (September 2006 release). The database contained a total of 619 CRMs, among which some were redundant or overlapping. Pairwise sequence similarity was computed using ClustalW [Pavesi et al. (2001)] and CRM sequences with more than 40% similarity were treated as redundant. After removing the redundant sequences, 414 non-overlapping sequences were obtained. Out of these, 58 sequences were too long (>3.5 kb) to be useful in the Modulexplorer pipeline. These were taken as the test dataset. The remaining 356 CRM sequences were selected as the training dataset for Modulexplorer.

The training CRMs represent a diverse mix regulating gene expression in a variety of tissues and stages. Out of 356 training CRMs, 302 control gene expression in the embryo, 193 in the larva and 41 in the adult fly respectively. Of 302 CRMs active in the embryo, 87 are expressed in the blastoderm and 215 in the post-blastoderm stages. The 215 post-blastoderm CRMs are expressed in one or more of the tissues integumentary system, imaginal precursor, nervous system, digestive system, muscle system, circulatory system, tracheal system, reproductive system, excretory system, edipose system and endocrine system as shown in Figure VI-3.

Also for the training of Modulexplorer three different background sequence sets were created from coding, intron and intergenic sequences selected randomly from the whole Drosophila genome. Each of the three background sets consisted of 356 sequences size-matched with the 356 training CRMs.

Experimental annotation of 1066 TFBSs for 83 known TFs in the vicinity of 85 genes was obtained from the Drosophila DNase I Footprint Database v2.0 (FlyReg database) [Bergman et al. (2005)]. This is a subset of the FlyReg database, leaving out entries with unknown transcription factor or gene information. The FlyReg and REDfly databases had 52 genes in common, so that the experimentally annotated TFBSs and CRMs could be related for these 52 genes. Interestingly, the annotated TFBSs overlapped the annotated CRM regions for all genes except one. There were thus 778 known TFBS falling within 155 known CRMs across 51 genes [Narang et al. (2006)]. Based on the survey of literature from which FlyReg annotations were compiled, 19 out of these 155 CRMs are fully annotated with TFBSs while 136 CRMs are partially annotated with TFBSs. This TFBS annotation has been used in this study to validate the de novo TFBS annotations generated by Modulexplorer in the first step of the pipeline.

The BDGP release 5 genome assembly (2007) was used for the whole genome prediction and for all other analyses.

All the datasets and the whole genome prediction results are available for free access at our website http://www.comp.nus.edu.sg/~bioinfo/Drosophila.

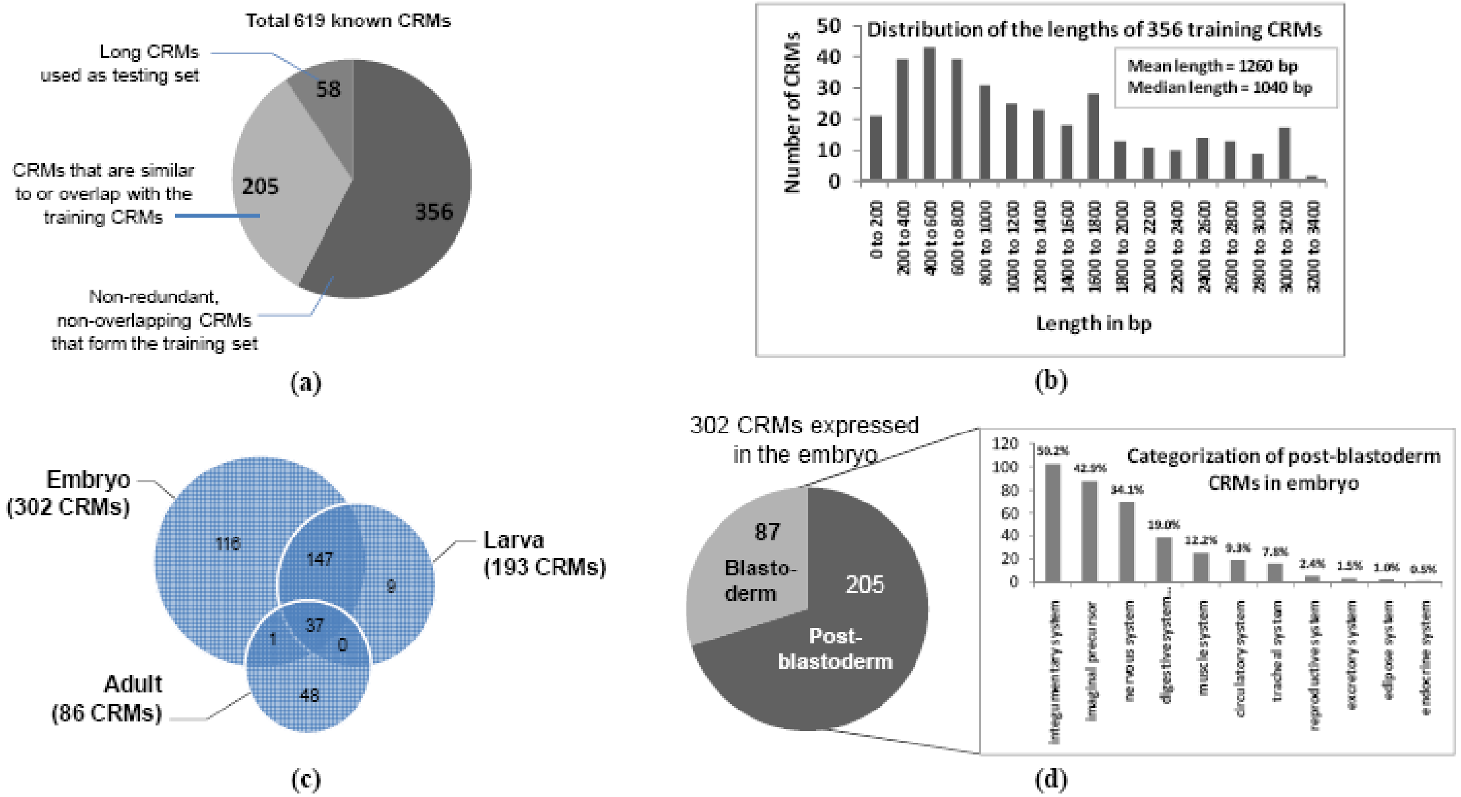


Figure VI-3. (a) From a total of 619 experimental CRM sequences obtained from the REDfly database, 205 redundant CRMs were discarded, 58 long CRMs (>3.5 kbp) were used as a testing set and remaining 356 form the training set. The length distribution of the 356 training CRMs is shown in (b). Most CRMs are between 200 to 1200 bp long with 1040 bp as the median length. The functional diversity among the training CRMs is shown in (c) and (d). Out of the 356 training CRMs, 302 are expressed in the embryo stage, 193 in the larva stage, and 86 in the adult fly. Among the 302 CRMs expressed in the embryo, 87 are expressed in the blastoderm stage (stages 3-5) and 205 in the post-blastoderm stages (stages 6 to 16). Categorization of the 205 post-blastoderm CRMs in terms of the developing organ system where they express is shown in (d). The integumentary system (ectoderm), imaginal precursor (wing disc, retinal disc etc.), nervous system, digestive system (abdomen) and muscle system are over-represented classes among the known CRMs

VI-3 Methods

VI-3.1 TFBS Characterization and Motif Extraction

To construct the Modulexplorer model, we first developed a method to robustly identify TFBSs within CRMs. Drosophila CRMs show homotypic clustering of TFBSs, i.e. they contain multiple binding sites for the same transcription factor [Davidson et al. (2002); Markstein et al. (2002); Lifanov et al. (2003); Ochoa-Espinosa et al. (2005)]. Therefore TFBSs often appear as redundant subsequences within the CRM. We studied the correlation between the redundant sites and TFBSs in 155 Drosophila CRMs for which TFBSs have been previously characterized by DNAse footprinting experiments as described in the Data section. Each of these CRMs contains 3-6 binding sites per TF in general (Figure VI-4(a)). The fluffy tail test statistic [Abnizova et al. (2005)] of these CRMs also indicated a high level of sequence redundancy in the CRMs (Figure VI-4(b)). Based on the observation of sequence redundancy in CRMs, we developed a novel algorithm to recover the TFBSs as redundant sites in a CRM with high accuracy.

The experimentally annotated TFBSs in Drosophila CRMs are 6 to 140 bp long and contain multiple short conserved segments of length 6-10 bp with variable gaps or spacers. We found the dyad motif representation [van Helden et al. (2000), Eskin and Pevzner (2002), Rombauts et al. (2003), Favorov et al. (2005)], similar to the one proposed by Sinha and Tompa (2000) to model spaced motifs in yeast, suitable to represent these TFBSs. A dyad motif is a pair of monad motifs separated by at most D bp. Each monad is written in degenerate IUPAC alphabet {A,C,G,T,R,Y,S,W,N} [Sinha and Tompa (2000)]. A dyad motif is associated with an order. If A and B are two monads in a dyad motif (A,B), the dyad is said to appear in the left (L) (or right (R)) order if A appears

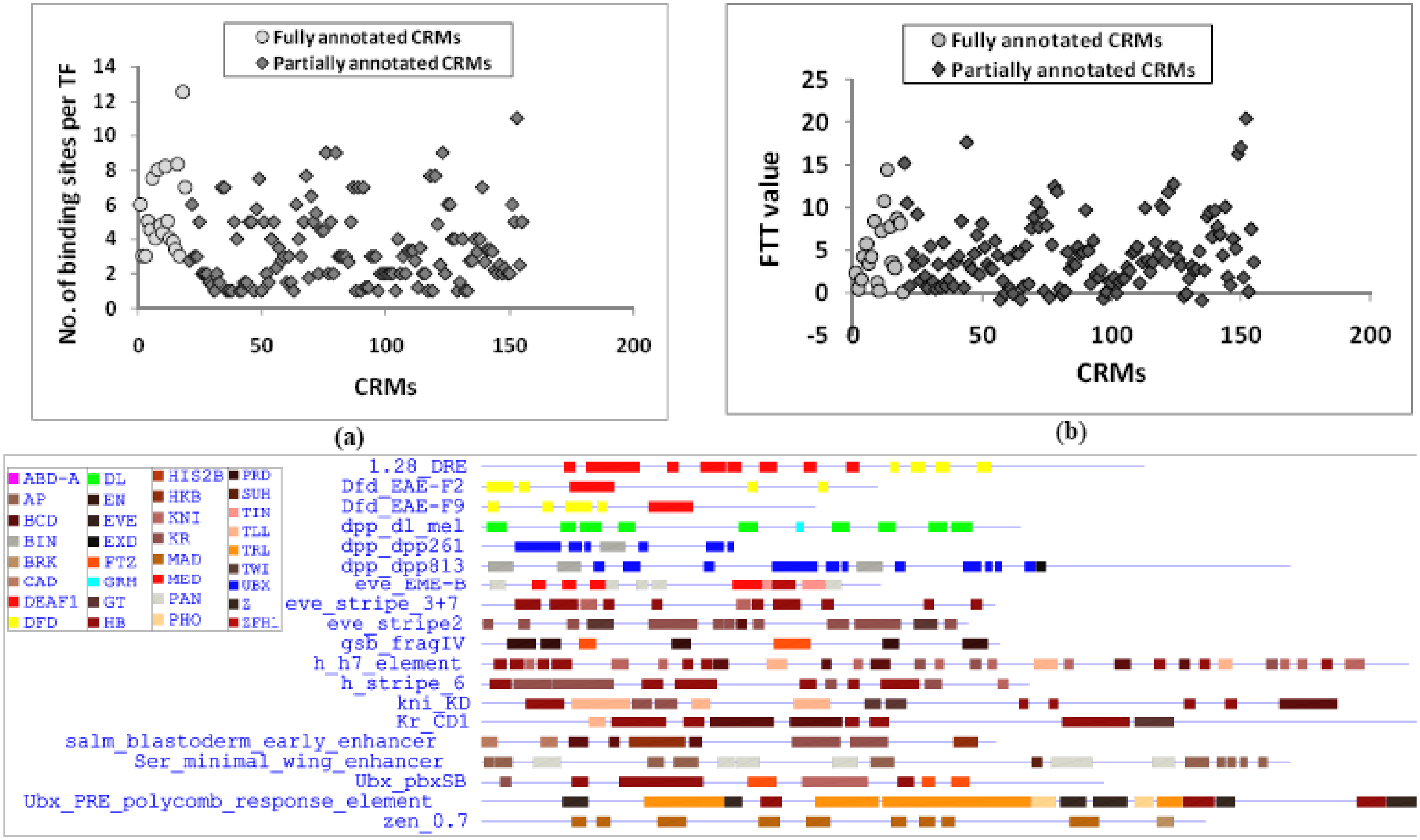


Figure VI-4. Drosophila CRMs have high redundancy of transcription factor binding sites. The number of binding sites per transcription factor in a CRM is shown in (a) for 19 CRMs having full experimental TFBS annotation (average 5.4 binding sites per TF) and 136 partially annotated CRMs (average 3.6 binding sites per TF). The fluffy tail test (FTT) scores [Abnizova et al. (2005)] for these sequences are shown in (b). The sequences were repeatmasked before computing the FTT to eliminate tandem repeats that may erroneously cause a high FTT value. FTT scores of most CRMs are greater than 2.0, indicating significant redundancy. The FTT scores of fully and partially annotated CRMs are similar, indicating that partially annotated CRMs may have greater redundancy than observed in the partial annotation. The full annotation of 19 CRMs is shown in (c).

to the left (or right) of *B*. We characterized the TFBSs with dyad motifs individually in each CRM. The procedure has the following steps, which are illustrated by an example in Figure VI-5:

(1) We first find all oligonucleotides of length 6 over the alphabet {A,C,G,T} which are over-represented in the CRM as compared to the background. The over-representation of an oligonucleotide is measured by the Z-score formula:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}},$$

where n1 and n2 are its number of occurrences in the CRM and background respectively (allowing one mismatch), N_1 is the CRM length, N_2 is the total length of background sequences, $\hat{p}_1 = \frac{n_1}{N_1}$, $\hat{p}_2 = \frac{n_2}{N_2}$, and $\hat{p} = \frac{n_1 + n_2}{N_1 + N_2}$.

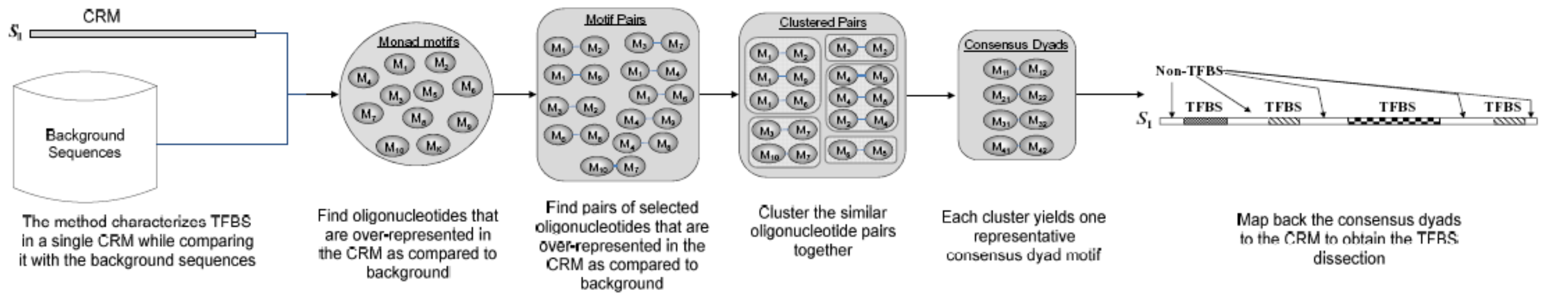
(2) From the selected oligonucleotides, we find pairs occurring at short distance (0 to 15 bp gap) from each other. Over-represented oligonucleotide pairs are selected again by Z-score. The selected pairs are then clustered according to their similarity. For clustering, the highest scoring unclustered pair is chosen as a new cluster center, and any other pair with at most two mismatches with it is added to the new cluster. This procedure is repeated until all pairs are clustered. Clusters of size less than 5 are dropped.

(3) Using the clusters of oligonucleotide pairs, we identify redundant site in the CRM. A redundant site will usually give rise to several similar over-represented oligonucleotides in a CRM. Thus we mark sites where more than 90% of the oligonucleotide pairs in the same cluster simultaneously match as redundant sites.

Figure VI-5. Over the next three pages, the figure illustrates the novel procedure used in Modulexplorer for characterizing TFBSs de-novo in a CRM.

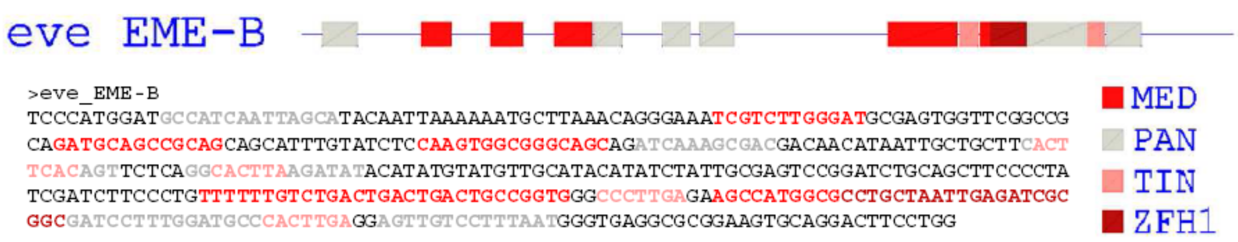
Modulexplorer characterizes TFBS de-novo in an individual CRM by extracting repeating or redundant sites. It represents these sites by spaced dyad motifs. Figure 2(a) summarizes the method.

Extraction of TFBSs from a CRM using de-novo motif discovery


S_1
CRM
Background Sequences
The method characterizes TFBS in a single CRM while comparing it with the background sequences
Monad motifs
M_1 M_2 M_3 M_4 M_5 M_6 M_7 M_8 M_9 M_{10} M_K
Find oligonucleotides that are over-represented in the CRM as compared to background
Motif Pairs
Find pairs of selected oligonucleotides that are over-represented in the CRM as compared to background
Clustered Pairs
Cluster the similar oligonucleotide pairs together
Consensus Dyads
M_{11} M_{12} M_{21} M_{22} M_{31} M_{32} M_{41} M_{42}
Each cluster yields one representative consensus dyad motif
Non-TFBS
TFBS
TFBS
TFBS
TFBS
S_1
Map back the consensus dyads to the CRM to obtain the TFBS dissection

2(a)

The following example illustrates the above TFBS characterization procedure. The example is based on the EME-B enhancer of the even skipped gene, which is shown in Figure 2(b) below. This enhancer contains 16 binding sites for 4 different TFs.



```
>eve_EME-B
TCCCATGGATGCCATCAATTAGCATACAATTAAAAAATGCTTAAACAGGGAAATCGTCTTGGGATGCGAGTGGTTCGGCCG
CAGATGCAGCCGCAGCAGCATTTGTATCTCCAAGTGGCGGGCAGCAGATCAAAGCGACGACAACATAATTGCTGCTTCACT
TCACAGTTCTCAGGCACTTAAGATATACATATGTATGTTGCATACATATCTATTGCGAGTCCGGATCTGCAGCTTCCCCTA
TCGATCTTCCCTGTTTTTTGTCTGACTGACTGACTGCCGGTGGGCCCTTGAGAAGCCATGGCGCCTGCTAATTGAGATCGC
GGCGATCCTTTGGATGCCCACTTGAGGAGTTGTCCTTTAATGGGTGAGGCGCGGAAGTGCAGGACTTCCTGG
```

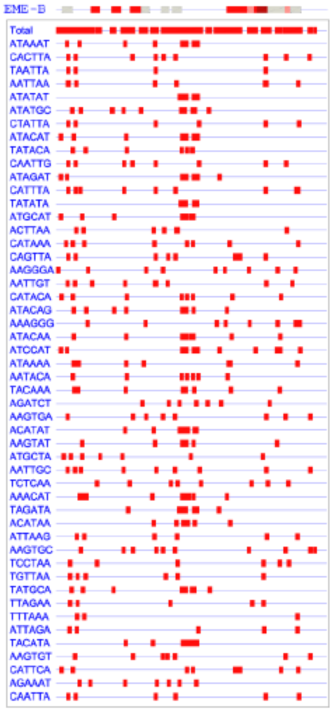
2(b)

Step 1

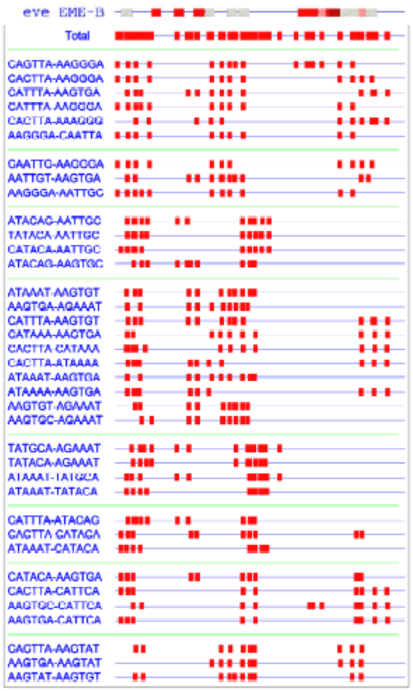
The TFBS extraction procedure begins with obtaining the set of most over-represented (l=6, d=1) oligonucleotides. The over-representation of oligonucleotides is measured relative to a set of background sequences. Figure 2(c) shows the over-represented oligonucleotides in the eve EME-B enhancer and marks the positions where each of these oligonucleotides occurs as red boxes. The line "Total" shows the union of all positions where these oligonucleotides occurs.

Step 2

From the selected oligonucleotides, pairs of oligonucleotides occurring at short distance from each other (0 to 15 bp gap) are searched and the most over-represented pairs are selected. Then the selected pairs are clustered according to their sequence similarity. The top scoring pair is the cluster center in each cluster, and any pair in the cluster will have at most two mismatches with the cluster center. Figure 2(d) shows the clusters of over-represented pairs. The occurrence positions of each pair are marked as red boxes.



2(c)



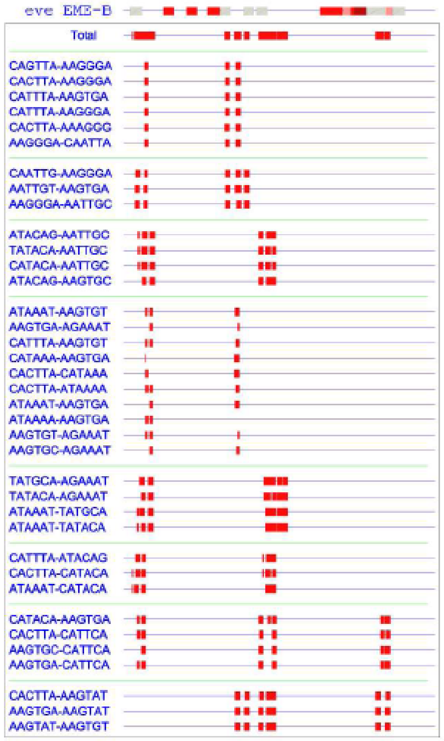
2(d)

Step 3

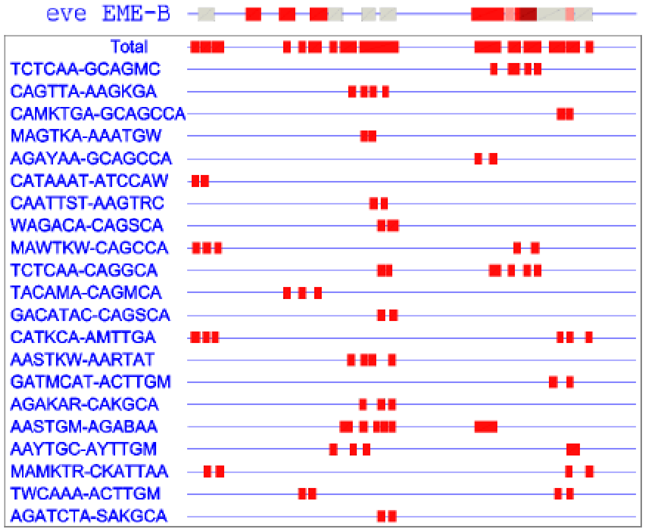
For each cluster of oligonucleotide pairs, we mark positions in the CRM where at least 90% of the pairs in the cluster are simultaneously mapped. These positions occur frequently in the over-represented oligonucleotides. Therefore they represent conserved sites. We call these sites as the *redundant sites*. Figure 2(d) shows redundant sites derived from 8 clusters of motif pairs in the eve EME-B enhancer. In this figure the number of redundant sites is small since only the top 8 motif clusters are shown. Redundant sites from the full set of oligonucleotide clusters is shown in 2(e).

Step 4

The oligonucleotide pairs in a cluster are aligned to derive a consensus motif which closely represents all the oligonucleotide pairs within the cluster. The consensus motif for a cluster closely represents the redundant sites in the CRM. Consensus dyads extracted from 21 motif clusters are shown in Figure 2(f). The union of their sites, shown in the row labeled "Total", closely match with the known TFBS in the enhancer.



2(e)



2(f)

(4) We represent each cluster of oligonucleotide pairs by a single consensus dyad motif derived by aligning together all oligonucleotide pairs in the cluster. The consensus dyad motif represents the redundant sites or the TFBSs in the CRM.

VI-3.2 Bayesian network model

The Modulexplorer Bayesian network model as shown in Figure VI-2 describes a CRM as a combination of dyad motifs with mutual order and distance constraints. The dyad motifs D_i, $i \in \{1, 2, \ldots, K\}$, with the states 0 (absent) or 1 (present), are the parent nodes while their common child node is the CRM node *Y* with states True (CRM) or False (non-CRM). Each dyad motif node D_i itself has two parent monad motif nodes $\left(M_{i,1}, M_{i,2}\right)$, where the nodes $M_{i,j} : i = 1, 2, \ldots, K; j = 1, 2$ take the states 0 or 1 depending upon whether the motif $M_{i,j}$ is absent or present in the sequence respectively. The node D_i has a noisy-AND relationship with its parent monad motif nodes. The noisy-AND relationship is implemented as described in [Vomlel (2006)]. The intermediate dummy variables $M'_{i,1}, M'_{i,2}$ are inserted between the node D_i and its parents $M_{i,1}, M_{i,2}$. The dummy variables also take the states 0 and 1. The relationship between D_i and the dummy variables $M'_{i,1}, M'_{i,2}$ is a deterministic AND. However, the dummy variables depend stochastically upon the actual variables $M_{i,1}, M_{i,2}$ as $\Pr\left(M'_{i,j} = 1 \middle| M_{i,j} = 0\right) = \alpha^0_{i,j}$ and $\Pr\left(M'_{i,j} = 1 \middle| M_{i,j} = 1\right) = \alpha^1_{i,j}$. Thus D_i depends stochastically upon the motif nodes as $\Pr\left(D_i = 1 \middle| M_{i,1} = a, M_{i,2} = b\right) = \alpha^a_{i,1}\alpha^b_{i,2}$, where $a, b \in \{0,1\}$.

Additionally the nodes *O_i* and *d_i* impose distance and order constraints upon the monad motifs in the dyad *D_i*. The order constraint O_i defines a bias in the relative positions (left or right) of the monad motifs $M_{i,1}$ and $M_{i,2}$ whenever they both occur in a CRM, whereas the distance node d_i models the distance between the adjacent occurrences of the pair of motifs $M_{i,1}$ and $M_{i,2}$. The order node has states ‘left’ or ‘right’, while the distance is discretized into two levels – ‘low’ and ‘high’ – with distance up to 6 bp considered ‘low’ and above that as ‘high’. The order and distance nodes are have the conditional probabilities $\Pr\left(O_i = left \middle| D_i = 0\right) = 0.5$, $\Pr\left(O_i = left \middle| D_i = 1\right) = \omega_i$, $\Pr\left(d_i = low \middle| D_i = 0\right) = 0.5$, $\Pr\left(d_i = low \middle| D_i = 1\right) = \delta_i$.

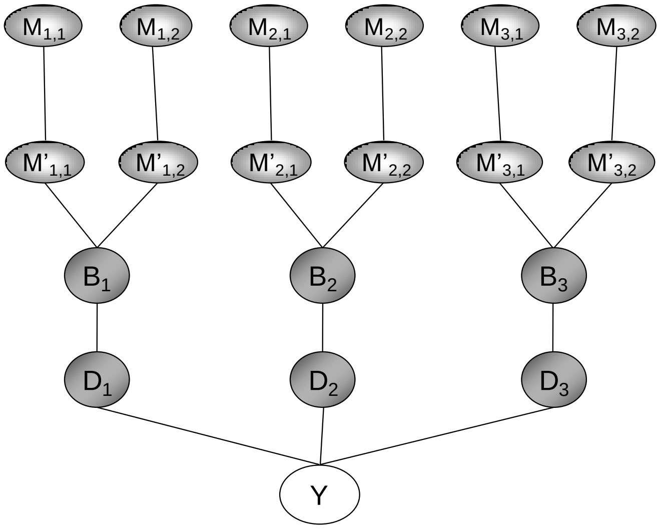
The Bayesian network encodes the joint probability:

$$\begin{aligned}&\Pr\left(Y, D_1, \ldots, D_K, M'_{1,1}, M'_{1,2}, \ldots, M'_{K,1}, M'_{K,2}, M_{1,1}, M_{1,2}, \ldots, M_{K,1}, M_{K,2}, O_1, \ldots, O_K, d_1, \ldots, d_K\right)\\ &= \Pr\left(Y \middle| D_1, \ldots, D_K\right) \prod_{i=1}^{K} \Pr\left(D_i \middle| M'_{i,1}, M'_{i,2}\right) \Pr\left(M'_{i,1} \middle| M_{i,1}\right) \Pr\left(M'_{i,2} \middle| M_{i,2}\right) \Pr\left(M_{i,1}\right) \Pr\left(M_{i,2}\right) \prod_{i=1}^{K} \Pr\left(O_i \middle| D_i\right) \Pr\left(d_i \middle| D_i\right)\end{aligned}$$

which contains the parameters $\alpha^0_{i,j}, \alpha^1_{i,j}, \omega_i, \delta_i$ and $\Pr\left(Y \middle| D_1, \ldots, D_K\right)$

In the training of the Bayesian network model, first we learn the parameters ω_i, δ_i of order and distance nodes directly from the training CRMs. For each occurrence of the dyad motif *D_i* in the training CRMs, the order of occurrence of its monad motif parents and their distance is identified. The frequencies of these occurrences are used to compute the probabilities ω_i and δ_i. These parameters are henceforth kept fixed.

Thereafter the noisy AND parameters $\alpha^0_{i,j}, \alpha^1_{i,j}$ and the parameters $\Pr\left(Y \middle| D_1, D_2, \ldots, D_K\right)$ are estimated. The order and distance nodes are temporarily removed. The reduced model is shown as an undirected graph in Figure VI-6. Each training sequence, $S_n \mid n = 1, 2, \ldots, N$, is represented as an ordered pair $\left(\vec{m}^{(n)}, y^{(n)}\right)$, where $\vec{m}^{(n)} = \left(m^{(n)}_{1,1} \quad m^{(n)}_{1,2} \quad \ldots \quad m^{(n)}_{K,1} \quad m^{(n)}_{K,2}\right)$ is a binary vector representing the presence or absence of each of the monad motifs in the sequence, and $y^{(n)}$ is a label 1 or 0 depending upon whether the sequence is a CRM or non-CRM respectively. We use factorization of the probability potentials to achieve efficiency in training [Vomlel (2006)]. The hidden



$\phi_{B_k M'_{i,j}}$	B_k=0	B_k =1
$M'_{i,j}$ =0	+1	0
$M'_{i,j}$ =1	+1	+1
$\phi_{B_k D_k}$	B_k =0	B_k =1
D_k =0	+1	0
D_k =1	–1	+1

Figure VI-6. Potentials $\Pr\left(D_i \middle| M_{i,1}, M_{i,2}\right)$ factorized using the hidden nodes B_i.

nodes $B_1, \ldots, B_K$ in the undirected graph serve to factorize the probability potentials. The expectation maximization (EM) algorithm is used to learn the model parameters from the data $\left(\vec{m}^{(n)}, y^{(n)}\right)$. The EM equations are written as [Vomlel (2002), Vomlel (2006)]:

M step:

$$\alpha_{i,j}^{a} = \frac{n\left(M'_{i,j}=1, M_{i,j}=a\right)}{n\left(M_{i,j}=a\right)}, \quad \Pr\left(Y=b \middle| \mathbf{D}=\mathbf{d}\right) = \frac{n\left(\mathbf{D}=\mathbf{d}, Y=b\right)}{n\left(\mathbf{D}=\mathbf{d}\right)},$$

where $a, b \in \{0,1\}$, $i = 1, 2, \ldots, K$, j=1,2 and **D** is the vector of all D_i

E-step:

$$n\left(\mathbf{D}=\mathbf{d}, Y=b\right) = \sum_{n=1}^{N} \begin{cases} \Pr\left(\mathbf{D}=\mathbf{d} \mid \vec{m}^{(n)}\right) & \text{if } y_n = b \\ 0 & \text{otherwise} \end{cases},$$

where $\Pr\left(\mathbf{D}=\mathbf{d}\mid\vec{m}^{(n)}\right)=\prod_{i=1}^{K}\alpha_{i,1}^{m_{i,2}^{(n)}}\alpha_{i,2}^{m_{i,2}^{(n)}}$,

$$n\left(M'_{i,j}=1,M_{i,j}=a\right)=\sum_{n=1}^{N}\begin{cases}\Pr\left(M'_{j}=1\mid\vec{m}^{(n)}\right) & \text{if } m_{i,j}^{(n)}=a\\ 0 & \text{otherwise}\end{cases},$$

$$\Pr\left(M'_{i,j}\mid\vec{m}^{(n)}\right)\propto\prod_{k=1}^{K}\left[\sum_{\mathbf{D}}\Pr\left(Y=y^{(n)}\mid\mathbf{D}\right)\sum_{B_k}\left[\begin{matrix}\phi_{B_kD_k}\left(B_k,D_k\right)\phi_{B_kM'_{i,j}}\left(B_k,M'_{i,j}\right)\Pr\left(M'_{i,j}\mid M_{i,j}=m_{i,j}^{(n)}\right)\\ \left\{\sum_{M'_{i,x}}\phi_{B_kM'_{i,x}}\left(B_k,M'_{i,x}\right)\Pr\left(M'_{i,x}\mid M_{i,x}=m_{i,x}^{(n)}\right)\right\}\end{matrix}\right]\right],$$

where $M_{i,x}$ is the pair of $M_{i,j}$, and the potentials ϕ are shown in Figure VI-6.

After training, the Bayesian network CRM model can infer whether or not a given sequence is a CRM based on its motif content. The sequence is scanned to ascertain which of the 2*K* motifs $M_{i,j}, i=1,2,\ldots,K, j=1,2$ occur in the sequence, as well as the order and distance between the adjacent motifs. This evidence is provided to the Bayesian network and a standard inference algorithm is used to assign a value between 0 and 1 at the “CRM” node, which is the estimated probability of the given sequence being a CRM. To predict CRMs in a long uncharacterized sequence, a sliding window approach is used.

VI-3.3 Feature Based Clustering of CRMs

The aim of feature based clustering is to find clusters of CRMs having a common set of motifs. The dyad motifs are called “items” and the set of all dyad motifs that match a given CRM is called the “itemset” for that CRM. The *Closet* algorithm [Wang et al. (2003)] is used to determine the closed maximal subset of items that are common to at least *T* itemsets. This translates to finding the maximal set of motifs that are common to at least *T* CRMs. The number *T* is called *support*. In a single run, the *Closet* algorithm outputs all possible clusters of at least *T* itemsets (CRMs) along with their maximal common set of items (motifs). The *fitness* of a cluster is defined as the inverse of the probability of obtaining the cluster by chance. Let *S* be the total number of itemsets (or CRMs) and *M* be the total number of distinct items (motifs). Then,

$$\Pr\left(\text{at least } T \text{ itemsets contain all items } I\right)=\sum_{t=T}^{S}\binom{S}{T}\left[P\left(I\right)\right]^{t}\left[1-P\left(I\right)\right]^{S-t},$$

where $I=\left\{i_1,i_2,\ldots,i_N\right\}$ are the *N* items in the cluster and $P\left(I\right)$ is the probability that all these items are selected. If p_k is the frequency of the item i_k in all itemsets, then $P\left(I\right)=\prod_{k=1}^{N}p_k$. Fitness is taken as negative log of this probability. We run *Closet* for different values of support *T*. The fitness is a convex function of *T*. We select the cluster with the highest fitness across all *T*.

After obtaining a CRM cluster, we remove all these CRMs from the list and run the *Closet* algorithm again on the reduced set. Thus we get the next most conserved cluster. This iterative procedure finds several conserved clusters of CRMs.

VI-3.4 Derivation of Regulatory Code

By the abovementioned feature based clustering, CRMs having a common set of motifs are obtained. This common set of motifs is used to derive the regulatory code. The regulatory code is obtained as a minimal subset of the common motifs that can effectively discriminate between the CRMs in the cluster and the background sequences or other CRMs. We first translate each consensus motif to a PWM using all its occurrences within the CRMs in the cluster. Then we analyze the PWMs of all the common motifs with STAMP tool [Mahony and Benos (2007)], which computes the similarities among the motifs and hierarchically clusters them. The hierarchical cluster is shown as a phylogenetic tree. With a certain similarity cutoff, we separate the motifs into distinct clusters. From each cluster, we select one representative motif that is most over-represented in the CRMs in the cluster as compared to background sequences or other CRMs. The selected motifs comprise the minimal regulatory code. The regulatory code motifs can be used to discriminate CRMs in the cluster from other CRMs and background. The discrimination is based on the total count of matches of the motifs in a 1 kb window. The thresholds for the PWMs [Stormo (2000)] are fixed according to the number of random matches produced in a set of background sequences. The value chosen in this study is 5×10^{-4} probability of random match, i.e. 1 random match per 2 kb of sequence.

VI-4 Training of Modulexplorer

As the first step in the training of Modulexplorer we annotated TFBSs de novo in all 619 CRMs using the method described in Section VI-3.1. Within 19 CRMs which are fully experimentally annotated with TFBSs (Section VI-2, Figure VI-4), the predicted TFBSs overlapped 81% of the experimental TFBSs. In classifying each base in the sequence as TFBS or non-TFBS, the ROCs for the 19 fully annotated CRMs are shown in Figure VI-7. The overall sensitivity and false positive rate are 81.5% and 22% respectively. The p-value for this correlation is 9×10^{-32} compared with random sites of the same length. Thus the method robustly identified the TFBSs in the CRMs de novo.

To learn the Modulexplorer Bayesian network model, we obtained a non-redundant and non-overlapping set of 356 training CRMs and 58 test CRMs from the REDfly database as described in the Data section above. The CRMs were of several different types as described in Figure VI-3. In addition, three different background sets

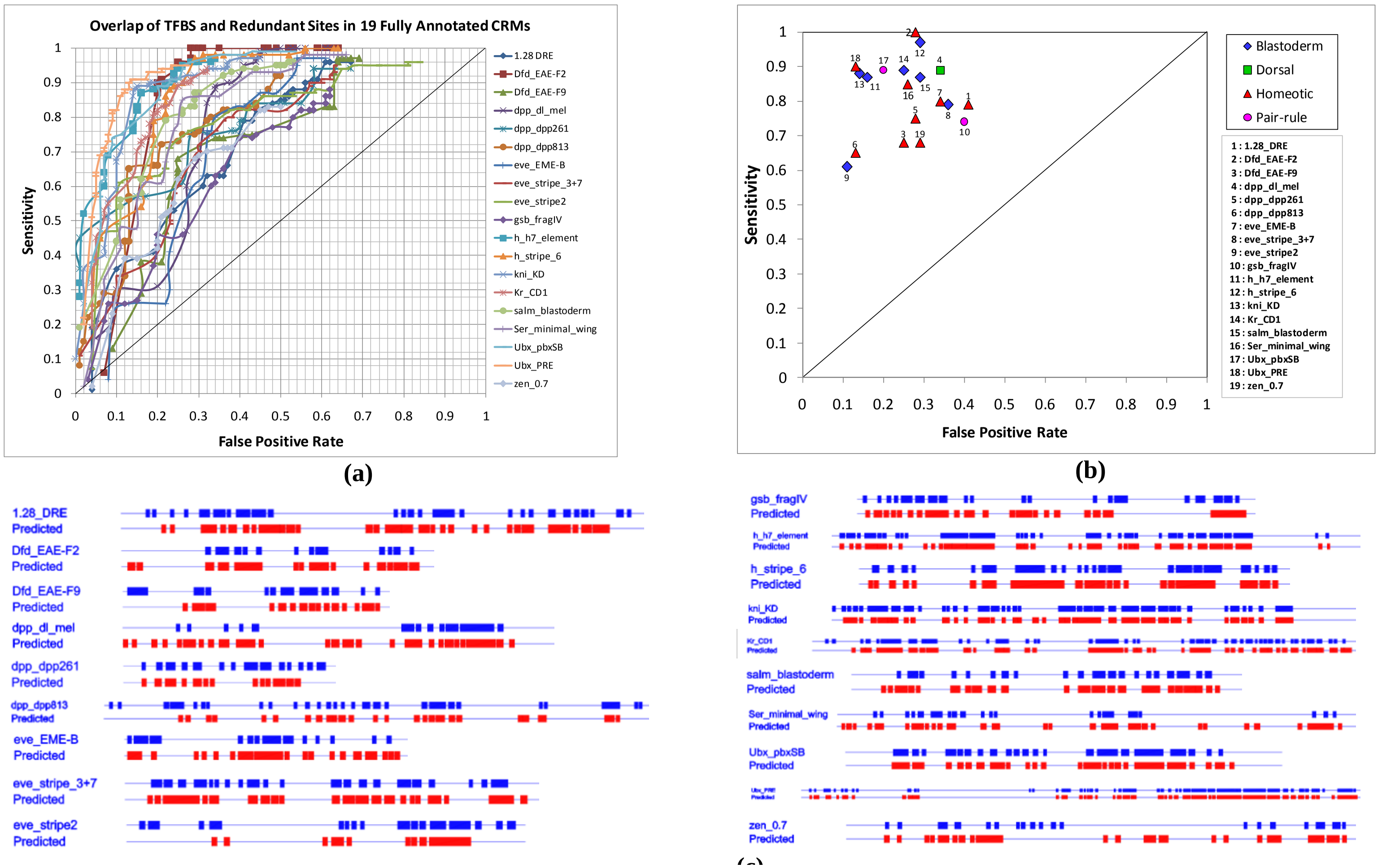


Figure VI-7. The TFBSs in Drosophila CRMs appear as repeated or redundant sites. Modulexplorer locates these redundant sites as potential TFBSs. The receiver-operating characteristic of predicting TFBSs using redundant sites in 19 fully annotated CRMs is shown in (a). Here sensitivity (y-axis) refers to the % of nucleotides in TFBSs that are overlapped by some redundant site, while false positive rate (x-axis) refers to the % of nucleotides in a redundant site that do not match any TFBS. The maximum effectiveness of TFBS characterization in each of the 19 CRMs is shown in (b), which is the point in the ROC curve where Matthew's correlation coefficient is maximized. At this maximum effectiveness, the visual overlap between the TFBS sites (blue boxes) and the redundant sites (red boxes) in each CRM is shown in (c).

from exon, intron and intergenic sequences respectively were prepared as discussed in the Data section.

Ten-fold cross-validation training of the Bayesian network was performed with this training data. The discrimination achieved between CRM and background sequences in cross-validation training is shown in Figure VI-8. The result is compared with the current best performing algorithm HexDiff [Chan and Kibler (2005)] and with a Markov model. Other CRM prediction algorithms (such as [Rajewsky et al. (2002); Bailey and Noble (2003); Sharan et al. (2004)]) could not be included for comparison since they require prior biological knowledge of the CRM model (such as the PWMs of the TFs) and are specific to a subset of CRMs of the same type. As shown in Figure VI-8(a), all three models showed high discrimination between CRM and coding (exon) sequences. However for non-coding (intron and intergenic) sequences (Figure VI-8(b)), the Markov model showed no discrimination (area under ROC=0.37), HexDiff showed marginal discrimination (area under ROC=0.58), while Modulexplorer had the highest discrimination (area under ROC=0.75).

Modulexplorer's prediction performance was then evaluated on a test dataset of 58 CRMs and 1000 random background sequences different from the training set. The test set is unbiased as it contains CRMs expressed in a variety of tissues and stages and is distinct from the training sequences (Figure VI-8(c)). The ROC, shown in Figure VI-8(d), resembles the performance on the training set (area under ROC=0.72).

VI-5 Pairwise TF-TF Interactions Learnt De-novo by the Modulexplorer

We investigated the conditional probabilities associated with the edges of the Modulexplorer Bayesian network model to obtain insight into the pairwise TFBS

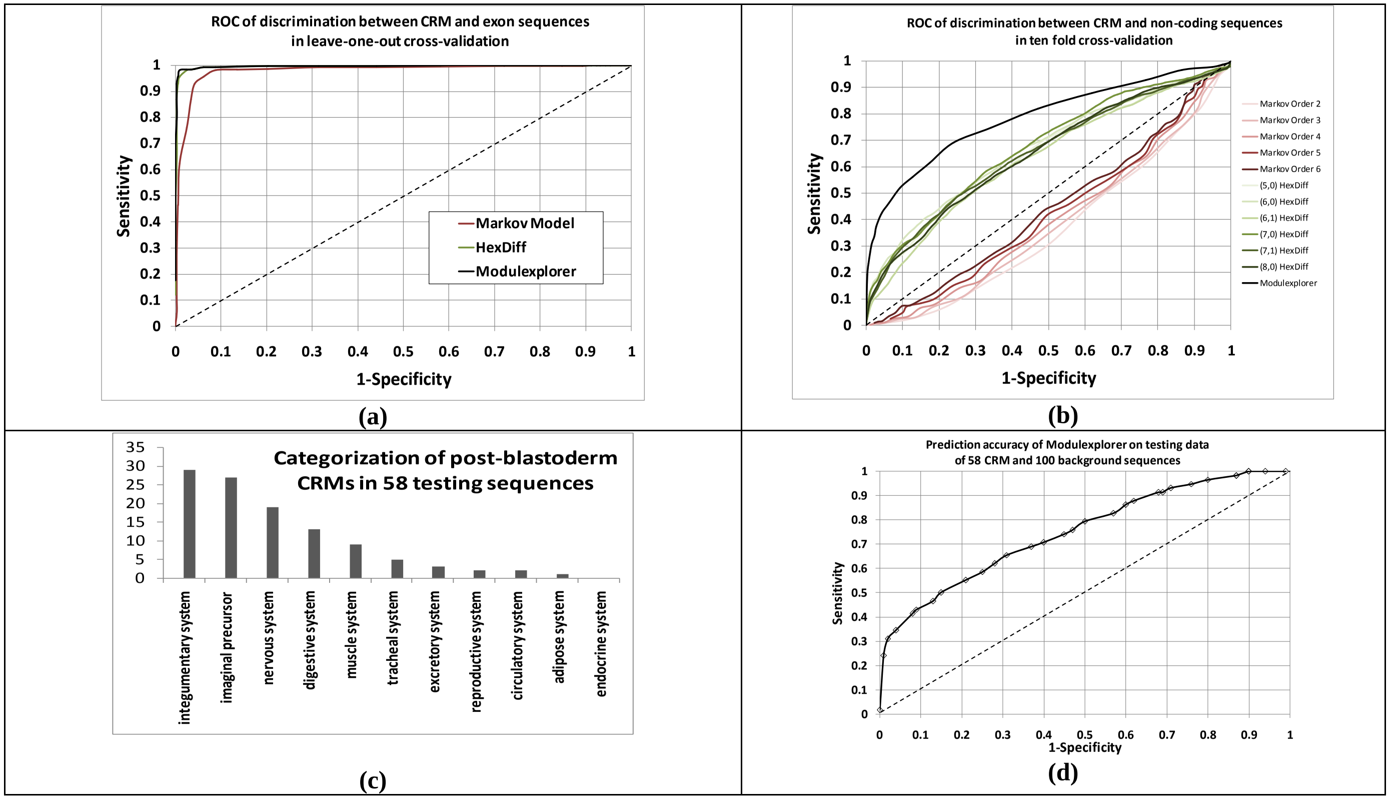


Figure VI-8. Performance of the Modulexplorer in discriminating between CRM and background sequences. Modulexplorer's performance is compared with two other methods: a Markov model (orders 2 to 6) and the HexDiff algorithm [Chan and Kibler (2005)]. The original Hexdiff algorithmuses (6,0) motifs, but it was extended in this comparison to try several different (l,d) motifs. Discrimination achieved between training CRMs and exon sequences in 10-fold cross-validation is shown in (a). The ROC shows that all three methods could easily discriminate CRMs from exons. Discrimination between CRMs and non-coding sequences (intron+intergenic) is shown in (b). Here Markov model shows no discrimination, HexDiff has marginal discrimination, while Modulexplorer achieves maximum discrimination. Modulexplorer was further evaluated on a separate testing set of 58 CRMs. The number of CRMs of different types in the test set according to their stage and tissue of expression is shown in (c). The performance of Modulexplorer on this test set, shown in (d), is similar to the training performance.

interactions represented in the model. We took from the model the representative motifs for 61 known TFs which best matched their known binding sites (listed in Figure VI-9). The strength of interaction between any two motifs M_1, M_2 was measured as the ratio of the marginal probabilities $\Pr\left(CRM \middle| M_1, M_2\right) \big/ \Pr\left(non-CRM \middle| M_1, M_2\right)$. The pairwise interaction matrix is shown in Figure VI-10. Based on the interaction matrix, the TFs were hierarchically clustered using UPGMA algorithm [Sokal and Michener (1958)].

According to their known biological functions, the 61 TFs may be grouped into

ABD-A	MAATTG-AATGGG
ABD-A	AAATTG-AATGGG
ABD-B	RTAAAA-AAWWTG
ABD-B	TCAAAA-AAYSRTA
ADF1	TKCGMA-AGCSGCTC
ADF1	CTGCG-CYGWWCA
AEF1	CTACTA-AATCBG
AEF1	AATCAG-GTACAA
ANTP	AAATAT-AWWWTG
ANTP	TAAAAA-ATWWAT
AP	AAATAA-AATKAT
AP	AATAA-AATTGC
ARA	ATWWAA-ATCAAA
ARA	GAAATA-AASTTR
BAP	MTTSAA-AATCGCA
BCD	ATTAAA-AWWWTS
BCD	AWWWTS-AAAATYY
BIN	AATCAA-AAATAG
BIN	TAACAA-GCAGACG
BRK	GAAAMC-GACAGCT
BRK	CGCKAG-ATTTSC
BYN	WTAAAA-AGTTGA
BYN	TDYAAA-CTGCTA
CAD	TAAAAA-ATAAMA
CAD	TAAAAA-AAYTAT
CF2-II	RTWWWA-CCAGAC
CF2-II	CATWTA-ACGCTAA
DEAF1	GCAAAA-AATCGM
DEAF1	RTYWAA-RAGTCA
DFD	AAATTA-ATWWWA
DFD	ATWAAT-AAAYTA
DL	AAAATA-CARAAA
DL	ATWWWA-CKAAAA
DSX-F	AATCA-GACTACA
DSX-F	TRATCA-CACAAAT
DSX-M	AATCA-GACTACA
DSX-M	TRATCA-CACAAAT
EMS	TCAAAA-ARTGWCA
EMS	TCAAAA-AAYSRTA
EN	AATAAA-AAATGT
EN	AMAWKKA-ATCAAA
ESPL	AGTAAAA-ACMAAT
ESPL	AAAAATM-AGCAAA
EVE	AATAAA-TRWTAA
EVE	AATAAA-AAATGT
EXD	MAATTG-AATGGG
EXD	AATTAG-TCCWAA
EY	TSSSAA-AAGTCA
EY	TSSSAA-AAATGA
FTZ	ATAAAA-AAYTAT
FTZ	ATAAAA-TRTAAA
FTZ-F1	CAATTA-ATTGTC
FTZ-F1	AMYTARG-ATKGTC
GL	ATTSTG-GRAGAA
GL	AGGAAT-ACABAT
GRH	ATAAAA-GAATAA
GRH	AATGA-CTTTCC
GT	AAAASA-AAAGGY
GT	TAAAAA-CCGCGA
HB	AAAAAA-CTAAAA
HB	AAAAAA-ATAMAA
HIS2B	CCTAAG-ACGCTG
HIS2B	AGGTA-AGCTGGA
HKB	GGAHWWAKC-CCACGC
HKB	ACAAWT-KGCAAA
KNI	TAAAAA-AWWWTG
KNI	AAAAAT-ATTAAA
KR	CAAWTC-AAATGG
KR	AAAWAG-CVAAAA
MAD	CCGWCGC-SKCGMM
MAD	MGCGACM-SKCGMM
MED	MASTKA-ATMCAT
MED	TCGAGAC-GKCGMA
NUB	GCCAAA-AATCAR
NUB	CATMGA-GCCAAA
OVO	TTAAAAA-ACAAKA
OVO	TAGAAA-AAWGGA
PAN	TSAAAA-AWSAAA
PAN	ACAAAT-TSAAAA
PHO	ATAAAA-GAAATAC
PHO	ACATAA-AAAATGA
PRD	AWWWTR-CCATGA
PRD	CRATTA-YGTCAAA
SD	ATTTAA-AAAAAT
SD	AAAAAT-AATGAA
SLBO	TGATMA-AYCWGV
SLBO	AATCA-GACTACA
SLP1	CWWHGA-AACACT
SLP1	MTBWSA-SGAGGAC
SNA	GCGAAA-ACGYRCG
SNA	CGGGAA-ACGYRCG
SNA	TGGAAA-GCCAYA
SO	ARGATG-TSAYMWC
SRP	AGCCAA-GCGAAA
SRP	AAWWWT-AGCCAA
SUH	TCGTAA-CAGAAA
SUH	CAGAAA-CATCGA
TIN	GATCCA-GCAGMC
TIN	TGSSMA-GAGAAA
TLL	CAAAAA-TCAAAA
TLL	TAAAAA-TCAAAA
TOY	SGWWWC-GGRGAA
TOY	TSSSAA-AAGTCA
TRL	AWWWTG-AATAAA
TRL	ASATAA-AAAAGW
TTK	GCAAAA-CCYGCG
TTK	GAAGGA-CGAASG
TWI	TCAAAA-AAGGCC
TWI	TATGGA-ATGCAA
UBX	AATAAA-TRWTAA
UBX	AATAAA-AAAAAT
VND	TTSAAA-AAGAKA
VVL	GMATKC-TCSTCA
VVL	AGKATG-ATCSTCA
Z	AAAASRA-ATRAAT
Z	TTAAAAA-AAATTA
ZEN	AATAAA-TRWTAA
ZEN	AATAAA-AAATGT
ZFH1	GCTTCCC-AAYTGC
ZFH1	CAKAAAT-CAMKTRA

Figure VI-9 Dyad motifs in Modulexplorer most closely resembling the binding sites of known TFs.

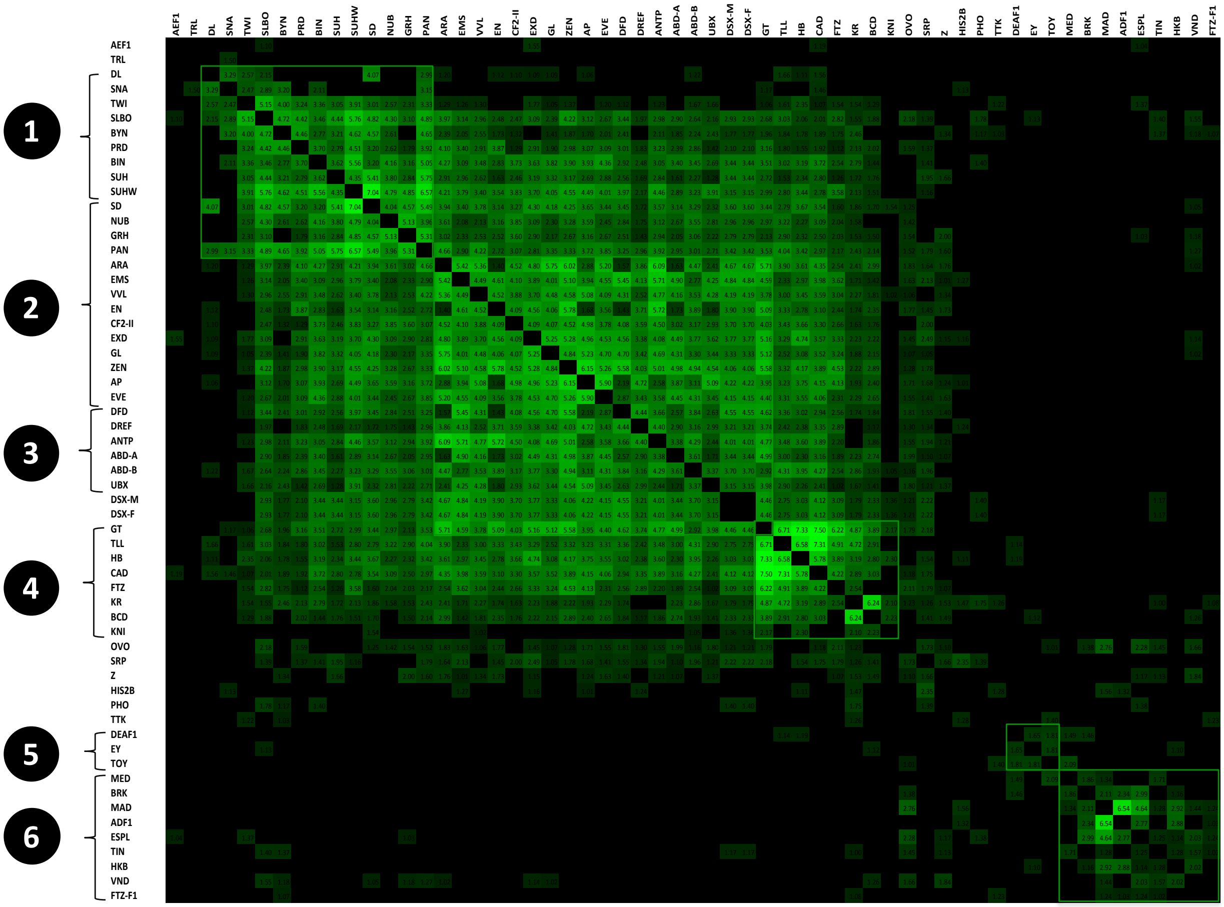


Figure VI-10 Pairwise interactions between 61 different TFs learnt de-novo by the Modulexplorer probability model. Based on the interaction matrix, the TFs were hierarchically clustered. Six functionally related groups of TFs were formed: (1) cofactors of twist in mesoderm and nervous system development, (2) TFs involved in imaginal disc development, (3) the antennapedia complex, (4) TFs expressed in the blastoderm, (5) TFs for eye development and (6) a miscellaneous set of TFs. Five distinct clusters are seen in the interaction matrix. Three of the clusters contain mixed set of TFs from groups 1-4, while two other clusters correspond to the TF groups 5 and 6.

five broad categories. The TF Twist (*twi*) and its cofactors *dl*, *sna*, *byn*, *slbo*, *prd*, *bin*, *su(H)*, *su(Hw)* in mesoderm and nervous system development [Kusch and Reuter (1999); Furlong et al. (2001); Markstein et al. (2004); Borghese et al. (2006)] are placed in the first group. The TFs *sd*, *pan*, *nub*, *ap*, *grh* and *ara* which are involved in the development of imaginal discs such as the wing disc were placed in the second group. The third and fourth groups were formed by the TFs of the antennapedia complex (*Antp, abd-A, abd-B, ubx*, *dfd*) and the blastoderm (*bcd*, *hb*, *cad*, *kni*, *Kr*, *tll*, *gt*) respectively. The fifth group consists of the TFs *ey* and *toy* involved in eye development.

In the TF-TF interaction matrix, the first four TF groups showed high mutual interaction values in general. The overlap is expected as these TFs are known to function cooperatively [Mann and Morata (2000); Morata (2001)]. However, closer analysis of the interaction matrix by hierarchical clustering indicated that the TFs of these four groups form three distinct clusters in the TF-TF interaction matrix. The first cluster included all TFs from the first group and the TFs *sd*, *nub*, *grh* and *pan* from the second group. The second cluster contained the remaining TFs from the second group and all TFs of the antennapedia complex. In addition, the TFs *ems*, *vvl*, *en*, *exd*, *cf2-II*, *gl*, *Dref*, *zen* and *eve* also came together with this cluster according to the hierarchical clustering. There are some supports that these TFs may be related to the known factors in the second cluster. *exd*, *en*, *ems*, *zen* and *eve* are known regulators in the development of appendages including the legs and wings [Mann and Morata (2000)]. *Vvl* also has known function in wing development [de Celis et al. (1995)], while little information is available on the TFs *gl* and *cf2-II*. The primary function of *Dref*, is DNA replication. Though there is no support presently of its association with antennapedia complex, recent studies have shown its various diverse roles [Hirose et al. (2001)] and hence it might be related to the antennapeida complex.

The third cluster contains all the blastoderm TFs. Moreover the antennapedia pair-rule TF *ftz* was also found in this cluster. This association is not surprising as *ftz* cooperates with blastoderm TFs in known CRMs [Zhang et al. (1991)].

The TFs *ey* and *toy* of the fifth group appeared as a distinct cluster by the UPGMA clustering. The TF *deaf1* was also found in this cluster. Little is known in the existing literature about *deaf1*. Surprisingly we found from a survey of recent literature that *deaf1* seems to have a role in eye development [Veraksa et al. (2002)].

We found another separate cluster formed by a miscellaneous set of TFs *med*, *mad*, *brk*, *adf1*, *espl*, *tin*, *hkb*, *vnd* and *ftz-f1*. Some of these TFs have known interactions, e.g. *mad* and *brk* are co-regulators of *zen* [Rushlow et al. (2001)], while *mad* and *med* cooperate in the regulation of *bam* gene [Song et al. (2004)]. However the clustering of these TFs is a subject for further study.

In summary, TFs with high interaction probability in Modulexplorer were found to have close interaction with each other in the same biological process and developmental stages.

VI-6 Genome Wide Scan for Novel CRMs

The Modulexplorer model was used to search for novel CRMs within BDGP Release 5 assembly of the Drosophila genome. In a sliding window like approach, the complete 120 Mb genomic sequence was divided into 24,000 windows of length 1000 bp each with the adjacent windows overlapping by 500 bp. The Modulexplorer model assigns to each window a probability that it may contain a CRM. A small set of high confidence windows that were assigned high probability value by the model were shortlisted for analysis as shown in Figure VI-11(a). We chose a probability threshold so that the model has a small false positive rate of 1% in cross-validation. At this threshold the expected sensitivity is about 20%. Thus 240 false positive windows are expected in the predicted set. A total of 1298 windows were found above the threshold, which is more than 5-fold the number of expected false positives. The P-value for this recovery (Bonferroni corrected) is 4.0×10^{-16}. Out of 1298 windows, 472 windows were overlapping the training CRMs and 13 windows overlapped the 58 test sequences (Figure VI-11(b)). The remaining 813 windows are novel predictions. These novel predictions are listed in Supplementary Figure 4 at the end of this dissertation.

As an initial validation, the novel predictions were compared with computational CRM predictions reported by other authors [Berman et al. (2002); Markstein et al. (2002); Berman et al. (2004); Schroeder et al. (2004)] and with new CRMs added to the REDfly database in version 2. Mild overlap with these predictions was found as shown in Table VI-1. Out of 28 predicted CRMs reported by Berman et al. [Berman et al. (2002)], 6 were also reported by Modulexplorer. In a subsequent in-vivo validation by Berman et al. [Berman et al. (2004)], 9 of the 28 predicted CRMs were validated as active enhancers while the remaining 19's were not. Five of the six common predictions between Modulexplorer and Berman et al. corresponding to the genes *gt*, *odd*, *sqz* and *CG9650* (two overlapping Modulexplorer windows) were among the validated modules, while one prediction corresponding to the gene antp was inactive. Similarly 7 CRMs were common between Modulexplorer and the predictions of Markstein et al. [Markstein et al. (2002)] corresponding to the genes *run*, *zen*, *brk*, *sog*, *CG12444*, *osm-6* and *ady43a*. Of these, the *zen*, *brk*, *sog* and *ady43a* enhancers have been validated as active in-vivo. Nine other CRMs reported by Modulexplorer corresponding to the genes *fkh*, *sim*, *wg*, *mir-309*, *grh*, *phyl*, and *cluster_at_55C* are confirmed by the updated REDfly database (version 2), while four CRMs for the genes *gt*, *kni* and *pdm2* are validated by Schroeder et al. (2004).

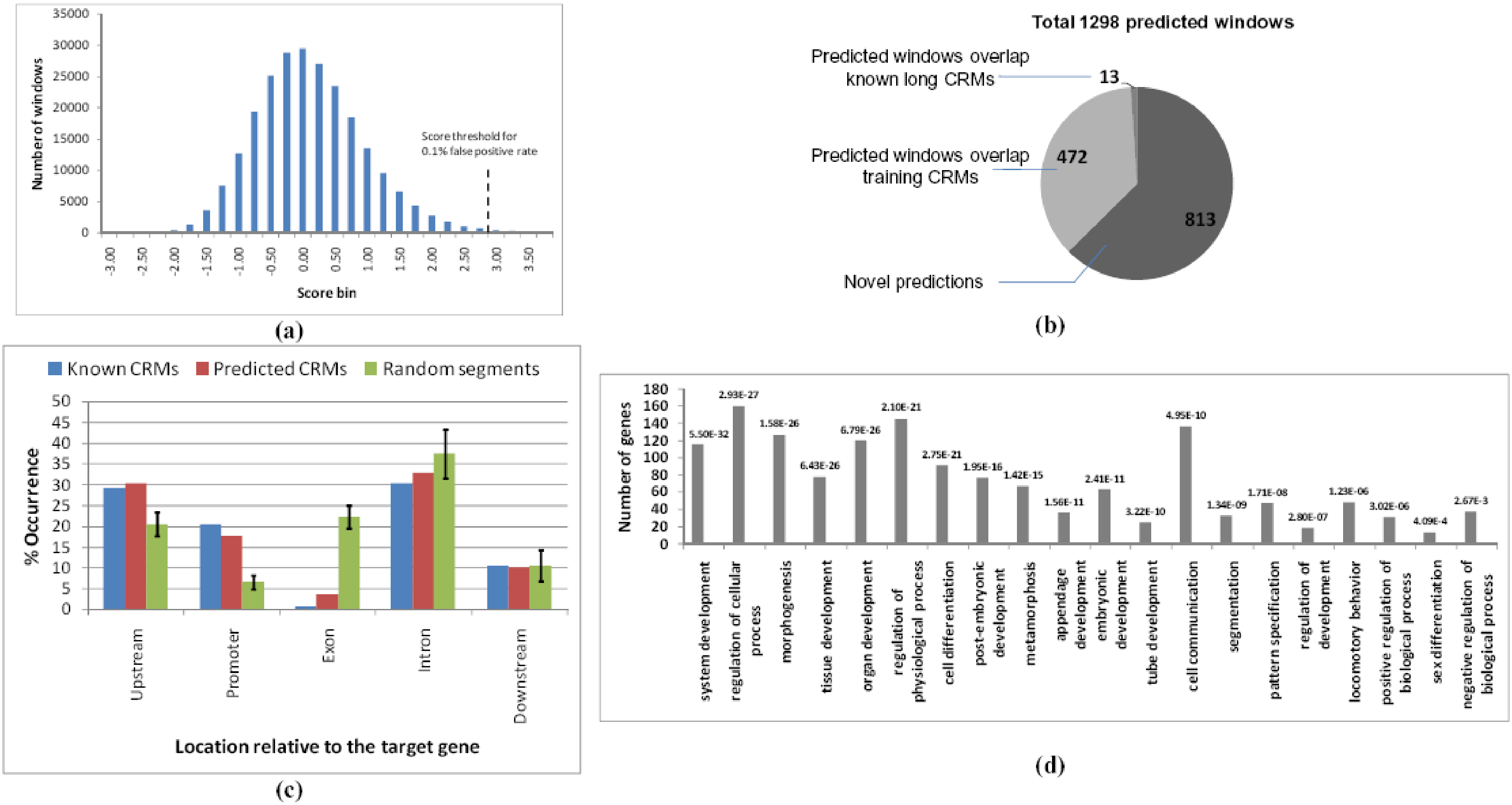


Figure VI-11. Summary of Modulexplorer's whole genome CRM predictions: (a) A stringent score threshold was used for shortlisting predicted CRM windows such that the false positive rate is about 0.1%. (b) A total of 1298 windows were predicted above the chosen threshold, out of which 813 are novel predictions. (c) The predicted CRMs are significantly over-represented in the promoter and upstream intergenic regions. (d) This is the list of level 3 gene ontology (GO) categories statistically over-represented in the target genes of the predicted CRMs. They show enrichment in development and regulatory functions (Bonferroni corrected P-values of the GO associations are shown alongside).

Table VI-1. Overlap of novel CRMs predicted by Modulexplorer with CRMs predicted in previous computational studies.

Reference	No. of CRMs predicted	No. of CRMs validated	No. of CRMs found active	No. of overlapping Modulexplorer predictions	No. of Modulexplorer CRMs validated	No. of Modulexplorer CRMs found active
Berman et al. (2002, 2004)	28	28	9	6	6	5
Markstein et al. (2002)	15	15	5	7	7	4
Schroeder et al. (2004)	32	20	15	16	5	4
REDfly version 2 (new)	-	-	34	9	9	9

The Modulexplorer predictions were over-represented in upstream regulatory regions of genes (Figure VI-11(d)) indicating a strong bias towards transcriptional control. Of the 813 predicted CRM windows, 391 (48.1%) fell in the upstream intergenic and promoter region, which is significantly higher (p-value = 1.1×10^{-182}) compared to randomly distributed size-matched segments (mean 26%, stdev 4.4% over 100 trials). Known CRMs show a similar bias, with 49.6% CRMs overlapping upstream intergenic and promoter regions (p-value = 1.1×10^{-226}). The known and predicted CRMs also show significant under-representation in the exon regions as compared to random segments.

In many cases, multiple predicted CRMs were found clustered around a gene. The trend is similar to that for known CRMs, where out of 619 known Drosophila CRMs, 398 occur as a cluster of 4 or more CRMs around 51 genes. Monte carlo simulations were performed to assess the statistical significance of the clustering of CRMs in intergenic gaps. The number of clusters of different sizes in 50kb windows formed by randomly distributing 813 segments of 1000 bp length across the Drosophila genome is shown in Figure VI-12 (averaged over 100 simulations). In comparison, the corresponding distributions for the predicted and known CRMs show significant clustering around their target genes.

Putative target genes were assigned to the predicted CRM windows based on proximity. Though a CRM can regulate distant genes, it is an uncommon occurrence, for instance, 81% of the known CRMs target their most proximal gene. In this study, CRMs lying within the intron of a gene were assigned the same gene as their target, whereas CRMs lying in the intergenic region were assigned both the closest upstream and downstream genes as their possible targets. Gene ontology classification of the target genes obtained using the online tool GOToolBox [Martin et al. (2004)] is shown in Figure VI-11(d) with the GO terms sorted according to their significance (Bonferroni corrected P-value). The terms show highly significant enrichment in the GO categories related to development and gene regulation (morphogen activity) [Martin et al. (2004)]. This distribution is consistent with the GO categories of the target genes of the training CRMs.

The G+C content of the predicted CRM windows is shown in Figure VI-13. The known and the predicted CRMs have similar GC content, which is higher than those in intron and intergenic sequences but lower than that in exons. The same trend has been previously reported in [Li et al. (2007)].

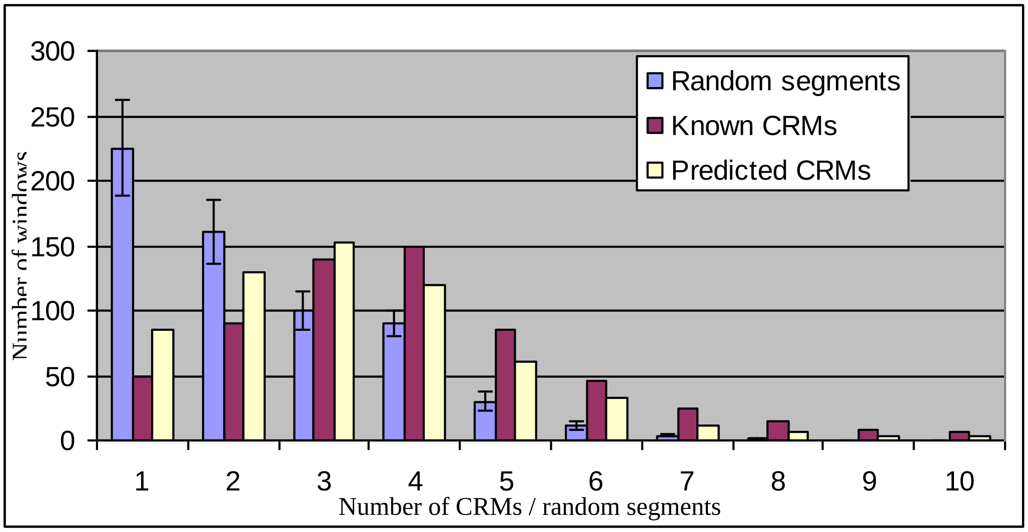


Figure VI-12. The 619 known REDfly CRMs, the 813 CRM windows predicted by Modulexplorer and a set of 813 randomly distributed segments were analyzed for their clustering around genes. A 50 kb long sliding window was scanned over the genome. The number of windows which contained one or more CRMs or random segments is shown below. The histogram shows the number of CRMs or random segments in the window on *x*-axis and the number of such windows on *y*-axis. The known and predicted CRMs come across in clusters of 3 to 4 CRMs in a window, whereas the randomly distributed segments are not usually clustered.

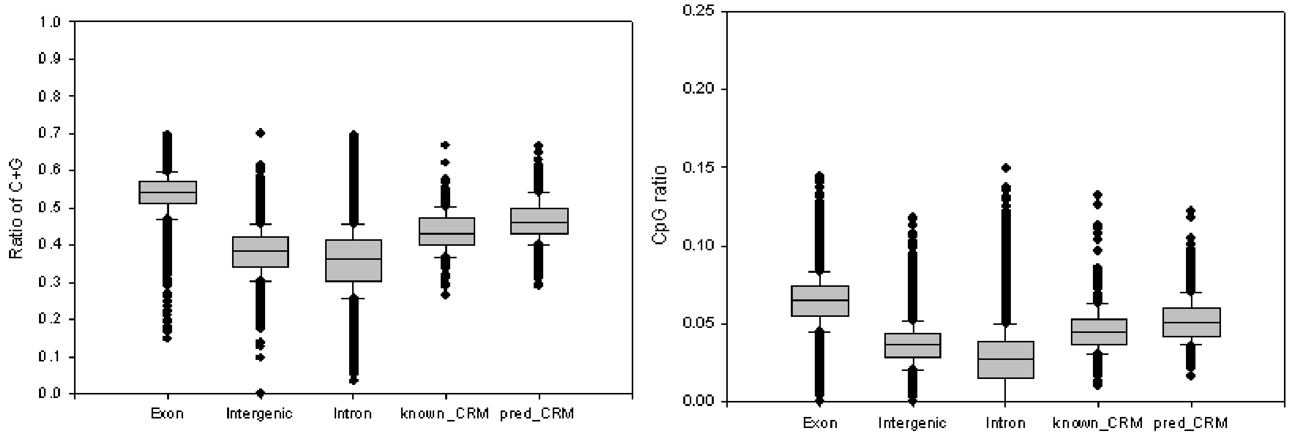


Figure VI-13. The GC content of the predicted CRMs is similar to that of the known CRMs and higher in general compared to intron and intergenic sequences.

VI-7 Feature Based Clustering of CRMs

To characterize the CRMs predicted by Modulexplorer into functional categories, the 813 predicted CRMs and 356 training CRMs were together clustered based on their motif content. The clustering was performed by an iterative frequent itemset mining clustering procedure as described in Section VI-3.3. It was observed that the CRMs of every CRM cluster consistently regulate target genes expressed in the same tissue and development stage. This supports the hypothesis that CRMs with similar motifs regulate target genes within the same tissue and developmental stage.

The major CRM clusters are described below. For each CRM cluster discovered, first we validated from the REDfly database if the known CRMs in the cluster are functional within the same tissue and developmental stage. This check also deduced the type of the CRM cluster. Then for the novel CRMs, we validated if their target genes are expressed in the same tissue and developmental stage using in-situ gene expression profiles from BDGP [Tomancak et al. (2007)] or Flybase [Wilson et al. (2008)] annotation. Finally, the common motifs for the cluster were used to derive a concise regulatory code (see Methods section). We show that the regulatory code specifically distinguishes CRMs that confer the common gene expression pattern from other CRMs and background sequences.

Table VI-2 summarizes the clusters. The first three iterations of the clustering procedure produced a mixed set of CRMs rich in AT motifs. These CRMs represented two major categories with target gene expression in the blastoderm embryo and in the wing imaginal disc of 3rd instar larva. Subsequent iterations produced clusters with predominant target gene expression in the embryonic mesoderm, ventral nerve cord and eye-antennal tissues. The clusters are individually described below.

VI-7.1 Early mesoderm development

The mesoderm cluster consisted of 11 training CRMs for 9 genes and 34 novel CRMs for 27 genes as shown in Figure VI-14. All 11 training CRMs express their target genes in the developing mesoderm during stages 8-12 (nine in the visceral mesoderm and two in somatic mesoderm). Recovering all 11 mesoderm CRMs from the 356 training CRMs by random is highly unlikely (Bonferroni corrected p-value=1.1×10^{-8}).

For the novel CRMs, in-situ expression profiles of 19 out of 27 target genes were available in the BDGP database as shown in Figure VI-15, including *CG2493, sob, traf1, ush, eya, pvf2, wg, fus, rib, egfr, cpr49ac, sens, SP1173, emc, pxb, fer2lch, rst, dm* and *gnf1*. All of these showed expression in the mesoderm during stages 8-12. Of these, the genes *traf1, ush, eya, pvf2, wg, rib, egfr, emc, fer2lch,* and *rst* have known involvement in mesoderm development (confirmed with the Interactive Fly website) while the genes *CG2493, sob, fus, cpr49ac, sens, SP1173, pxb, dm* and *gnf1* are novel. Of the remaining 8 target genes, four genes *sna, knrl, htl* and *fer1hch* were confirmed by Flybase annotations for their involvement in mesoderm development, while the other four are unknown to function in the mesoderm. The recovery of at least 23 out of 27 genes as functional in the mesoderm is again highly unlikely by chance (Bonferroni corrected p-value$<7\times10^{-9}$).

The regulatory code derived for the CRMs in this cluster contained 12 motifs (Figure VI-14). The occurrence of all these motifs within a 1 kb fragment at a PWM match threshold of 5.0×10^{-4} (see Section VI-3.4) was sufficient to classify a sequence as a mesoderm enhancer. The code is specific, reporting zero false positive against other REDfly CRMs and two false positives against 1000 random sequences. The dpp 813 bp enhancer was the lone available CRM in this cluster that has experimental TFBS

Table VI-2. Clusters of CRMs sharing a common regulatory code (motifs) obtained using iterative frequent itemset mining. Five major clusters are listed with their (i) predominant tissue and stage of expression, (ii) number of known and predicted CRM target genes, (iii) number of predicted CRM target genes with validation, (iv) number of validated genes which are novel for their role in development, and (v) false positive rate of the regulatory code on other training CRMs and random background sequences.

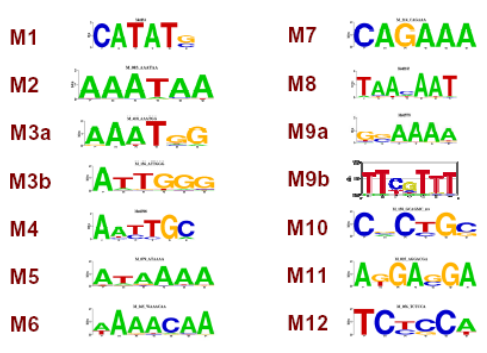
Cluster name (by dominant gene expression pattern)	Known target genes	Predicted target genes	Predicted target genes with validation	Validation type	Novel target genes (out of the predicted genes with validation)*	Specificity of regulatory code	
						# false positives in 1000 random sequences	# matches with other CRMs (out of 356)
Mesoderm (stage 8-12)	9	27	23	BDGP in-situ (19) +Flybase (4)	9	0	2
Ventral nerve cord (stage 11-16)	14	44	30	BDGP in-situ (23) + Flybase (7)	18	2	0
Eye-antennal disc (stage 12-16)	17	21	9	BDGP in-situ (7) + Flybase (2)	4	0	0
Wing imaginal disc (3rd instar larva)	9	31	15	Microarray (12) + Flybase (3)	-	12	8
Blastoderm (stage 4-6)	29	79	50	BDGP (37) + Microarray (13)	-	-	-

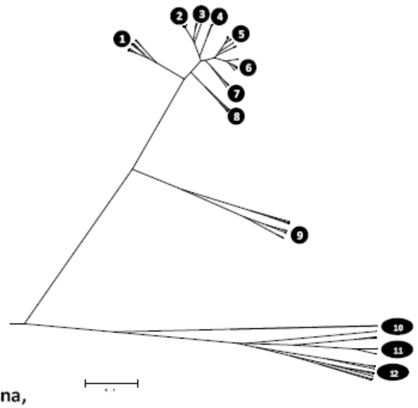
* Novel target genes refers to the predicted target genes which have been validated from BDGP in-situ images but have not previously been cited in the literature for their role in development. The updated list of genes with a known role in development was obtained from the "Interactive Fly" website:

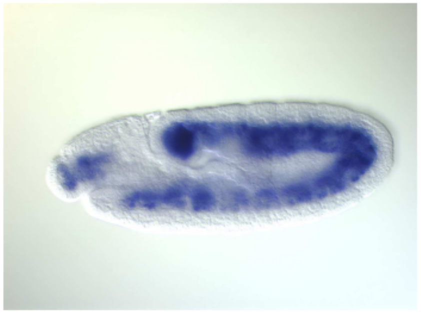
Mesoderm: http://www.sdbonline.org/fly/aimorph/mesoderm.htm
Ventral nerve cord: http://www.sdbonline.org/fly/aimorph/cns.htm
Eye-antennal disc: http://www.sdbonline.org/fly/aimorph/eye.htm

Cluster 3: Mesoderm (Gene expression in the mesoderm primordium in embryonic stages 9-12)

Regulatory code:







M1 = twi, M2 = eve, M3a = prd, M3b = kni, M4 = sna,
M5 = abd-b, M6 = bin/grh/exd, M7 = abd-b,
M9a = dl, M9b = exd/bin, M11 = ems, M12 = tin

ACT57B_-539/+2|chr2R|16830939|16831534	ACT57B
BAP_BAP3|chr3R|17216733|17217194	BAP
DAP_DAP-SB|chr2R|5600577|5602440	DAP
DPP_DPP813|chr2L|2445769|2446581	DPP
MEF2_II-E|chr2R|5825113|5826217	MEF2
RST_F5P|chrX|2872501|2874950	RST
RST_F6P|chrX|2874944|2877475	RST
TIN_TIN56|chr3R|17207306|17208183	TIN
TM1_PME|chr3R|11136386|11136738	TM1
EVE_EME-B|chr2R|5872866|5873261	EVE
chr2L|533501|534500	USH
chr2L|3583501|3584500	SOB;ODD
chr2L|4366001|4367000	TRAF1
chr2L|6531001|6532000	EYA
chr2L|7082001|7083000	PVF2
chr2L|7323501|7324500	WG;WNT6
chr2L|15461001|15462000	CG4161;SNA
chr2L|15483501|15484500	SNA;TIM17B2
chr2L|20483501|20484500	CG2493;MIR-1
chr2R|5589001|5590000	UBA1;CG30002
chr2R|7445001|7446000	EN;TOU
chr2R|8275501|8276500+	CPR49AC
chr2R|8276001|8277000	CPR49AC
chr2R|11552001|11553000	FUS
chr2R|15152501|15153500	CG7229;RIB
chr2R|17432501|17433500	EGFR
chr3L|746001|747000	CG13897;EMC
chr3L|6697001|6698000	SP1173;CG8519
chr3L|6697501|6698500	SP1173;CG8519
chr3L|13384501|13385500	CG10191;SENS
chr3L|19675501|19676500	TEY;CG8765
chr3L|20622501|20623500	KNRL;CG13251
chr3L|20628001|20629000	KNRL;CG13251
chr3R|705501|706500	LAF;GNF1
chr3R|11469001|11470000	CG5302;PXB
chr3R|11473001|11474000	CG5302;PXB
chr3R|13890501|13891500	HTL;CG14317
chr3R|26189501|26190500	CG34300;FER1HCH
chr3R|26240001|26241000	FER2LCH;CG2217
chrX|2915001|2916000	RST;CG4116
chrX|2915501|2916500	RST;CG4116
chrX|3238501|3239500	CG10793;DM
chrX|20584501|20585500	RUN;CG1324
chrX|20597001|20598000	RUN;CG1324

Figure VI-14. Cluster of CRMs controlling target gene expression in the embryonic mesoderm, and their regulatory code.

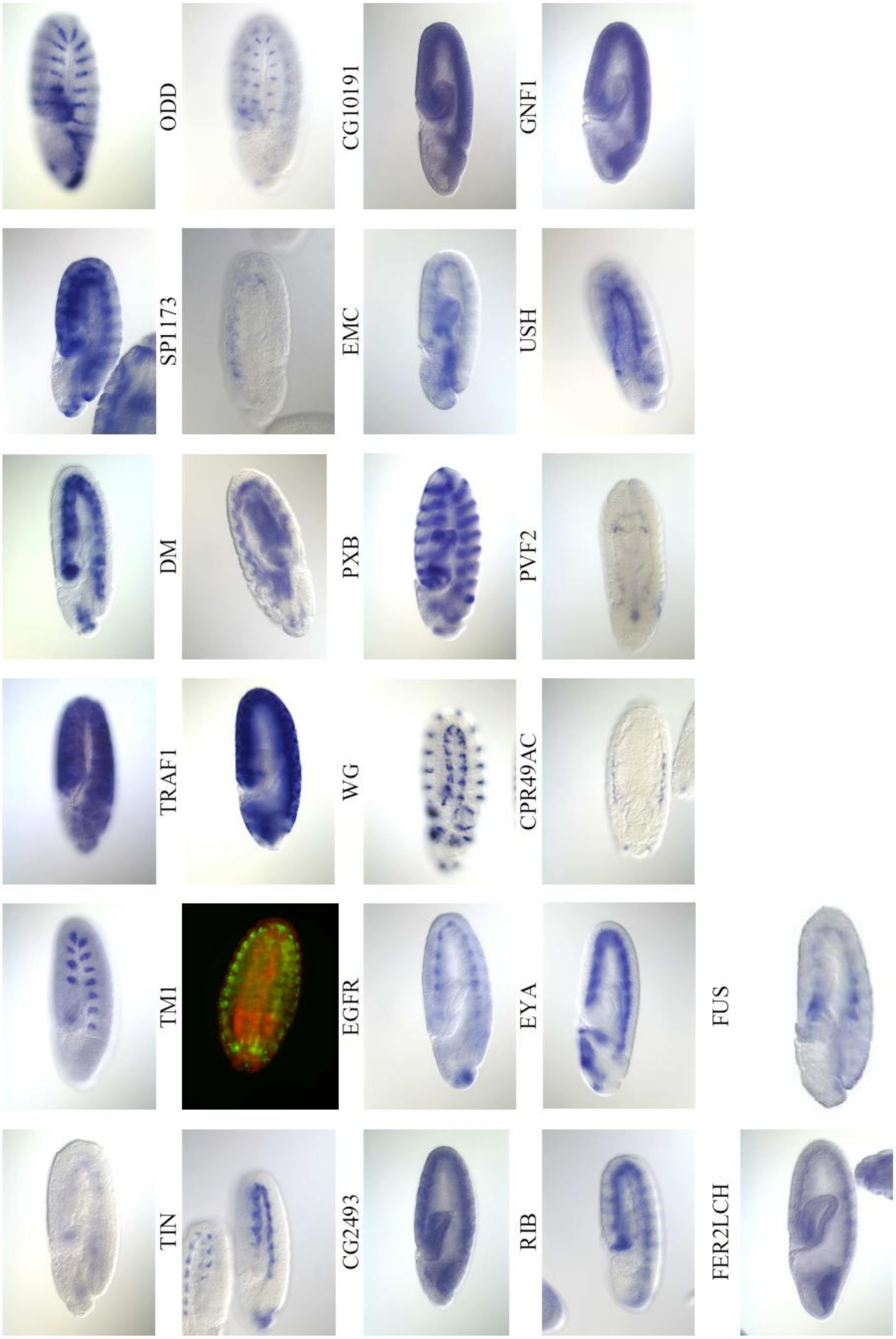
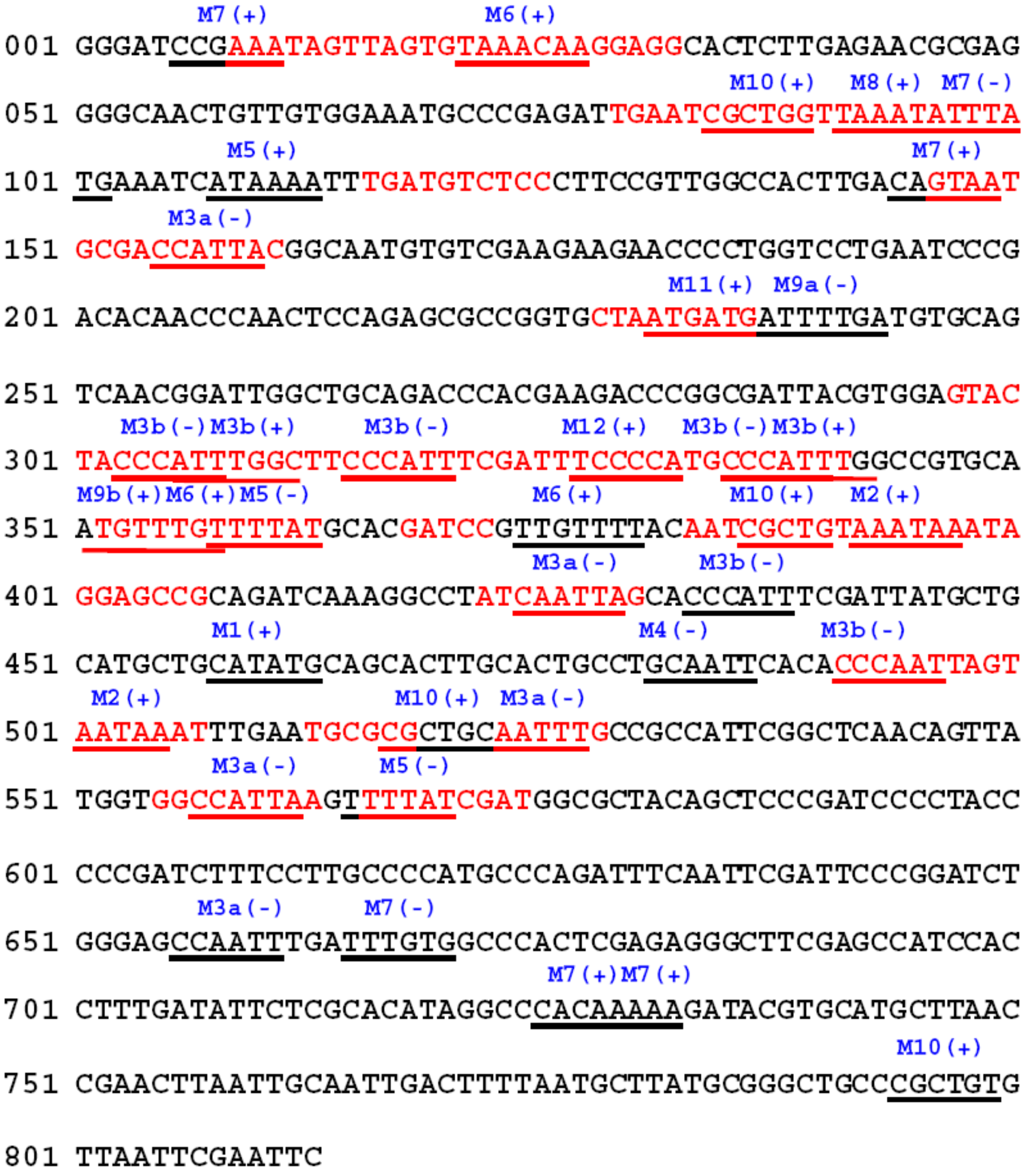


Figure VI-15. BDGP in-situ expression images for the target genes of novel CRMs in the mesoderm cluster.



```
          M7(+)           M6(+)
001 GGGATCCGAAATAGTTAGTGTAAACAAGGAGGCACTCTTGAGAACGCGAG
                                       M10(+) M8(+) M7(-)
051 GGGCAACTGTTGTGGAAATGCCCGAGATTGAATCGCTGGTTAAATATTTA
            M5(+)                               M7(+)
101 TGAAATCATAAAATTTGATGTCTCCCTTCCGTTGGCCACTTGACAGTAAT
         M3a(-)
151 GCGACCATTACGGCAATGTGTCGAAGAAGAACCCCTGGTCCTGAATCCCG
                                   M11(+) M9a(-)
201 ACACAACCCAACTCCAGAGCGCCGGTGCTAATGATGATTTTGATGTGCAG

251 TCAACGGATTGGCTGCAGACCCACGAAGACCCGGCGATTACGTGGAGTAC
      M3b(-)M3b(+) M3b(-)     M12(+) M3b(-)M3b(+)
301 TACCCATTTGGCTTCCCATTTCGATTTCCCCATGCCCATTTGGCCGTGCA
    M9b(+)M6(+)M5(-)        M6(+)      M10(+) M2(+)
351 ATGTTTGTTTTATGCACGATCCGTTGTTTTACAATCGCTGTAAATAAATA
                            M3a(-)   M3b(-)
401 GGAGCCGCAGATCAAAGGCCTATCAATTAGCACCCATTTCGATTATGCTG
           M1(+)                  M4(-)    M3b(-)
451 CATGCTGCATATGCAGCACTTGCACTGCCTGCAATTCACACCCAATTAGT
     M2(+)           M10(+) M3a(-)
501 AATAAATTTGAATGCGCGCTGCAATTTGCCGCCATTCGGCTCAACAGTTA
           M3a(-)   M5(-)
551 TGGTGGCCATTAAGTTTTATCGATGGCGCTACAGCTCCCGATCCCCTACC

601 CCCGATCTTTCCTTGCCCCATGCCCAGATTTCAATTCGATTCCCGGATCT
          M3a(-)   M7(-)
651 GGGAGCCAATTTGATTTGTGGCCCACTCGAGAGGGCTTCGAGCCATCCAC
                             M7(+)M7(+)
701 CTTTGATATTCTCGCACATAGGCCCACAAAAAGATACGTGCATGCTTAAC
                                                M10(+)
751 CGAACTTAATTGCAATTGACTTTTAATGCTTATGCGGGCTGCCCGCTGTG

801 TTAATTCGAATTC
```

Figure VI-16. Matches of the mesoderm regulatory code motifs within the dpp 813 bp enhancer are shown by underlines. For comparison the known TFBS in this enhancer, available only for the first 600 bp, are shown in red color text. Out of 32 matches of the regulatory code motifs in first 600 bp, 26 overlapped known TFBS.

annotation. The sites of regulatory code motifs in this CRM matched closely with the known TFBS annotation as shown in Figure VI-16.

The motifs in the regulatory code showed similarity to the known motifs for the TFs *dl, twi, sna, tin, bin, abd-B, exd, eve, ftz, prd* and *ems*. The TFs *dl, twi* and *sna* are known to establish the identity of mesoderm cells [Ganguly et al. (2005)]. The TFs *eve, tin* and *bin* promote the differentiation of mesoderm cells post-gastrulation (stages 8-12) to form different muscle progenitor types (somatic, visceral, tracheal, etc.). These TFs act together with pair-rule and segment polarity TFs including *dpp, wg, en, exd, eve, prd, ftz, abd-B, ems* which form gradients along anterior-posterior (AP) and dorso-ventral (DV) axes. The AP and DV gradients allow formation of different muscle progenitors in different parasegments along these axes [Borkowski et al. (1995)]. Though it has been suggested that the independent influences of the above TFs could be integrated together via CRMs [Furlong (2004)], so far no specific regulatory code is known for these CRMs. The current regulatory code is thus novel to characterize CRMs in mesoderm development during stages 8-12 when muscle progenitors differentiate.

VI-7.2 Ventral nerve cord

The ventral nerve cord (VNC) cluster consisted of 15 known CRMs for 14 genes and 44 novel CRMs for 44 genes as shown in Figure VI-17. 11 out of 15 known CRMs have known involvement in ventral nerve cord development during stages 11-16 (p-value 6×10^{-6}).

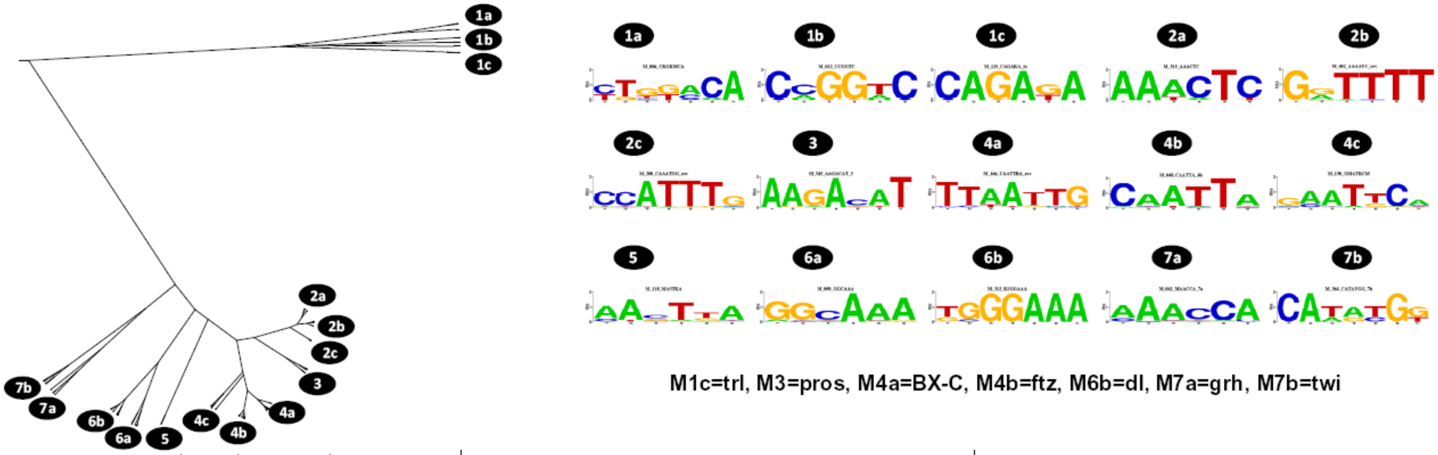
Among the 44 targets genes of the 44 novel CRMs, 23 genes had in-situ confirmation of expression in the VNC during stages 11-16 as shown in Figure VI-18, while another 7 genes were annotated in Flybase as functional in the VNC. Thus 30 out of 44 genes were validated in VNC development. Out of these, 12 genes *pdm2, tsh, traf1, wg, phyl, klu, mirr, D, Ap-2, B-h1, run,* and *rst* have known function in VNC development while 18 genes *ceng1a, slp2, ush, rx, pk, sens, comm2, CG6897, fz2, CG11347, klar, ets65a, pxb, ptx1, hth, corto, Ca-alpha1T,* and *CG9650* are novel. For the rest 14 genes, 10 have no information available while 4 show no expression in the VNC.

The regulatory code for the VNC cluster consists of 15 motifs. The regulatory code could separate the known neuronal enhancers in the VNC cluster from other known REDfly CRMs and 1000 random sequences with 100% specificity.

The motifs in the regulatory code closely matched the known consensus for the TFs *dl, twi, grh, trl, ftz, pros* and the bithorax complex TFs which have known involvement in VNC regulation. The TF *twi* specifies the neuroectoderm cells. About a quarter of the neuroectodermal cells eventually differentiate as neuroblasts while the rest form the ectoderm. During gastrulation (stage 7), the neuroectodermal cells migrate to the ventral region. The fate of neuroectodermal cells as neuroblasts or epidermal cells is decided during stages 8-12 by lateral inhibition. Neuroblast formation is promoted by the proneural genes (*ac, sc, lsc, ase*) and inhibited by neurogenic genes (*notch, delta, su(H)* etc.). The neuroblasts in different parasegments along the AP and DV axes develop subtypes by expressing different sets of genes under the control of various TFs. The regulatory inputs of these TFs are combined by CRMs. The TFs *pros, grh, ftz,* bithorax complex TFs are known to function together in neuroblast differentiation, such as the formation of ganglion mother cells [Prokop et al. (1998); Skeath and Thor (2003)]. The different neurblasts proliferate in the interior of the embryo during stages 13-16 to form

Cluster 4: Ventral Nerve Cord

Gene expression in the ventral nerve cord primordium in embryonic stages 11-12.



M1c=trl, M3=pros, M4a=BX-C, M4b=ftz, M6b=dl, M7a=grh, M7b=twi

BETATUB56D_AS1|chr2R|15337665|15338984
BRK_neurogenic|chrX|7190966|7191464
CHN_SOP|chr2R|11019805|11020918
POXN_4|chr2R|11724005|11726807
RHO_NEE-600|chr3L|1461606|1462196
RUN_STRIPE_5|chrX|20552648|20553988
RUN_STRIPE3+7|chrX|20554214|20556618
SALM_BAT|chr2L|11410694|11413358
SIM_MESECTODERM|chr3R|8895835|8896466
SLI_1.8_NV|chr2R|11776614|11778398
SO_SO7|chr2R|3318590|3320202
SOG_broad_neurogenic |chrX|15518730|15519122
TWI_DL_MEL|chr2R|18932427|18933842
VN_neurogenic|chr3L|5828770|5829267
VND_early_embryonic|chrX|485982|487688

chr2L:12666000-12667000	PDM2
chr2L:13227000-13228000	CG16813
chr2L:13889000-13890000	CENG1A
chr2L:21889500-21890500	TSH;CG11629
chr2L:3860000-3861000	SLP2;CG3964
chr2L:4366000-4367000	TRAF1
chr2L:532500-533500	USH
chr2L:7300000-7301000	CG31909;WG
chr2R:10318500-10319500	PHYL
chr2R:16828000-16829000	RX;ACT57B
chr2R:19078000-19079000	CG34371
chr2R:19094000-19095000	CG13539;CG3162
chr2R:3050000-3051000	PK
chr3L:10994000-10995000	KLU
chr3L:12644500-12645500	CG32111;MIRR
chr3L:13384500-13385500	CG10191;SENS
chr3L:14149500-14150500	SOX21B;D
chr3L:15692500-15693500	COMM2
chr3L:18616000-18617000	CG6897
chr3L:18842500-18843500	CG32027;MIR-315
chr3L:19224000-19225000	FZ2;CG33647
chr3L:21591000-21592000	AP-2
chr3L:4409000-4410000	CG11347
chr3L:450000-451000	KLAR
chr3L:6109000-6110000	ETS65A
chr3L:9603000-9604000	CG3280
chr3R:11468500-11469500	CG5302;PXB
chr3R:26746000-26747000	PTX1
chr3R:27318000-27319000	CG11333;CG12063
chr3R:3961000-3962000	CG7891;GRN
chr3R:6321500-6322500	CYP12E1;HTH
chr3R:674500-675500	CG14659;OPA
chr3R:9293500-9294500	CG17025;CG12538
chr3R:935500-936500	CORTO;CG12007
chrX:10578000-10579000	X11LBETA
chrX:1276500-1277500	CG32813
chrX:15540500-15541500	CG8117;CG8119
chrX:17302000-17303000	B-H1;CG8611
chrX:17650500-17651500	CG15816;UNC-4
chrX:20595500-20596500	RUN;CG1324
chrX:2915000-2916000	RST;CG4116
chrX:6050500-6051500	CA-ALPHA1T;CG32750
chrX:7118000-7119000	CG9650
chrX:7168000-7169000	CG1958;CG1677

Figure VI-17. Cluster of CRMs controlling target gene expression in the embryonic ventral nerve cord, and their regulatory code.

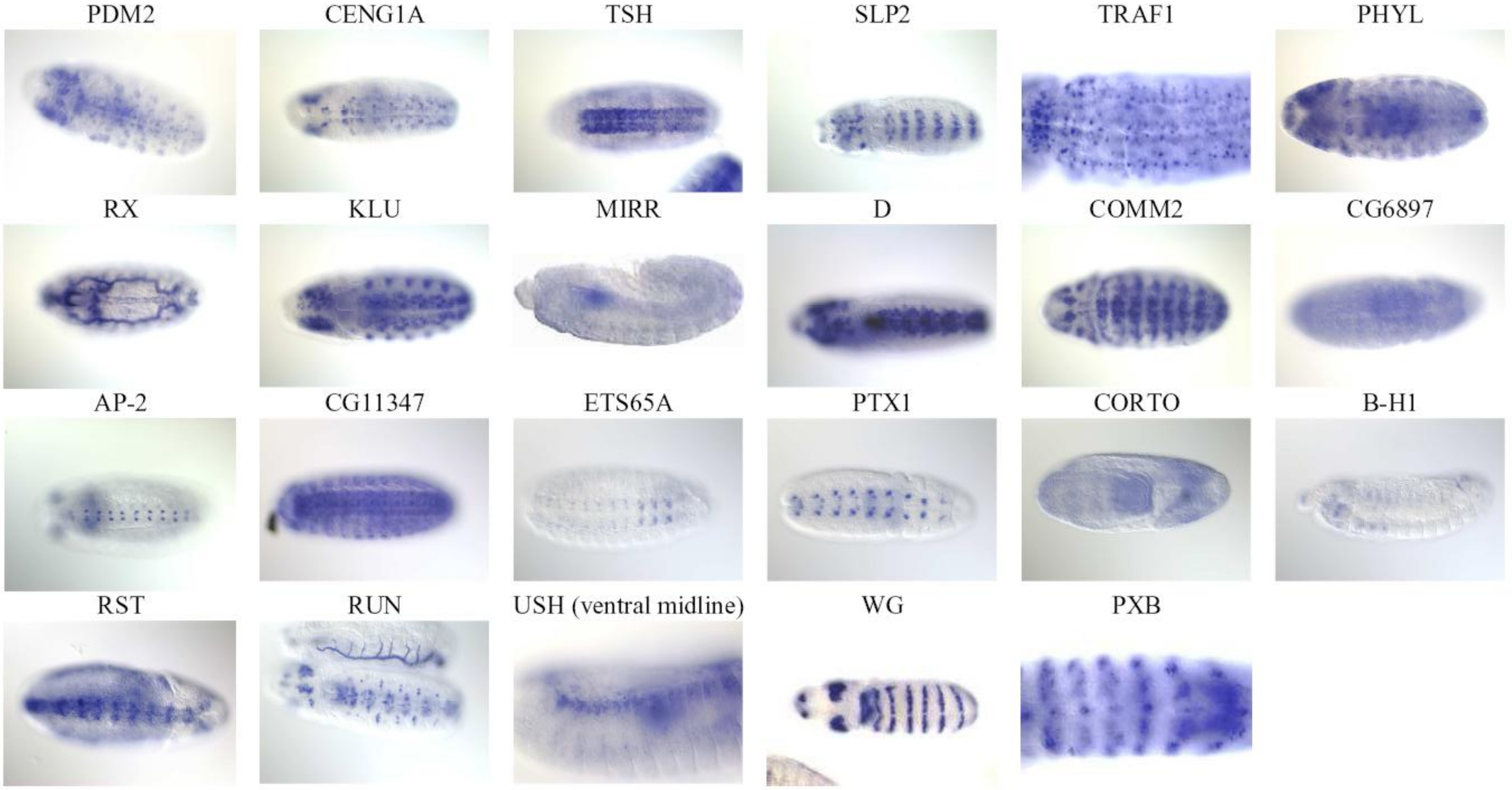


Figure VI-18. BDGP in-situ expression images for the target genes of novel CRMs in the ventral nerve cord cluster.

the VNC. Thus from the identities of motifs in the current regulatory code, it again appears that the CRMs in this cluster regulate neuroblast differentiation.

VI-7.3 Eye-antennal disc

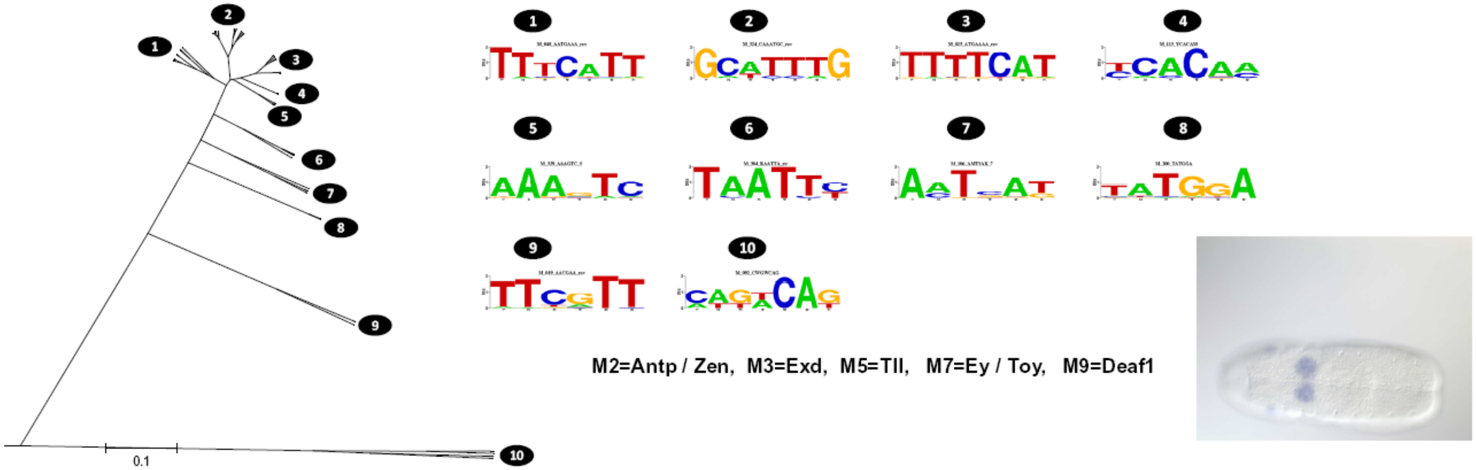
The eye-antennal expression cluster consisted of 18 known CRMs for 17 genes and 21 novel CRMs for 21 genes as shown in Figure VI-19. 12 of the 18 known CRMs confer expression in the eye-antennal disc during stages 12-16 (p-value 8.3×10^{-6}).

Among the novel CRMs, the 9 target genes *ceng1a, lola, D, fz, spn, cas, fer2lch, opa, skpd* were confirmed as expressed in the eye-antennal disc (Figure VI-20). This has a p-value of 1.0×10^{-6}. For 9 other target genes, no expression or functional annotation information was available. Three genes showed no expression in the embryonic eye-antennal disc. Out of the 9 validated genes, *lola, D, fz, spn* and *cas* have known involvement in eye-antennal development while the genes *ceng1a, fer2lch, opa* and *skpd* represent novel targets.

The eye-antennal regulatory code had 10 motifs. The regulatory code gave 100% specificity over other known CRMs and 1000 random sequences. Motifs in the regulatory code were recognized as closely resembling the known binding sites for the TFs *Antp/zen, Exd, tll* and *ey/toy*. The TFs ey and toy specify the optic primordium cells. The eye-antennal imaginal disc is formed during stage 12 by the invagination of optic primordium cells to produce a monolayer epithelium. Commitment of the imaginal disc cells towards eye or antenna fates occurs in a series of steps from stage 12 embryo until the second instar larva. Several eye or antennal determinant genes such as *eyg, ey, toy, dac, optix, salm, exd, dll,* etc. are expressed from embryonic stage 12 onwards in the imaginal disc. The TFs *dpp, zen, tll, otd, wg* etc. are active in this process. From the

Cluster 5: Eye-Antennal Disc

Gene expression in the eye-antennal disc during stages 11-16 of the developing embryo.



ABD-B_IAB7|chr3R|12741361|12742091	chr2L:13885500-13886500	CENG1A
AOP_C-LACZ|chr2L|2174335|2176640	chr2L:19623000-19624000	LAR
BRD_1.5|chr3L|14964318|14965805	chr2R:18119000-18120000	OATP58DC;DVE
DAC_5EE|chr2L|16469684|16471421	chr2R:19078000-19079000	CG34371
DFD_EAE|chr3R|2611056|2613714	chr2R:20936500-20937500	GSB-N
EY_UE2.3|chr4|722314|724597	chr2R:6424500-6425500	LOLA
GCM_-7.4/-4.4|chr2L|9586087|9589095	chr3L:11864500-11865500	CG5906;CG14128
GTP-BP_srpralpha|chrX|11022609|11024686	chr3L:14141500-14143000	SOX21B;D
KNI_+1_construct|chr3L|20687054|20688533	chr3L:14283500-14284500	FZ
KR_PP3.OHZ|chr2R|21104944|21108078	chr3L:18843000-18844000	CG32027;MIR-315
KR_SN1.7KRZ|chr2R|21113280|21115023	chr3L:2553000-2554000	SPN
OBP18A_PROM|chrX|19029874|19032842	chr3R:1548500-1549500	CAS;CG1239
OBP56A_PROM|chr2R|15582271|15585291	chr3R:17379500-17380500	SLOU
OBP56B_PROM|chr2R|15586805|15588818	chr3R:18053500-18054500	CG31163
OBP56E_PROM|chr2R|15596902|15599903	chr3R:26240000-26241000	FER2LCH;CG2217
OBP56G_PROM|chr2R|15671523|15674474	chr3R:27018000-27019000	CG31004;BNK
OBP57B_PROM|chr2R|16388770|16391772	chr3R:4618000-4619000	PIF1B;PIF1A;CG11776
OTD_eye_enhancer|chrX|8530863|8533429	chr3R:674500-675500	CG14659;OPA
	chr3R:7137000-7138000	CG31386;KP78B
	chrX:13250500-13251500	CG15747;CG10617
	chrX:19702500-19703500	CG12701;SKPD

Figure VI-19. Cluster of CRMs controlling target gene expression in the embryonic eye-antennal disc, and their regulatory code.

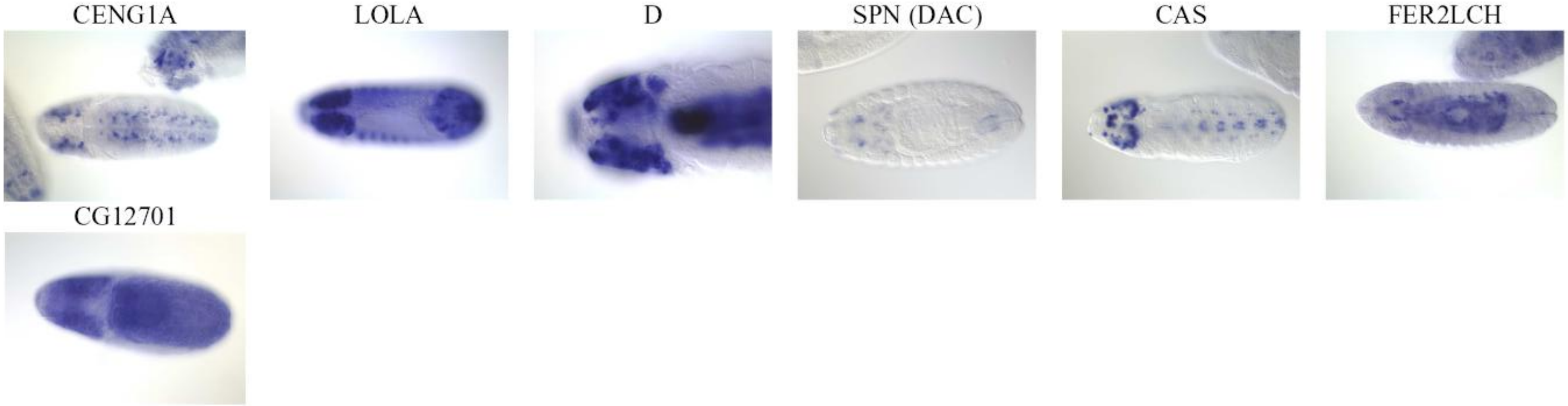


Figure VI-20. BDGP in-situ expression images for the target genes of novel CRMs in the eye-antennal disc cluster.

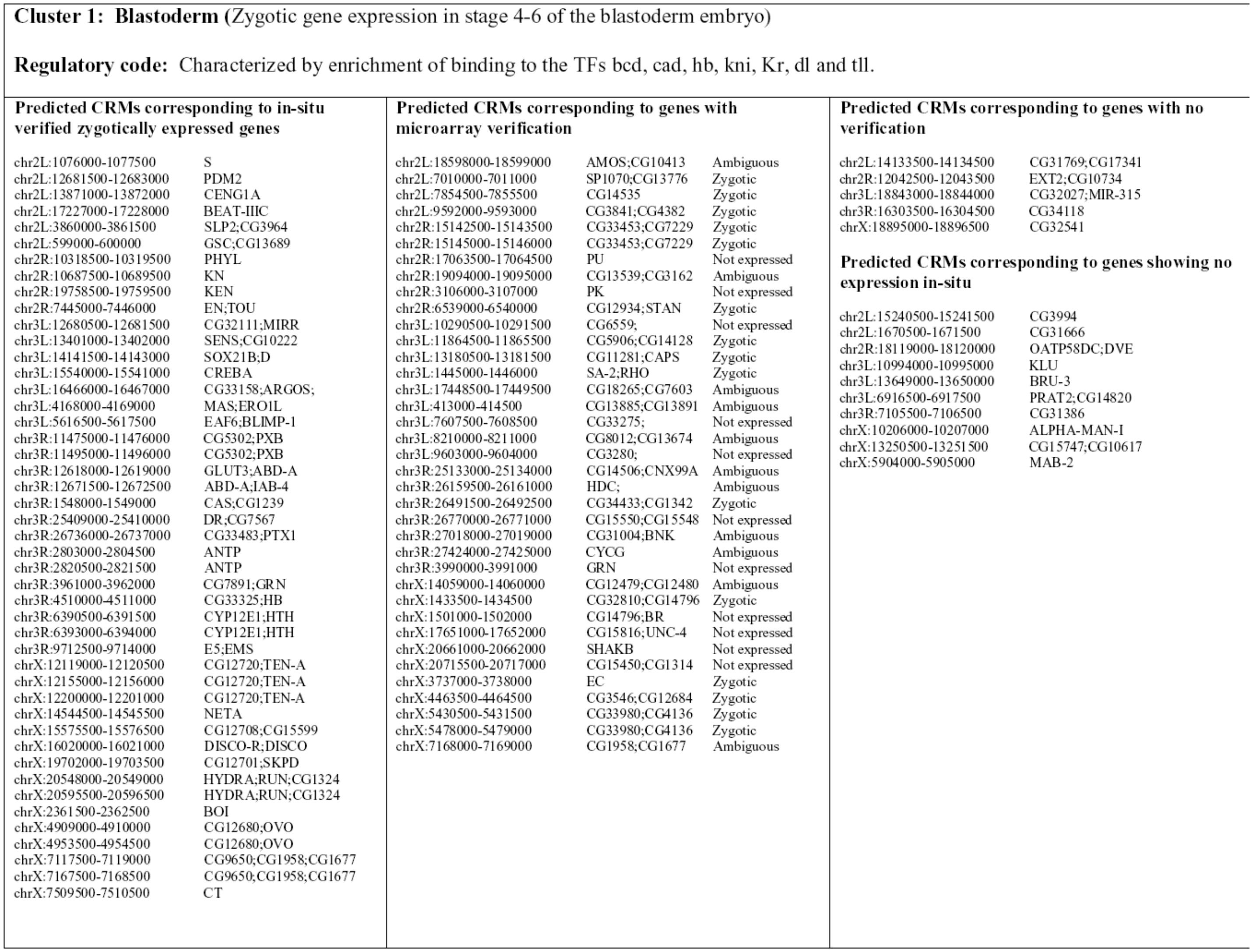
regulatory code, it therefore appears that the CRMs in this cluster regulate genes in the embryonic stage of eye-antennal specification.

The appearance of *deaf1* motif in the regulatory code is surprising. *Deaf1* was also observed in the previous section to interact with *ey* and *toy* in the TF-TF interaction matrix. Its role in eye development is therefore a subject for further study.

VI-7.4 Blastoderm embryo

The CRMs containing AT-rich motifs obtained in the first three iterations of the clustering procedure consisted of two major CRM types controlling target gene expression in the blastoderm embryo and the wing imaginal disc. The blastoderm CRMs were separated manually on the basis of their enrichment in binding sites for the known blastoderm TFs *hb, bcd, cad, Kr, kni, dl* and *tll*. The binding sites were annotated using the PWMs for these TFs reported in previous studies [Berman et al. (2002); Rajewsky et al. (2002)]. A total of 33 known CRMs for 29 genes and 98 novel CRMs for 79 genes were recovered as shown in Figure VI-21.

The novel blastoderm CRMs showed a 2-fold enrichment of TF-binding as compared to their flanking -5 kb to +5 kb regions as shown in Figure VI-23. The target genes of these CRMs were studied for zygotic expression in stage 4-6 developing embryos. Out of 79 genes, zygotic expression could be confirmed for at least 50 genes. 37 of these were validated from in-situ images in BDGP in-situ [Tomancak et al. (2007)] and Fly-FISH [Lecuyer et al. (2007)] databases as shown in Figure VI-22. 13 other genes were confirmed from microarray expression data [Arbeitman et al. (2002); Pilot et al. (2006)]. Since microarray data does not clearly identify tissue localized zygotic gene



Cluster 1: Blastoderm (Zygotic gene expression in stage 4-6 of the blastoderm embryo)

Regulatory code: Characterized by enrichment of binding to the TFs bcd, cad, hb, kni, Kr, dl and tll.

Predicted CRMs corresponding to in-situ verified zygotically expressed genes

chr2L:1076000-1077500	S
chr2L:12681500-12683000	PDM2
chr2L:13871000-13872000	CENG1A
chr2L:17227000-17228000	BEAT-IIIC
chr2L:3860000-3861500	SLP2;CG3964
chr2L:599000-600000	GSC;CG13689
chr2R:10318500-10319500	PHYL
chr2R:10687500-10689500	KN
chr2R:19758500-19759500	KEN
chr2R:7445000-7446000	EN;TOU
chr3L:12680500-12681500	CG32111;MIRR
chr3L:13401000-13402000	SENS;CG10222
chr3L:14141500-14143000	SOX21B;D
chr3L:15540000-15541000	CREBA
chr3L:16466000-16467000	CG33158;ARGOS;
chr3L:4168000-4169000	MAS;ERO1L
chr3L:5616500-5617500	EAF6;BLIMP-1
chr3R:11475000-11476000	CG5302;PXB
chr3R:11495000-11496000	CG5302;PXB
chr3R:12618000-12619000	GLUT3;ABD-A
chr3R:12671500-12672500	ABD-A;IAB-4
chr3R:1548000-1549000	CAS;CG1239
chr3R:25409000-25410000	DR;CG7567
chr3R:26736000-26737000	CG33483;PTX1
chr3R:2803000-2804500	ANTP
chr3R:2820500-2821500	ANTP
chr3R:3961000-3962000	CG7891;GRN
chr3R:4510000-4511000	CG33325;HB
chr3R:6390500-6391500	CYP12E1;HTH
chr3R:6393000-6394000	CYP12E1;HTH
chr3R:9712500-9714000	E5;EMS
chrX:12119000-12120500	CG12720;TEN-A
chrX:12155000-12156000	CG12720;TEN-A
chrX:12200000-12201000	CG12720;TEN-A
chrX:14544500-14545500	NETA
chrX:15575500-15576500	CG12708;CG15599
chrX:16020000-16021000	DISCO-R;DISCO
chrX:19702000-19703500	CG12701;SKPD
chrX:20548000-20549000	HYDRA;RUN;CG1324
chrX:20595500-20596500	HYDRA;RUN;CG1324
chrX:2361500-2362500	BOI
chrX:4909000-4910000	CG12680;OVO
chrX:4953500-4954500	CG12680;OVO
chrX:7117500-7119000	CG9650;CG1958;CG1677
chrX:7167500-7168500	CG9650;CG1958;CG1677
chrX:7509500-7510500	CT

Predicted CRMs corresponding to genes with microarray verification

chr2L:18598000-18599000	AMOS;CG10413	Ambiguous
chr2L:7010000-7011000	SP1070;CG13776	Zygotic
chr2L:7854500-7855500	CG14535	Zygotic
chr2L:9592000-9593000	CG3841;CG4382	Zygotic
chr2R:15142500-15143500	CG33453;CG7229	Zygotic
chr2R:15145000-15146000	CG33453;CG7229	Zygotic
chr2R:17063500-17064500	PU	Not expressed
chr2R:19094000-19095000	CG13539;CG3162	Ambiguous
chr2R:3106000-3107000	PK	Not expressed
chr2R:6539000-6540000	CG12934;STAN	Zygotic
chr3L:10290500-10291500	CG6559;	Not expressed
chr3L:11864500-11865500	CG5906;CG14128	Zygotic
chr3L:13180500-13181500	CG11281;CAPS	Zygotic
chr3L:1445000-1446000	SA-2;RHO	Zygotic
chr3L:17448500-17449500	CG18265;CG7603	Ambiguous
chr3L:413000-414500	CG13885;CG13891	Ambiguous
chr3L:7607500-7608500	CG33275;	Not expressed
chr3L:8210000-8211000	CG8012;CG13674	Ambiguous
chr3L:9603000-9604000	CG3280;	Not expressed
chr3R:25133000-25134000	CG14506;CNX99A	Ambiguous
chr3R:26159500-26161000	HDC;	Ambiguous
chr3R:26491500-26492500	CG34433;CG1342	Zygotic
chr3R:26770000-26771000	CG15550;CG15548	Not expressed
chr3R:27018000-27019000	CG31004;BNK	Ambiguous
chr3R:27424000-27425000	CYCG	Ambiguous
chr3R:3990000-3991000	GRN	Not expressed
chrX:14059000-14060000	CG12479;CG12480	Ambiguous
chrX:1433500-1434500	CG32810;CG14796	Zygotic
chrX:1501000-1502000	CG14796;BR	Not expressed
chrX:17651000-17652000	CG15816;UNC-4	Not expressed
chrX:20661000-20662000	SHAKB	Not expressed
chrX:20715500-20717000	CG15450;CG1314	Not expressed
chrX:3737000-3738000	EC	Zygotic
chrX:4463500-4464500	CG3546;CG12684	Zygotic
chrX:5430500-5431500	CG33980;CG4136	Zygotic
chrX:5478000-5479000	CG33980;CG4136	Zygotic
chrX:7168000-7169000	CG1958;CG1677	Ambiguous

Predicted CRMs corresponding to genes with no verification

chr2L:14133500-14134500	CG31769;CG17341
chr2R:12042500-12043500	EXT2;CG10734
chr3L:18843000-18844000	CG32027;MIR-315
chr3R:16303500-16304500	CG34118
chrX:18895000-18896500	CG32541

Predicted CRMs corresponding to genes showing no expression in-situ

chr2L:15240500-15241500	CG3994
chr2L:1670500-1671500	CG31666
chr2R:18119000-18120000	OATP58DC;DVE
chr3L:10994000-10995000	KLU
chr3L:13649000-13650000	BRU-3
chr3L:6916500-6917500	PRAT2;CG14820
chr3R:7105500-7106500	CG31386
chrX:10206000-10207000	ALPHA-MAN-I
chrX:13250500-13251500	CG15747;CG10617
chrX:5904000-5905000	MAB-2

Figure VI-21. List of novel CRMs separated from the AT-rich clusters which control target gene expression in the blastoderm embryo.

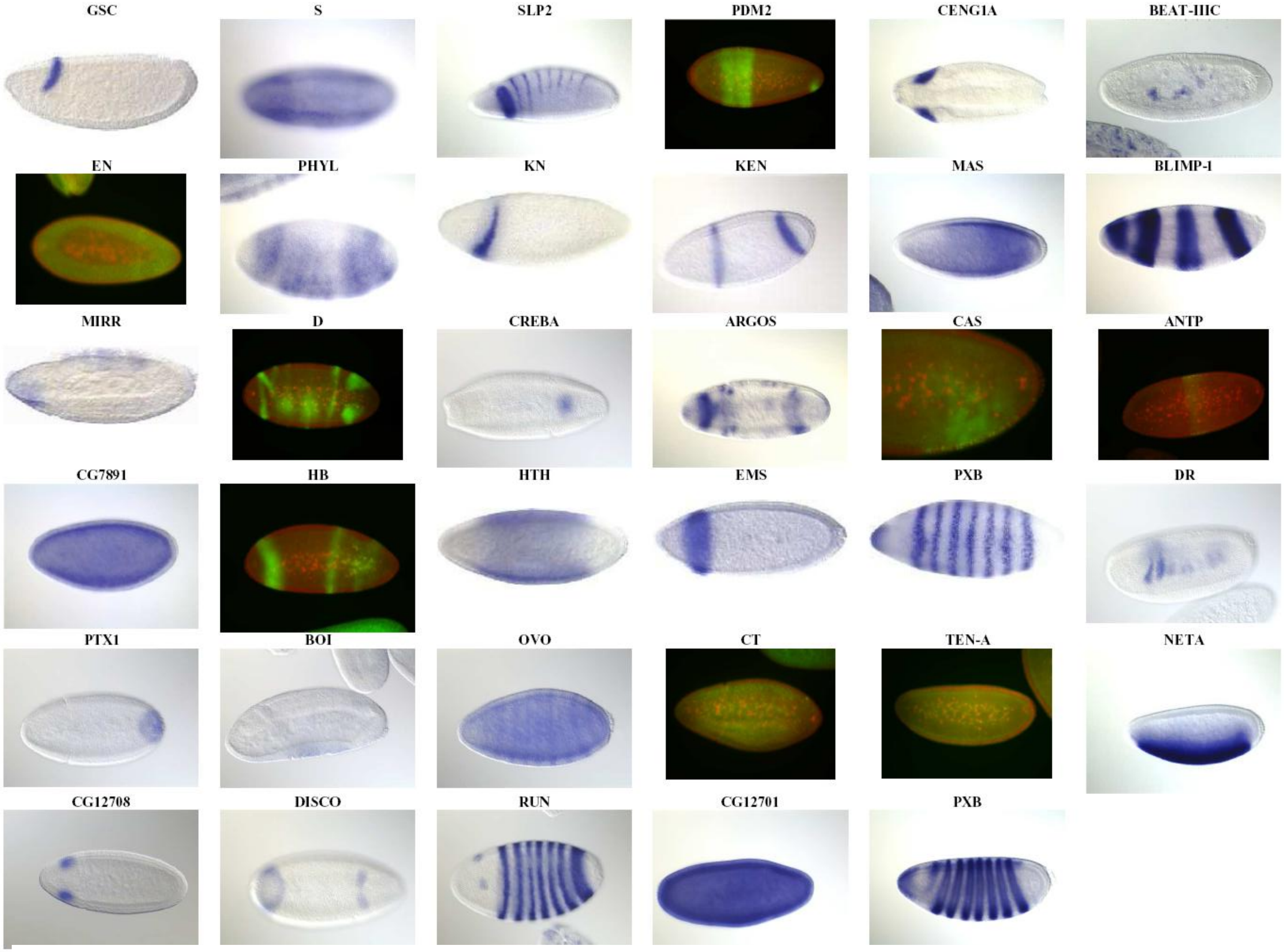


Figure VI-22. BDGP in-situ expression images for the target genes of novel CRMs in the blastoderm cluster.

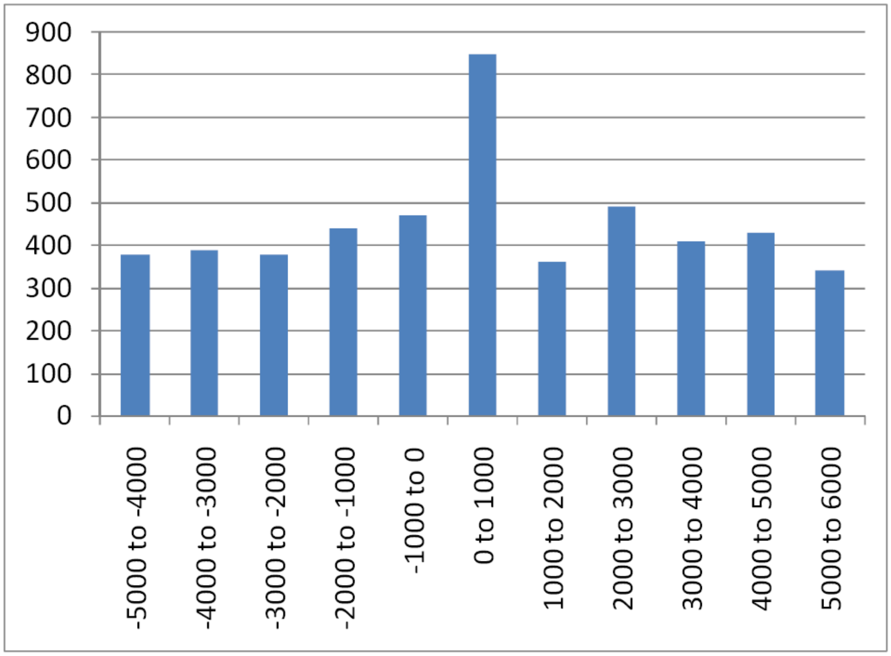


Figure VI-23. Binding sites for 10 blastoderm TFs were searched in the region -5000 to +5000 around the 98 predicted blastoderm CRMs. The CRMs are in the location 0 to 1000. In the CRM region the binding sites were over-represented by a factor of around 2. The y-axis shows the total number of binding sites found in the window in all 98 CRMs.

expression, a general rule was used to separate the genes into three classes – genes down-expressed in stages 1-3 but up-expressed in stages 4-6 were classified as zygotic genes, genes down-expressed throughout stages 1-6 were classified as not expressed and the rest were classified as ambiguous.

Considering 15% of all 14,000 Drosophila genes to be zygotically expressed in the blastoderm embryo, which is a generous estimate [Lecuyer et al. (2007)], the confirmation of at least 50 genes out of 79 for zygotic expression in blasatoderm is statistically significant with a P-value of 1.8×10-18 (hypergeometric probability with Bonferroni correction of factor 14,000).

VI-7.5 Wing imaginal disc

The wing imaginal disc specific CRMs were again manually separated from the AT-rich clusters. Since there is no known regulatory code for wing imaginal disc specification, we derived a regulatory code from known CRMs of the genes *ct, dpp, kn, kni, salm, ser, vg, pfe* and *chn*. All these CRMs confer gene expression in the wing disc in 3rd instar larva. The regulatory code was derived from the common motifs among these CRMs. 33 novel CRMs for 31 genes were separated from the AT-rich clusters using this regulatory code. In these novel CRMs, 15 target genes including *pdm2, drm, cg25c, act57b, dve, inv, rho, emc, c15, hh, CG12063, grn, CG8483, B-h1* and *bi* were validated by their enrichment in the wing imaginal disc in the 3rd instar larva using microarray analysis [Butler et al. (2003)]. For the rest 16 genes, no means of validation was available. All the above validated genes have known function in wing development.

The regulatory code included 11 distinct motifs with 7 motifs resembling TFs *ubx, ap, ara, sd, mad, pan, su(H)* and *nub* which are known to regulate wing imaginal disc development in the larval stage. The wing imaginal disc is formed from the embryonic ectoderm by an invagination at the compartment where DV stripe of *wg* intersects with AP stripe of *dpp*. The primordium of the wing disc is established in late stages 13-16 of the developing embryo when TFs such as *hth, exd, vg, sna, esg* become transcriptionally active in the wing imaginal cells. Growth and pre-patterning of the imaginal disc takes place in the larva with a number of genes expressed presaging the development of adult structures. The TFs *en, hh, dpp, wg, antp, ubx, exd, hth, dll* etc. have been implicated in pre-patterning of the imaginal disc into compartments, while the TFs *ap, pan, su(H), nub,* *mad, sd, ara, e(spl), ubx* are known to occur in CRMs mediating spatio-temporal specific expression of genes in the wing imaginal disc.

VI-8 Implications of Modulexplorer

The Modulexplorer Bayesian network model describes a CRM as a cluster of TFBSs for TFs that co-regulate gene expression in a particular tissue and development stage. The TFBS combination defines a regulatory code. The regulatory codes were learnt de-novo in this study from a repository of CRMs of unknown types. CRMs sharing a common set of motifs were found to regulate the same spatio-temporal specific gene expression. In previous studies [Li et al. (2007)], low sequence similarity has been reported among CRMs. This is true as the 414 CRMs comprising the training and test data in this study had at most 40% sequence similarity, while in average lower than 20%. However in this study we observed similarity of CRMs in terms of their shared TFBS or motif content. Therefore a new notion of similarity among CRMs emerges.

Though we used the common motifs among "similar" CRMs to specify a regulatory code, the Modulexplorer model originally learns regulatory codes in the form of probabilistic interaction among the TFBSs or motifs. We studied such interactions at the most basic level in the pairwise TF-TF interaction matrix. The observed pairwise interactions could be corroborated with known biology. The Modulexplorer model thus suggests that regulatory codes exist as rules of probabilistic TF-TF interactions.

Modulexplorer also gives clues for the improvement of sequence-based modeling of regulatory sequences such as using oligonucleotide motifs. It was observed during this study that oligonucleotide motifs produce large number of matches in non-TFBS segments of the CRMs (e.g. Figure VI-5), which reduces the effectiveness of modeling based on motifs. The performance of the model improved when the TFBSs were accurately annotated and used to train the model, e.g. Modulexplorer showed better discrimination between CRMs and background. Similar enhancements may be possible in other applications of sequence-based modeling.

Modulexplorer contributes new biological information of regulatory codes for CRMs associated with the development of mesoderm, ventral nerve cord, eye-antennal disc and the wing imaginal disc. It also provides functional annotation of genes, for instance 31 new genes have been classified in the above developmental functions. The roles of some TFs in these regulatory mechanisms were also suggested, such as the novel role of deaf1 in eye-antennal disc development.

The regulatory codes currently discovered are few in number as the application of Modulexplorer model is restricted to CRMs that have been previously characterized. Also the current method of CRM clustering using frequent itemset mining is not robust enough as it can only discover clusters where a sufficient number of CRMs share a large number of common motifs. Also the model currently relies on homotypic clustering to accurately discover the TFBS. This may not be possible in other species or in CRMs where homotypic clustering is absent. In such scenarios, one of the possibilities could be to characterize TFBS by motif discovery in a set of known CRMs having the same motif module [Gupta and Liu (2005)].

CHAPTER - VII

CONCLUSIONS AND FUTURE WORK

This research utilized position localization of TFBSs in regulatory sequences to enhance their computational modeling and prediction. Three different applications were addressed in particular – DNA motif detection, general promoter prediction, and cis-regulatory module prediction. Although positional bias of TFBSs in regulatory element has been known, it has not been adequately studied and exploited. The present research focused on this aspect and contributed three new tools to the bioinformatics community – LocalMotif and BayesProm, Modulexplorer. The salient research conclusions are summarized below with some directions for future work.

VII-1 Role of Positional Localization of TFBSs

Positional localization of TFBSs has been observed in a number of situations in gene regulatory sequences. In this dissertation, the following specific scenarios were considered:

(1) The positional localization of TFBSs with respect to the gene promoter in the mechanism of transcriptional initiation.

(2) The positional localization of TFBSs of co-regulating transcription factors with respect to the binding sites of the main transcription factor.

(3) The positional localization of closely packed TFBSs with respect to each other in an enhancer sequence (CRM).

There are other scenarios where positional localization of TFBSs may occur not because of biological reasons but due to the nature of the experiment itself. For instance in the emerging ChIP sequencing technology, the main TFBS and the binding sites of the co-regulating TFs are found localized with respect to the “peaks” or the positions of maximal overlap of the sequenced ChIP fragments.

In this dissertation, it was shown in all of the above scenarios that positional localization can be utilized to improve the quality of bioinformatics analysis. In addition the results of this study also enhanced our understanding of the nature of positional localization of TFBSs that exists in these scenarios. The contributions of each chapter in these aspects are summarized below.

In the localized motif detection problem of Chapter 3, the positional localization of the TFBSs relative to the TSS or a related TFBS was used to improve the performance of motif finding. In the formulation of this problem, localized motifs were distinguished from randomly locally over-represented patterns by their spatial confinement within a certain position interval of the sequences when compared to the full sequence. The *Spatial confinement score* (SCS) was derived as a statistical measure of the significance of the observed localization. The SCS was found very useful to discover biologically meaningful patterns. In cases where the biological motif becomes subtle for a usual motif detection algorithm, its high SCS still makes it conspicuous to a localized motif detection algorithm. Based on this concept, the software tool LocalMotif consistently showed higher accuracy compared to general motif finders in detecting localized motifs. Spatial confinement score also reduces the chances of detecting false positive patterns. Thus motif finding tools can improve their accuracy near the TSS or other such biological contexts where a specific biological landmark exists, such as splice site, ribosome binding site, etc. by considering positional localization information.

On datasets where binding sites of a known TF are available, the co-regulatory motifs could be detected by LocalMotif by the virtue of their positional localization though these were invisible to other motif finders. This was shown by the example of the dataset containing estrogen response elements (ERE), where the forkhead binding sites are often present near the ERE. In this scenario the use of localization information is promising and the same concept can be applied to datasets derived from ChIP sequencing. In ChIP-seq, the main motif is highly localized at the center of the peak and thus the co-regulatory motifs can also be found localized around the peak center. This is an emerging area of research and a localization scoring function similar to that in LocalMotif could be used to detect co-regulatory factors with high accuracy.

In the analysis of real genomic sequences in this chapter, motifs in the core, proximal and distal promoter regions were detected by LocalMotif automatically by the virtue of their localization around the TSS. This was observed with the set of 1941 promoters of Drosophila Melanogaster as well as with promoters of orthologous genes in vertebrate genomes. The results showed that real motifs near the TSS are positionally localized. This confirms that TFBSs are positionally distributed around the TSS, a fact which is used in Chapter 4 of this dissertation for promoter prediction.

The computational modeling and prediction of eukaryotic promoters in Chapter 4 was performed using oligonucleotide positional densities instead of oligonucleotide over-occurrence as the basis of the computational model. The use of preferred positions of various TFBSs in the promoter relative to TSS in this case could easily give good differentiation between promoter and non-promoter sequences. The program BayesProm did not require any background model to do the predictions and performed comparable to second generation promoter prediction tools which are based on extensive tuning of parameters and often use other biological information as well. BayesProm was in fact more sensitive than any other program. This not only confirms the observation of positional localization of TFBSs around the TSS, but also suggests that positional information of TFBSs is in itself a distinguishing feature of functional regulatory sequences and is very relevant in bioinformatics analysis of gene regulation. The positional localization of different features in different types of regulatory sequences is an important aspect to be researched further towards understanding the control circuitry embedded in these sequences.

In the case of cis-regulatory modules, which are distal elements and so far not much understood, researchers have generally emphasized high density of binding sites for co-operating TFBSs as their main feature. In the computational model Modulexplorer, firstly the TFBSs were discovered de-novo in a CRM with high accuracy when the motifs were considered in pairs rather than as single patterns. Secondly the motifs were found to occur in specific combinations and with specific mutual gap and order in the CRMs. Incorporating this information into the modeling improved the specificity of the CRM model (i.e. reduction in false positives). This indicates that positional information is also important in CRMs, or in general distal regulatory elements. In this case the positional localization of TFBSs is not with respect to a certain fixed biological landmark but mutually with respect to each other. The TFBSs were observed to be closer to each other in a CRM as compared to random patterns. The possible reason for this could be to allow interaction among the TFs. This subject needs to be explored in greater depth and could improve our understanding of the features of functional regulatory sequences distal to the

TSS. While in this study only pairwise order and distances of TFBSs were modeled, the actual situation could in fact involve interactions of greater complexity among the TFs.

Thus in summary this dissertation not only shows the advantage of using positional information in the bioinformatics analyses of regulatory sequences, but also throws light on the different natures of TFBS localization in gene regulatory sequences proximal and distal to the TSS.

VII-2 Nature of Regulatory Sequences

As described in the previous section, analyses performed using LocalMotif and BayesProm in this research confirm the current view about the nature of gene promoters. Both analyses showed that transcription factors bind to the promoter at specific positions relative to the TSS. Analysis of Drosophila core promoters with LocalMotif showed the localization of binding sites such as TATA box, initiator, DRE, DPE etc. Similarly in human core and proximal promoters BayesProm showed the localization of TATA box, CAAT box, GC box and initiator. Furthermore, the localization intervals of binding sites near to the TSS were shorter than the intervals of binding sites distal of the TSS, which supports the current understanding of proximal and distal promoter regions.

Modulexplorer analysis, on the other hand, provides novel information about the nature of gene enhancers or cis-regulatory modules. As in previous studies, the model confirms that a CRM is a cluster of TFBSs for TFs that co-regulate gene expression in a particular tissue and development stage. It also confirms that the TFBSs are positionally localized with respect to each other in a CRM. However, while previous studies report low sequence similarity among CRMs, Modulexplorer shows good similarity of CRMs in terms of their shared TFBS or motif content. Therefore it gives a new notion of similarity among CRMs. This conclusion could be extended in general towards assessing the similarity of regulatory sequences. According to this view, regulatory sequences can be considered as a combination of the functional (TF binding) and the non-functional (background) parts. While the non-functional part is usually dissimilar, the functional part could be quite similar or conserved across sequences. Measuring similarity over the functional part only could be a better idea.

The Modulexplorer model furthermore shows that there is organization within the functional (TF binding) part of the CRM. The occurrences of various TFBSs in a CRM are not random. There are probabilistic rules governing which TFBSs occur together and which do not. Such rules can further be used to characterize true regulatory sequences and to assess their similarity / dissimilarity.

Furthermore, the homotypic clustering of TFBSs in a CRM, i.e. occurrences of multiple binding sites for the same TF in a CRM, is confirmed in this study. This property was utilized to discover the TFBSs in the CRMs.

VII-3 Modeling Techniques

Statistical information theory and probabilistic graphical models were used in this dissertation for modeling. The conclusions drawn from the application of these techniques to the present research problems are described below.

In Chapter 3, information theory framework was found very useful for formulating the motif scores (over-representation score, spatial confinement score and relative entropy score). When formulated in the context of information theory, these scores measure the amount of surprise associated with the motif in the particular aspect. For example the over-representation score measures the amount of surprise in the observed number of instances of the motif as compared to its expected number of instances. In mathematical terms, the score represents the Kullback-Leibler distance between the observed and the reference distribution. The scores are normalized with respect to suitable bases so that they usually lie in the range 0 to 1. A score close to zero indicates little surprise while a score close to 1 indicates high surprise. Information theoretic definition of the scores is thus helpful in obtaining a clear quantitative picture of the goodness of the motif in the three different aspects. It also allows combining the three scores into a single score easily and logically. Since the three scores measure three independent characteristics of the motif, the score can be considered as a vector in a three dimensional space which each score measured along an orthogonal axis. The combined score can then be stated as the Euclidean distance or Hamming distance of the motif from the origin. Motifs distant from the origin are more interesting than those closer to the origin. The information theoretic scoring measure is also comparable for motifs of different lengths and mutations. Thus it allows pooling together the results of different (l,d) runs and then selecting the best motifs among all the runs. Future motif finding algorithms and other bioinformatics tools as well can take advantage of this information theoretic framework for computing and combining scores.

In Chapter 4, a mixture of Gaussians was used to represent the oligonucleotide positional density. The parameters of the Gaussian mixture were estimated using the EM algorithm. Instead of learning all the components of the Gaussian mixture simultaneously, the mixture was built component-wise. Initially the mixture has one component which is learnt by EM, then a new component is added and the EM is repeated to learn the two-component mixture, and similarly one component is added per step until the optimal solution is reached by AIC. Such component-wise building of the Gaussian mixture was found highly effective. Future applications using Gaussian mixture models can benefit from this idea.

In both Chapters 4 and 5, Bayesian networks were used for the modeling. In Chapter 4, a continuous naïve Bayes network was used, whereas in Chapter 5 a discrete Bayesian network was used. In both applications, Bayesian networks provided considerable advantages as compared to other AI modeling techniques. These advantages are described below.

The first advantage of Bayesian networks is in meaningfully representing physical entities or phenomena in the network structure. In Chapter 4, the nodes of the BayesProm model represent TFBSs. The parameters of the model encode the occurrence distributions of the TFBSs relative to the TSS. Thus upon learning the model from a set of human promoters, the Bayesian network automatically identified important TFBSs in these promoters and their occurrence positions relative to the TSS. The model gathered physical domain knowledge from the data into the network structure during training. Blind classifiers such as neural network, SVM etc. do not allow such a meaningful physical representation. In the Modulexplorer model of Chapter 5, the nodes of the network represented monad and dyad motifs and the CRM. The model structure represented biological knowledge of how the monad motifs form dyads and how these dyads combine together to form a CRM. Thus the Bayesian network model could meaningfully incorporate known biology knowledge into the model. This is again not possible in neural networks and SVM.

The second advantage of using Bayesian networks is in allowing validation of the model based on known biology. Since the nodes and parameters of the Bayesian network have physical interpretation, they can also be validated with existing biology knowledge. The parameters of the BayesProm model, which represent occurrence distributions of the TFBSs relative to the TSS, were verified after training the model. The parameters showed that the prominent features in BayesProm corresponded to known TFBSs, and the TFBS positions were also correctly determined. In the Modulexplorer model, the parameters of the Bayesian network represent interaction probabilities among the dyad motifs in a CRM. Since the dyad motifs correspond to binding sites of known TFs, in effect the model parameters represent TF-TF interactions. The pairwise TF-TF interactions were verified after training the Modulexplorer model. The TF-TF interactions in the model compared well with known TF-TF interactions, confirming the validity of the model. Such validations are not possible in other AI modeling tools.

A related advantage is that the Bayesian network can even provide new knowledge in its parameters after training. The TF-TF interactions in the Modulexplorer model not only confirmed with existing knowledge but also showed a novel interaction of the TF *Deaf1* with the TFs *ey* and *toy* which are involved in eye-antennal development in Drosophila melanogaster. Thus they implicated *Deaf1* in eye-antennal development.

Another crucial advantage of Bayesian network modeling in Modulexplorer is that the model could be trained with very less data. The training data of CRM sequences was few in number, with only 356 training sequences. However robust learning was possible since domain knowledge was incorporated into the modeling. Parameterization in the EM algorithm was reduced by (i) establishing a noisy-AND relationship between the motifs in a dyad, (ii) having preset values of gap and order CPTs by learning them directly from the data, and (iii) by intelligently choosing the dependencies in the network structure.

Thus in summary Bayesian networks are highly effective modeling tools in bioinformatics.

The Modulexplorer model also shows that sequence-based modeling of regulatory sequences can be considerably improved by considering only the functional part of the regulatory sequences (i.e. the TFBSs). Previous modeling techniques look for matches of motifs in the whole sequence. The motifs produce large number of matches in non-functional (non-TFBS) segments of the sequence, which reduces the accuracy of modeling. A better way of modeling is to accurately annotate the TFBSs and use only the motif matches in the TFBS segments. Based on this approach, Modulexplorer showed better discrimination between CRMs and background as compared to previous tools.

VII-4 Research Contributions

The main novel contributions of the present study towards bioinformatics research are summarized as follows.

This study introduced a new formulation of the motif finding problem as localized motif finding. The difference between locally over-represented and localized motifs was clearly defined, and based on this a clear definition of the localized motif finding problem was given. To solve this problem, a novel scoring function called the spatial confinement score was introduced. This score is computed in the form of an information criterion. The existing scoring functions of over-representation and relative entropy were also reformulated in an information theoretic form and normalized with suitable bases so that they usually lie within the range 0 to 1. The work has contributed a novel motif finding algorithm called LocalMotif to the bioinformatics community. The algorithm is published and available to the research community for free use. It has good potential of application for the analysis of co-regulatory motifs in ChIP-Seq datasets.

For modeling promoter sequences, this study introduced the idea of using positional localization of motifs relative to the TSS. In the current work the idea was implemented in the form of oligonucleotide positional densities and a continuous naïve Bayes model. However, other better representations may be possible. This study shows that positional information of TFBSs in itself gives high accuracy of promoter prediction, and thus when combined with biological knowledge as in other available tools, can further improve the accuracy. The work has also resulted in a new published tool called BayesProm which can be used by the research community to analyze TSSs in the human genome.

The work on Modulexplorer makes some major contributions to the research on CRMs. Existing research has focused on CRMs of a single type which express their target genes in same tissue and developmental stage. This is the first work to attempt the study of multiple types of CRMs that express their genes in several different tissues and developmental stages.

The first contribution of Modulexplorer study is the compilation of a comprehensive database of Drosophila CRMs with full TFBS annotation. Prior to this study the experimental TFBS annotation was available only for 19 CRMs fully and 136 CRMs partially. This study introduced a novel method for de-novo discovery of TFBSs, which has more than 80% sensitivity and about 20% false positive rate. Using this method, full TFBS annotation of all 619 Drosophila CRMs has been produced.

The second contribution is in highlighting a clear definition of regulatory codes and finding novel regulatory codes govening Drosophila CRMs. Though the term regulatory code and its concept are available in the published literature in different places, a comprehensive description and application of the concept are not available in any single publication. This study introduces the concept clearly and presents a model that can learn regulatory codes de novo from training data of CRMs. The study has contributed new regulatory codes for Drosophila CRMs associated with the development of mesoderm, ventral nerve cord, eye-antennal disc and the wing imaginal disc.

A related novel contribution is the use of a database of in-situ expression profile images of genes in Drosophila embryos to validate the functions of the CRMs. It was hypothesized that CRMs sharing the same regulatory code, i.e. the same motif modules, must show the same expression profiles of their target genes in a certain developmental stage and tissue. This was indeed confirmed with the help of in-situ expression images of these genes.

The third contribution is the prediction of 813 novel CRMs in Drosophila. A majority of these CRMs are also classified with their possible roles in development, i.e. the tissue and development stage in which they express their target genes. With this discovery, a related contribution is the novel functional annotation 31 Drosophila genes in development of mesoderm, ventral nerve cord, eye-antennal disc and wing imaginal disc.

Another biological contribution made by the Modulexplorer study is to suggest the roles and interactions of different TFs in regulatory mechanisms, such as the novel role of *deaf1* in eye-antennal disc development.

Finally the study contributes the tool Modulexplorer to the research community, which can be used to model and discover CRMs in Drosophila. The pipeline of Modulexplorer gives a systematic description of how enhancer modeling and prediction can be performed with suitable validations at every step. The ideas presented in this dissertation have the potential to stimulate future works in the modeling and prediction of enhancers.

VII-5 Recommendations for Further Study

The research problems addressed in this thesis are currently of active interest in bioinformatics. Some specific research directions motivated by the present research and some ideas for extending the present work are described below.

1. LocalMotif is presently based on the (*l*,*d*) motif model. There are emerging opinions about improving the motif representation to reduce false positives or to give a more accurate description of the motif. While it is still not clear which representation is the best, it has been suggested that (i) motifs with possible gaps, (ii) motifs with mismatches restricted to specific binding site positions, or (iii) motifs based on IUPAC characters, might lead to a more accurate model. The motif model of LocalMotif may be revised to study this effect while retaining the same scoring function.
2. The LocalMotif algorithm is currently derived as a modification of the SP-STAR algorithm of Pevzner and Sze (2000). It would be more efficient to use a faster

algorithm with a broader search space such as the suffix tree algorithm of Weeder [Pavesi et al. (2001)].

3. The BayesProm program was trained on 1796 human promoters from EPD version 74. Much more extensive public repositories of human promoters such as DBTSS have become available. The model may be retrained on the extended dataset for improved performance. In addition, a whole genome search could be performed to output a list of predicted genome wide binding sites.
4. Feature selection can be attempted on the BayesProm naïve Bayes model to remove unprofitable oligonucleotides from the model. Introducing dependencies among the attribute nodes in the Bayesian network model (such as a TAN Bayesian model) could also be tried.
5. Specific biological knowledge could be added to the BayesProm software to further improve its performance. For example (i) separately training two different models for CpG island and non-CpG island related promoters, (ii) coupling a gene prediction algorithm with BayesProm as has been done in FirstEF [Davuluri et al. (2001)] and Dragon gene start finder [Bajic and Seah (2003)].
6. A major research problem could be to apply the Modulexplorer concept to vertebrate CRMs. This would firstly require developing a procedure to annotate TFBSs in vertebrate CRMs. Vertebrate CRMs are different from insect CRMs in that they contain binding sites for a larger variety of TFs and have fewer instances of homotypic clustering of TFBSs. Thus they would require a different approach for TFBS annotation. The Bayesian network model would also have to be modified accordingly.

7. In Modulexplorer, feature based clustering of CRMs has currently been performed using frequent itemset mining. However the Bayesian network model originally models probabilisitic interactions among the motifs in CRMs. Thus better ways to perform clustering could be explored, which may allow separation of CRM clusters of smaller sizes with common function. This will lead to the discovery of more regulatory codes and refinement of the existing ones.
8. The novel enhancers discovered by Modulexplorer can be validated in the wet lab by generating P-element constructs fused to *eve* basal promoter and a *lacZ* reporter gene and examining the expression of these constructs by in-situ RNA hybridization to the *lacZ* transcript in the embryo in the desired tissue and stage of embryogenesis.
9. The new motifs reported by Modulexplorer in the development of mesoderm, ventral nerve cord, eye-antennal disc and the wing imaginal disc could be studied to see if they give any new biological information about transcription factors in Drosophila. Deletion studies could be carried out to validate these motifs.

APPENDIX

APPENDIX A. Spatial Confinement Score in LocalMotif

A.1 Spatial confinement

Consider a $\left(l,d\right)$ motif M with its instances (relative to the anchor point) observed in a large set of sequences, $\tilde{S}$, of length L each, aligned relative to an anchor point *A*. *Spatial confinement* of *M* within a position interval $\left(p_1,p_2\right)$ is defined as the difference between the fraction of binding sites actually observed within the interval $\left(p_1,p_2\right)$ and the fraction that would be expected to lie in it if binding sites were uniformly distributed across the entire sequence length. For instance a length $L/2$ interval $\left(p,p+L/2\right)$ is expected to contain 50% of the observed binding sites if they were uniformly distributed. But if this interval contains 65% of the total binding sites, then it has +0.15 spatial confinement of *M*.

Spatial confinement always lies in the range $\left(-1,1\right)$. Its positive value in an interval signifies higher than expected binding site concentration in that interval. Figure A-1 shows the spatial confinement of the motif TTGACA in E. coli promoter sequences for various intervals. The interval length $\left(p_2-p_1\right)$ is shown on the *x*-axis and the interval beginning position p_1 is shown on the *y*-axis. Spatial confinement is shown as a surface in the *z*-axis. Maximum spatial confinement is observed for the interval (30,50) indicating that the motif is confined within this interval. The interval is indeed biologically accurate [Harley and Reynolds (1987)].

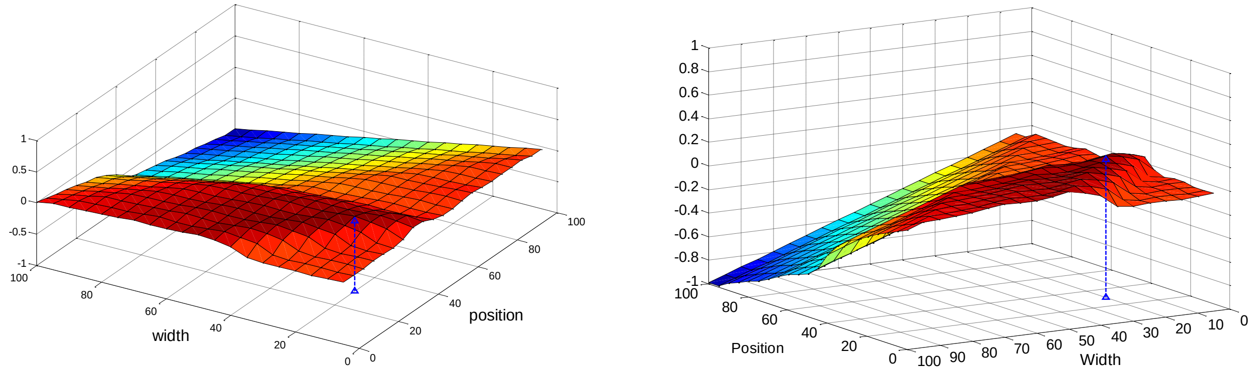


Figure A-1 Spatial confinement of the motif TTGACA in different intervals $\left(p_1, p_2\right)$ in a set of 471 E. coli promoter sequences of length 101 each. The *x*-axis denotes position p_1 and the *y*-axis denotes the interval width $\left(p_2 - p_1\right)$. Maximum is observed at $p_1 = 30$ and width=20, indicating that the motif is confined within the interval (30,50), which agrees with the literature [Harley and Reynolds (1987)].

Thus spatial confinement gives a picture of the relative concentration of binding sites for a motif in different position intervals, and can be used to identify the position interval where the motif is maximally confined. However in practice it is difficult to accurately compute it because the number of input sequences provided to the algorithm is mostly limited. The limited information can be utilized most effectively using statistical procedures. A statistical measure for spatial confinement is therefore derived as the *spatial confinement score*.

A.2 Spatial confinement score

Instead of the large sequences set $\tilde{S}$, let only a subset $\mathbf{S} \subset \tilde{S}$ be available as input to the algorithm. Thus **S** is a sample data from the population $\tilde{S}$. Let c be the concentration of binding sites for the motif M in position interval $\left(p_1, p_2\right)$ within the population $\tilde{S}$. An estimate $\hat{c}$ of c may be obtained from the sample **S**. Let n denote the total number of binding sites for M in the sequence set **S**, of which n_1 lie within the interval $\left(p_1, p_2\right)$ and $n_0 = n - n_1$ lie outside this interval. The maximum likelihood estimate is given as $\hat{c} = n_1 / \left(n_0 + n_1\right)$.

The spatial confinement of M in the interval $\left(p_1, p_2\right)$ is measured as the difference $c - c_0$, where c_0 is the concentration of binding sites expected in $\left(p_1, p_2\right)$ according to uniform density, given by $c_0 = \left|p_2 - p_1\right| / L$. Since the exact value of c is unknown, the problem is to assess from the sample estimate $\hat{c}$ whether or not $c > c_0$ in the interval $\left(p_1, p_2\right)$ and to what degree c exceeds c_0. This would be a statistical measure of the spatial confinement of M in $\left(p_1, p_2\right)$.

A statistical hypothesis test is defined to assess whether $c > c_0$ with the following elements: the null hypothesis, the alternate hypothesis, the test statistic and the rejection region. The two hypotheses are:

$$H_0 : c = c_0$$

$$H_1 : c > c_0 \quad (\text{one tailed})$$

The test statistic is derived via likelihood ratio procedure. The complete derivation is described in the Appendix B.2. In essence, it is stated as the Kullback-Leibler distance between $\hat{c}$ and c_0 in equation B.10, which is thus used as the statistical measure for spatial confinement, called as the *spatial confinement score*. The rejection region for the hypothesis test is also described in the Appendix B.2.

APPENDIX B. Normalization of Scoring Functions in LocalMotif

B.1 Normalization of the relative entropy score (RES)

The relative entropy of the motif is the Kullback-Leibler divergence between the motif *M* and the background *B*:

$$D(M \,||\, B) = \sum_{i=1}^{l} \sum_{b} f_{b,i} \ln\left(\frac{f_{b,i}}{p_b} \right), \tag{B.1}$$

which can be decomposed as

$$D(M \,||\, B) = \sum_{i=1}^{l} \sum_{b} f_{b,i} \ln\left(f_{b,i} \right) - \sum_{i=1}^{l} \sum_{b} f_{b,i} \ln\left(p_b \right), \tag{B.2}$$

i.e.,

$$D(M \,||\, B) = \sum_{i=1}^{l} \sum_{b} f_{b,i} \ln\left(f_{b,i} \right) - \sum_{b} \left(\sum_{i=1}^{l} f_{b,i} \right) \ln\left(p_b \right) \tag{B.3}$$

If $x_1 + x_2 + \ldots + x_n = 1$ then the maximum of the entropy function $-\sum_{i=1}^{n} x_i \ln\left(x_i\right)$ occurs for $x_1 = x_2 = \ldots = x_n = 1/n$ and the maximum value is $\ln(n)$. Therefore the first term can be normalized by the factor $(1/l \ln 4)$. Normalizing the second term by the same factor $(1/l \ln 4)$, it appears as $-\frac{1}{\ln 4} \sum_{b} \overline{f}_b \ln\left(p_b \right)$, where $\overline{f}_b = \left(\frac{1}{l} \sum_{i=1}^{l} f_{b,i} \right)$. For a uniform background, where p_b=0.25, this term reduces to 1 since $\sum_{b} \overline{f}_b = 1$. Another special case is when $\forall b, \overline{f}_b = p_b$. Then after normalization the term becomes ≤1. As the difference between $\overline{f}_b$ and p_b increases, the term can become >1.

B.2 Derivation and normalization of the spatial confinement score (SCS)

The spatial confinement score is derived as follows. According to uniform density, the proportion of binding sites of a motif that lie within the interval of interest $\left(p_1, p_2\right)$ will be $c_0 = \left|p_2 - p_1\right|/L$, where L is the sequence length. Considering that the population distribution is uniform (hypothesis H_0), in a randomly chosen sample, the likelihood of observing n_1 binding sites within the interval $\left(p_1, p_2\right)$ and n_0 outside this interval is given by the binomial formula:

$$\Pr\left(n_1 \text{ sites in } \left(p_1, p_2\right) \middle| c_0\right) = L\left(c_0 \middle| n_0, n_1\right) = {}^{n}C_{n_1}\left(c_0\right)^{n_1}\left(1 - c_0\right)^{n_0} \qquad \text{(B.4)}$$

Considering that the population distribution is non-uniform (hypothesis H_1), let the concentration of binding sites in the interval $\left(p_1, p_2\right)$ be *c*. The binding site observations are outcomes of a binomial experiment where a binding site lies within the interval $\left(p_1, p_2\right)$ with probability c and outside it with probability $\left(1 - c\right)$. If the total number of observed binding sites is *n*, of which n_1 lie within $\left(p_1, p_2\right)$ and $n_0 = n - n_1$ lie outside, then the likelihood of observing n_1 binding sites within the interval $\left(p_1, p_2\right)$ and n_0 outside this interval is again given by

$$\Pr\left(n_1 \text{ sites in } \left(p_1, p_2\right) \middle| c\right) = L\left(c \middle| n_0, n_1\right) = {}^{n}C_{n_1}\left(c\right)^{n_1}\left(1 - c\right)^{n_0} \qquad \text{(B.5)}$$

The maximum likelihood estimate $\hat{c}$ of *c* is thus obtained as

$$\frac{dL}{dc} = 0 \quad \Rightarrow \quad \hat{c} = \frac{n_1}{n_0 + n_1} \qquad \text{(B.6)}$$

The likelihood ratio test statistic λ_1 for the hypothesis test defined in Section IV-2.3 is then obtained as

$$\lambda_1 = \frac{L(c_0)}{L(\hat{c})} = \frac{c_0^{n_1}(1-c_0)^{n_0}}{\hat{c}^{n_1}(1-\hat{c})^{n_0}} \tag{B.7}$$

and the rejection region is determined by

$$RR: \lambda_1 \le k \tag{B.8}$$

where k is chosen according to the desired level of significance α of the test. According to the Wilks' theorem [Rice (1995)], $-2\ln\lambda_1$ is approximately χ^2 distributed with one degree of freedom. This information can be used to derive the value of k given a fixed level of significance α. If λ_1 lies in the rejection region then there is sufficient evidence to conclude that the concentration of binding sites for the motif M in the interval (p_1, p_2) is greater than what would be expected from uniform density. As the value of λ_1 approaches zero, the hypothesis H_1 is favoured increasingly over H_0. The likelihood ratio test statistic λ_1 is related to the Kullback-Leibler distance between $\hat{c}$ and c_0 as

$$D(\hat{c} \,||\, c_0) = -\frac{1}{n}\ln\lambda_1 \tag{B.9}$$

which can be shown to be equal to

$$-\frac{1}{n}\ln\lambda_1 = \hat{c}\ln\left(\frac{\hat{c}}{c_0}\right) + (1-\hat{c})\ln\left(\frac{1-\hat{c}}{1-c_0}\right) \tag{B.10}$$

The above equations are used as the statistical measure for the spatial confinement score. It is already in a normalized form being independent of motif length etc.

B.3 Derivation and normalization of the over-representation score (ORS)

Searching for motif instances (TFBS) in a set of sequences can be considered as a binomial experiment where patterns of length *l* are drawn from the sequences and each pattern is classified as either a motif instance or a non-instance. The probability of observing k instances of the motif among a total of n samples is given by:

$$P(k,n) = \Pr(k \text{ "true" in } n) = {}^{n}C_{k}\, p^{k} (1-p)^{n-k}, \tag{B.11}$$

where *p* is the proportion of TFBS in the sequences. For example, under the (*l*,*d*) motif representation, the chance proportion p_0 of the TFBS of a motif according to uniform background is computed theoretically as follows:

$$n_0 = 4^{l},\ k_0 = \sum_{i=0}^{d} {}^{l}C_i (3)^{i}, \text{ and } p_0 = k_0 / n_0 . \tag{B.12}$$

The background probability distribution of the TFBS is then

$$P_0(k_0, n_0) = {}^{n_0}C_{k_0}\, p_0^{k_0} (1-p_0)^{n_0-k_0} . \tag{B.13}$$

If the background distribution is not uniform, the expression will be modified in a suitable manner to incorporate the individual probabilities of each of the k_0 patterns that match the (*l*,*d*) motif.

As *n* grows to be large, specifically if both $np > 5$ and $n(1-p) > 5$, the binomial distribution may be approximated by the Gaussian distribution $\mathcal{N}(np, np(1-p))$. Thus,

$$P_0(k_0, n_0) = {}^{n_0}C_{k_0}\, p_0^{k_0} (1-p_0)^{n_0-k_0} \approx p_0(x) \sim \mathcal{N}(n_0 p_0, n_0 p_0 (1-p_0)), \tag{B.14}$$

If in n_1 actual trials (i.e., upon searching the set of sequences consisting of n_1 oligonucleotides of length *l*) the observed number of matching patterns be k_1. This represents an observed proportion $p_1 = k_1/n_1$. Hence the observed probability distribution of the TFBS is:

$$P_1(k_1, n_1) = {}^{n_1}C_{k_1} p_1^{k_1} (1-p_1)^{n_1-k_1} \approx p_1(x) \sim \mathcal{N}(n_1 p_1, n_1 p_1 (1-p_1)). \tag{B.15}$$

The Z-score for computing the over-representation is based on the Gaussian approximation:

$$z = \frac{n_1 p_1 - n_1 p_0}{\sqrt{n_1 p_0 (1-p_0)}}. \tag{B.16}$$

The Z-score is not a normalized measure as it depends upon the number of samples n_1.

An entropy measure for over-representation derived directly (without Gaussian approximation) from the binomial distribution in a normalized form is used in LocalMotif. It is obtained as the Kullback-Leibler divergence between the two binomial distributions:

$$D(P_0 \,||\, P_1) = \sum_{k=0}^{n} P_0(k,n) \ln\left(\frac{P_0(k,n)}{P_1(k,n)}\right) \tag{B.17}$$

Upon expanding the above expression for KL divergence and normalizing, it turns out that the expression may be simplified as:

$$D(P_0 \,||\, P_1) = p_0 \ln\left(\frac{p_0}{p_1}\right) + (1-p_0)\ln\left(\frac{1-p_0}{1-p_1}\right), \tag{B.18}$$

which is independent of the number of samples *n*. This is used as the measure for over-representation in LocalMotif.

For a Gaussian approximation of the binomial distribution, the KL divergence between two Gaussians is given by

$$D\left(N_0 \,||\, N_1\right) = \frac{1}{2}\left(\frac{\sigma_0^2}{\sigma_1^2} + \frac{\left(\mu_0 - \mu_1\right)^2}{\sigma_1^2} - 1 - \ln\left(\frac{\sigma_0^2}{\sigma_1^2}\right)\right), \tag{B.19}$$

and thus

$$D\left(P_0 \,||\, P_1\right) = \frac{1}{2}\left(\frac{p_0\left(1 - p_0\right)}{p_1\left(1 - p_1\right)} + \frac{\left(p_0 - p_1\right)^2}{p_1\left(1 - p_1\right)} - 1 - \ln\left(\frac{p_0\left(1 - p_0\right)}{p_1\left(1 - p_1\right)}\right)\right) \tag{B.20}$$

which is approximately identical with the previous expression for most cases, except when p_0 or p_1 have extreme values that are close to 1 or 0, in which case the Gaussian approximation has significant error.

APPENDIX C. Fast Computation of Scores in LocalMotif

The equations for fast computation of score for a longer interval from scores for shorter constituent intervals are derived as follows. Let $p_1 < p_2 < p_3$, and let quantities for the interval $\left(p_x, p_y\right)$ be denoted with superscript *xy*. Then,

$$\begin{aligned} n_1^{13} &= n_1^{12} + n_1^{23} \quad \Rightarrow \quad c^{13} = c^{12} + c^{23} \\ n_0^{13} &= n_0^{12} + n_0^{23} \quad \Rightarrow \quad c_0^{13} = c_0^{12} + c_0^{23} \end{aligned} \tag{C.1}$$

$$\begin{aligned} p_0^{13} &= p_0^{12} + p_0^{23} \\ p_1^{13} &= p_1^{12} + p_1^{23} \end{aligned} \tag{C.2}$$

SUPPLEMENTARY FIGURES

Supplementary Figure 1. Results of running LocalMotif on the dataset of [Blanchette and Tompa (2002)]. Predicted binding sites reported by [Blanchette and Tompa (2002)] are shown in the left column, while the corresponding motifs predicted by LocalMotif are shown alongside to the right. Binding sites having experimental evidence are marked with an asterisk.

Gene	Motifs reported by Blanchette and Tompa		LocalMotif Predictions							
	Motif	Position	Motif	Pos Start	Pos End	Total Score	RES	SCS	ORS	Rank
c-fos	CAGGTGCGAATGTTC	-615	CAGGTGC	-650	-600	0.826	0.5	0.277	0.05	7
			GTGCGAA	-615	-565	0.707	0.474	0.183	0.05	29
1			TGTTCTCT	-605	-505	0.692	0.349	0.295	0.048	33
			TAATGTT	-705	-585	0.685	0.424	0.223	0.038	34
			GTTCGC	-630	-580	0.679	0.42	0.134	0.125	36
			ATGTTC	-610	-560	0.669	0.445	0.153	0.071	41
			CGAAAGT	-645	-595	0.625	0.444	0.131	0.05	54
			GTAATGTT	-625	-575	0.548	0.332	0.173	0.043	61
2	TTCCCGCCTCCCCTCCCC*	-583	CCCCCC	-575	-525	1.442	0.352	0.09	1.00	1
			CCCCCCCC	-580	-530	1.403	0.292	0.111	1.00	2
			CCCCCCC	-580	-530	1.39	0.318	0.072	1.00	3
			TCCCCG	-575	-525	0.783	0.375	0.064	0.343	11
			CCCGGC	-590	-540	0.744	0.382	0.04	0.323	17
			GCCGCC	-590	-540	0.734	0.361	0.05	0.323	19
			CTCGCCT	-595	-540	0.732	0.4	0.123	0.21	20
			CTTCCC	-585	-535	0.717	0.375	0.06	0.281	23
			CTTCTCCC	-580	-530	0.711	0.331	0.146	0.235	25
			CCTCCT	-590	-540	0.697	0.392	0.104	0.201	31
			TCGCCTTC	-600	-550	0.676	0.414	0.188	0.074	37
			TCGCCT	-590	-540	0.665	0.424	0.117	0.125	43
			TCCCGGCC	-595	-545	0.623	0.319	0.081	0.223	55
			CCCTCCTT	-600	-530	0.536	0.35	0.097	0.089	62
			CCCCTGCG	-575	-525	0.509	0.316	0.064	0.129	64
3	GAGTTGGCTGcagcc	-527	TTGGCCG	-625	-480	0.78	0.444	0.308	0.028	12
			GTTGTC	-565	-460	0.744	0.413	0.274	0.057	18
			GTTGTCT	-530	-455	0.683	0.423	0.202	0.058	35
			GTTGGCTG	-530	-480	0.592	0.376	0.173	0.043	58
4	GTTCCCGTCAATCcct*	-504	GTCAAT	-505	-455	0.838	0.475	0.275	0.088	6
			CCGTCAA	-505	-455	0.789	0.437	0.253	0.10	10
			GTCCATC	-505	-455	0.758	0.401	0.204	0.153	16
			TCGTCA	-500	-435	0.702	0.442	0.203	0.057	30
			TCCCGTT	-505	-455	0.675	0.391	0.131	0.153	38
			CCCGTT	-515	-450	0.67	0.405	0.14	0.125	40
			CGACAA	-530	-480	0.662	0.457	0.167	0.038	44
			GGTCCTGT	-525	-475	0.647	0.389	0.215	0.043	49
			CACGTC	-530	-420	0.646	0.426	0.175	0.046	50
			ATCCCGTT	-510	-460	0.566	0.345	0.169	0.053	59
			CCGGCAAG	-515	-465	0.517	0.322	0.142	0.053	63
			TTCCCCCC	-510	-460	0.497	0.339	0.029	0.129	65

5	**CACAGGATGTcc***	-479	CAGGATGT	-465	-415	0.817	0.409	0.346	0.063	8
			AAGGAT	-480	-415	0.771	0.461	0.266	0.044	13
			CAGGAG	-485	-435	0.721	0.379	0.18	0.162	21
			CATAGG	-480	-425	0.713	0.5	0.181	0.032	24
			GATGTA	-475	-415	0.71	0.472	0.186	0.052	26
			GGATGG	-480	-430	0.661	0.427	0.18	0.054	45
			CAGGCT	-480	-430	0.658	0.389	0.144	0.125	46
			TCACAG	-480	-430	0.652	0.41	0.17	0.071	47
			CACAGGA	-475	-425	0.648	0.443	0.131	0.075	48
6	**AGGACATCTG***	-462	GGATATCT	-475	-415	0.967	0.45	0.478	0.039	4
			GATATC	-500	-445	0.81	0.443	0.292	0.076	9
			GACAACT	-495	-445	0.707	0.474	0.183	0.05	27
			CATCTT	-495	-435	0.672	0.435	0.173	0.065	39
			AGCACGTC	-510	-430	0.605	0.32	0.219	0.067	56
7	**GTCAGCAGGTTTCCACG***	-439	ATGTATCC	-475	-415	0.84	0.438	0.37	0.031	5
			CAGGATGT	-465	-415	0.817	0.409	0.346	0.063	8
			TTTCCA	-475	-425	0.764	0.407	0.175	0.182	14
			GTATCC	-445	-390	0.76	0.442	0.258	0.06	15
			CAGGAG	-485	-435	0.721	0.379	0.18	0.162	21
			GCAGGTT	-480	-430	0.707	0.474	0.183	0.05	28
			TCGTCA	-500	-435	0.702	0.442	0.203	0.057	30
			CAGGCT	-480	-430	0.658	0.389	0.144	0.125	46
			CACGTC	-530	-420	0.646	0.426	0.175	0.046	50
			GGTTTCCA	-480	-430	0.633	0.358	0.212	0.063	52
			AGCACGTC	-510	-430	0.605	0.32	0.219	0.067	56
			TCATCAGC	-485	-435	0.599	0.357	0.179	0.063	57
8	**TACTCCAACCGC**	-159	ATACTCC	-185	-135	0.717	0.42	0.222	0.075	22
			ACTGCA	-165	-115	0.693	0.416	0.133	0.143	32
			CTAACC	-160	-70	0.669	0.438	0.201	0.03	42
			CTCCAAC	-185	-135	0.637	0.398	0.113	0.126	51
			ACTTCGAC	-160	-105	0.628	0.366	0.225	0.036	53
			GCTCCTAC	-200	-150	0.566	0.36	0.132	0.074	60
c-myc	**aGTTTATTC**	-611	GTTTAC	-650	-580	0.643	0.455	0.131	0.058	23
10	**TTGCTGGG**	-570	ATTTTGCT	-575	-520	0.798	0.436	0.241	0.121	9
			TTGTTG	-585	-520	0.605	0.4	0.14	0.066	28
11	**GGCGCGCAGT**	-359	CGCGTAGT	-385	-335	0.936	0.444	0.373	0.119	5
			TAGGCGC	-405	-355	0.746	0.423	0.191	0.132	16
			TAGGCG	-405	-355	0.612	0.413	0.129	0.071	27
12	**CAGCTGTTCCgc**	-325	AACTGTAC	-360	-310	0.787	0.455	0.253	0.079	11
			CAACTGT	-365	-315	0.689	0.418	0.167	0.104	21
13	**TGTTTACATCc***	-173	GTTTACA	-195	-145	0.686	0.471	0.139	0.076	22
			GTTAAC	-200	-145	0.621	0.427	0.121	0.073	25
			TTACTT	-200	-145	0.602	0.419	0.082	0.101	29
14	**ccaCCCTCCCC***	-105	ACCCCCCC	-125	-75	1.602	0.364	0.265	0.973	1
			CCCCCC	-125	-75	1.511	0.366	0.145	1.00	2
			CCCCCCC	-125	-75	1.429	0.334	0.095	1.00	3
			CCTCCCTA	-130	-80	1.043	0.432	0.244	0.366	4
			TCCCCTGC	-150	-70	0.863	0.401	0.255	0.208	7
			CCTTCC	-130	-80	0.834	0.444	0.136	0.254	8
			GGCTCCCC	-130	-75	0.793	0.362	0.128	0.302	10
			CTCCAC	-145	-95	0.778	0.411	0.113	0.254	12
			CCCCAA	-125	-70	0.745	0.407	0.13	0.207	17
			TTCCCCA	-130	-70	0.696	0.393	0.104	0.199	20
			TCCCCT	-135	-85	0.625	0.376	0.066	0.184	24

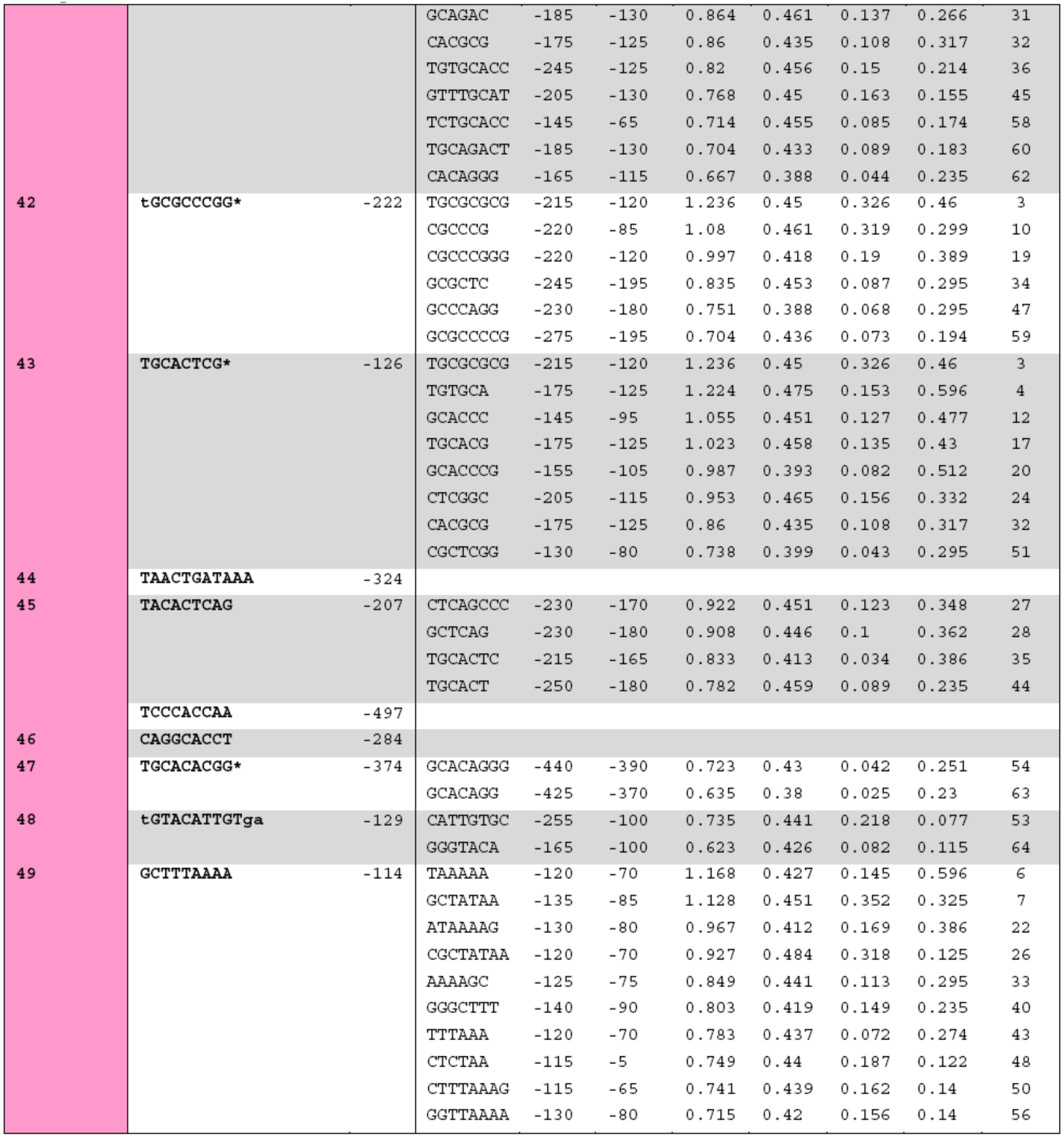
15	**AGCAGAGGGCG***	-69	AGAAAAGG	-75	-25	0.904	0.397	0.187	0.319	6
			AGAAGA	-80	-30	0.75	0.387	0.127	0.236	14
			ACGGCGT	-110	-60	0.748	0.493	0.23	0.026	15
			AACAGATG	-80	-30	0.721	0.414	0.187	0.119	19
16	**GGCGTGGG***	-62	GGGGGGGG	-105	-55	0.776	0.351	0.011	0.414	13
			ACGGCGT	-110	-60	0.748	0.493	0.23	0.026	15
17	**ATCTCCGCCCAcc**	-26	AGATCCGC	-75	-25	0.731	0.396	0.237	0.099	18
			CCGACC	-60	-5	0.615	0.443	0.071	0.101	26
gh1	**GGGAGGAG***	-198	GGAGGT	-205	-135	0.905	0.422	0.255	0.228	17
			GGGAAGGG	-270	-205	0.855	0.381	0.107	0.367	20
18			AGAGGAG	-235	-185	0.79	0.37	0.043	0.376	25
			GACGAG	-245	-195	0.727	0.416	0.144	0.167	43
			GGGACG	-265	-195	0.703	0.43	0.149	0.124	51
			GTAGGA	-215	-160	0.685	0.438	0.143	0.104	54
19	**ATTATCCAT***	-183	ATTAGC	-190	-140	0.826	0.459	0.26	0.107	24
			CAATTA	-210	-160	0.786	0.415	0.182	0.188	26
			ATACAT	-190	-140	0.738	0.432	0.053	0.253	38
			TTATCCC	-250	-170	0.726	0.455	0.181	0.09	44
			ATTATT	-235	-155	0.723	0.399	0.17	0.154	45
			TGTCCAT	-220	-170	0.52	0.387	0.018	0.115	59
20	**TTAGCACAA**	-174	GTAGCAC	-200	-135	1.023	0.458	0.435	0.13	10
			TAAACACA	-205	-155	0.892	0.397	0.063	0.432	18
			ATTAGC	-190	-140	0.826	0.459	0.26	0.107	24
			AACACA	-200	-150	0.747	0.418	0.031	0.298	36
			TAGCAC	-185	-125	0.743	0.481	0.191	0.071	37
			CACAAA	-195	-135	0.717	0.388	0.072	0.257	46
21	**GTCAGTGG***	-162	TGAGTG	-230	-180	0.83	0.431	0.147	0.253	23
			ATGAGTGG	-230	-180	0.785	0.401	0.133	0.251	27
			GGCAGTG	-170	-120	0.735	0.422	0.115	0.197	40
			CGGTGG	-170	-120	0.706	0.457	0.082	0.167	50
			CCGTCAG	-190	-130	0.64	0.401	0.124	0.115	57
22	**gcATAAATGTA***	-146	CATGTAT	-165	-95	1.153	0.364	0.192	0.597	7
			TATAAAT	-190	-140	0.984	0.378	0.043	0.563	11
			ATACATGT	-155	-105	0.983	0.394	0.129	0.459	12
			ACATGT	-155	-105	0.98	0.414	0.175	0.391	13
			ATGTAT	-165	-95	0.974	0.436	0.246	0.292	14
			GTAAAT	-190	-140	0.845	0.436	0.156	0.253	21
			AATCTA	-145	-95	0.833	0.417	0.228	0.188	22
			AAGGTA	-160	-90	0.774	0.412	0.18	0.182	29
			TATAAA	-190	-140	0.771	0.444	0.029	0.298	30
			TTAACGTA	-220	-135	0.765	0.455	0.208	0.102	33
			ATATAAAT	-190	-140	0.765	0.383	0.029	0.353	32
			TAGATGT	-175	-125	0.764	0.411	0.113	0.241	34
			TTAATG	-230	-135	0.76	0.411	0.154	0.195	35
			ATACAT	-190	-140	0.738	0.432	0.053	0.253	38
			ATGTATTT	-155	-105	0.737	0.391	0.07	0.276	39
			GAAAGGTA	-165	-105	0.691	0.39	0.099	0.202	53
			TAATTTTA	-155	-100	0.617	0.37	0.034	0.213	58
23	**GAAACAGGT**	-131	AAAAAGGG	-115	-65	1.559	0.366	0.194	1.00	1
			AAACAGG	-130	-80	1.195	0.392	0.097	0.707	6
			ATACATGT	-155	-105	0.983	0.394	0.129	0.459	12
			ACATGT	-155	-105	0.98	0.414	0.175	0.391	13
			AAACAT	-120	-65	0.863	0.422	0.06	0.381	19
			AGGTTT	-150	-100	0.734	0.459	0.087	0.188	41

			GAGAAA	-150	-95	0.729	0.41	0.084	0.235	42
			CACAGGTG	-180	-120	0.707	0.399	0.106	0.202	49
			TGAAACAG	-135	-85	0.676	0.434	0.04	0.202	55
24	**cagggTATAAAAAGggc***	-97	AAAAAGGG	-115	-65	1.559	0.366	0.194	1.00	1
			ATATAAA	-100	-50	1.491	0.372	0.119	1.00	2
			AAAAAA	-115	-65	1.467	0.38	0.087	1.00	3
			TTAAAAAA	-120	-70	1.439	0.356	0.083	1.00	4
			ATATAAAC	-120	-70	1.228	0.394	0.124	0.709	5
			TACAAA	-100	-50	1.096	0.454	0.179	0.463	8
			ATATAA	-125	-75	1.083	0.429	0.166	0.488	9
			ATAGAGAG	-150	-100	0.924	0.386	0.107	0.432	15
			TTTAAA	-130	-80	0.92	0.416	0.065	0.439	16
25	**TCATGTTTt**	-138	CATGTAT	-165	-95	1.153	0.364	0.192	0.597	7
			ACATGT	-155	-105	0.98	0.414	0.175	0.391	13
			ATGTAT	-165	-95	0.974	0.436	0.246	0.292	14
			CTCCTGT	-155	-100	0.777	0.392	0.135	0.249	28
			TGTTTA	-175	-120	0.766	0.406	0.085	0.275	31
			AGGTTT	-150	-100	0.734	0.459	0.087	0.188	41
			TCCTGT	-165	-115	0.717	0.441	0.13	0.147	47
			CTCCTGTT	-165	-115	0.711	0.406	0.104	0.202	48
			TCATGA	-140	-90	0.7	0.43	0.082	0.188	52
			CATGTTGG	-165	-115	0.668	0.403	0.086	0.178	56
histoneh1	**CAATCACCAC***	-107	ATCACCA	-135	-85	1.23	0.467	0.5	0.263	2
			ACCACGCA	-135	-85	1.177	0.33	0.253	0.594	3
26			CCACGC	-105	-40	1.048	0.451	0.297	0.3	4
			CCAATCA	-140	-90	1.027	0.432	0.214	0.381	5
			ATCAACCC	-110	-55	0.888	0.363	0.208	0.317	7
			ATCAAC	-110	-55	0.884	0.419	0.2	0.265	8
			CTATCA	-120	-70	0.861	0.434	0.185	0.242	9
			AATGACCG	-150	-85	0.837	0.5	0.243	0.094	10
			AACCAATC	-140	-90	0.72	0.403	0.096	0.22	12
27	**gAAACAAAAGTtt**	-427	AAGAAAA	-435	-350	1.395	0.363	0.203	0.829	1
			AAGGAAAA	-465	-415	1	0.34	0.066	0.594	6
			AAAAAA	-485	-415	0.735	0.356	0.019	0.361	11
			GAACAACA	-430	-350	0.7	0.398	0.082	0.22	13
insulin	**gttAAGACTCTAAtgacc***	-223	ACTCTAA	-245	-195	0.965	0.432	0.163	0.37	5
			GACTCG	-250	-185	0.867	0.457	0.331	0.08	11
28			ACTCTA	-235	-185	0.857	0.45	0.203	0.204	12
			TAAGACTC	-240	-190	0.828	0.448	0.247	0.133	18
			AAGACT	-240	-190	0.812	0.456	0.203	0.153	21
			GTCTAA	-225	-175	0.766	0.389	0.12	0.257	27
			CCCTAA	-235	-185	0.723	0.393	0.072	0.257	32
			AACCCTAA	-235	-185	0.627	0.335	0.121	0.171	40
29	**tcagcccccaGCCATCTGCC***	-122	CCATCTG	-120	-70	1.161	0.392	0.141	0.627	2
			GCCACC	-120	-70	0.923	0.394	0.076	0.453	7
			CAGCAGCC	-155	-105	0.923	0.31	0.069	0.544	6
			TCGGCC	-125	-75	0.907	0.396	0.142	0.368	9
			TGCCGA	-190	-90	0.881	0.428	0.349	0.104	10
			CAACTGCA	-135	-65	0.856	0.343	0.231	0.282	13
			CATCAG	-155	-95	0.836	0.386	0.073	0.377	14
			CCTCGGCC	-130	-80	0.835	0.319	0.122	0.394	15

			ATCTTC	-130	-80	0.827	0.42	0.203	0.204	19
			ATCTTCC	-130	-80	0.816	0.432	0.163	0.221	20
			CCTCTG	-130	-80	0.792	0.369	0.055	0.368	24
			GCCCTCAG	-125	-75	0.791	0.305	0.071	0.415	25
			TCTACC	-125	-75	0.756	0.426	0.177	0.153	28
			GCATCT	-130	-80	0.744	0.406	0.134	0.204	30
			CTTCTACC	-130	-75	0.707	0.414	0.178	0.115	33
			AACTGC	-145	-95	0.704	0.387	0.059	0.257	34
			GACATTTG	-130	-80	0.672	0.466	0.127	0.08	37
			ATCTGCCG	-170	-95	0.645	0.35	0.155	0.14	39
			CCCAGCTG	-120	-70	0.622	0.323	0.029	0.27	42
			GGCCATCT	-130	-80	0.583	0.346	0.065	0.171	45
30	**CTATAAAGcc***	-32	TATAACG	-70	-20	1.173	0.482	0.371	0.32	1
			TATAAC	-70	-20	0.833	0.486	0.243	0.104	16
			CCTATA	-75	-25	0.805	0.458	0.243	0.104	23
			GGGCTATA	-80	-25	0.748	0.375	0.225	0.149	29
			TATAAAGC	-70	-20	0.73	0.447	0.203	0.08	31
			TAAAGG	-85	-20	0.698	0.386	0.12	0.192	36
			ACTCTAA	-70	-20	0.649	0.432	0.044	0.173	5
			CCAGAAAG	-60	-10	0.647	0.332	0.144	0.171	38
			CTATCAAT	-75	-25	0.608	0.418	0.111	0.08	43
			CTTTGAAG	-75	-25	0.604	0.37	0.101	0.133	44
31	**GGGAAATG***	-145	CCGGAAA	-155	-105	1.13	0.409	0.5	0.221	3
			GAAATTG	-155	-105	1.022	0.399	0.203	0.421	4
			GGGAAGT	-165	-115	0.909	0.397	0.142	0.37	8
			AGGAAA	-175	-125	0.828	0.359	0.073	0.396	17
			CGGAAATT	-155	-105	0.811	0.396	0.224	0.19	22
			GAAAAT	-155	-105	0.766	0.378	0.103	0.285	26
			GGGAAT	-165	-115	0.698	0.4	0.093	0.204	35
			GAAAATGC	-155	-105	0.626	0.372	0.102	0.152	41
interleukin3 **32**	**TTGAGTACTagaaagt**	-228	TTGAATA	-230	-165	1.319	0.416	0.268	0.635	2
			GAGTAAT	-230	-180	1.155	0.414	0.295	0.447	6
			TAAGTAAT	-230	-180	0.913	0.352	0.258	0.304	13
			TTGAAT	-230	-165	0.866	0.426	0.268	0.173	17
			TAAGTA	-260	-160	0.798	0.365	0.29	0.143	25
			TTTTGA	-235	-185	0.752	0.4	0.242	0.11	33
			TTTTGAGT	-270	-220	0.727	0.401	0.202	0.124	39
			GAGTAA	-230	-170	0.704	0.395	0.135	0.174	42
			GAATAC	-225	-170	0.647	0.361	0.111	0.176	51
			TGAGTC	-230	-165	0.627	0.41	0.131	0.087	57
33	**GATGAATAATt***	-208	TTGAATA	-230	-165	1.319	0.416	0.268	0.635	2
			GAGTAAT	-230	-180	1.155	0.414	0.295	0.447	6
			TAAGTAAT	-230	-180	0.913	0.352	0.258	0.304	13
			TTGAAT	-230	-165	0.866	0.426	0.268	0.173	17
			ATAAAT	-210	-160	0.777	0.365	0.138	0.274	28
			TCGATG	-210	-135	0.757	0.397	0.235	0.125	31
			ATGGAT	-215	-150	0.733	0.402	0.176	0.155	38
			TCGATGAA	-215	-165	0.718	0.343	0.202	0.173	41
			GAGTAA	-230	-170	0.704	0.395	0.135	0.174	42
			GAATAC	-225	-170	0.647	0.361	0.111	0.176	51
34	**GTCTGTGGTTTtCTATGGAGGTTCCATGTCAGATAAAG***	-195	TTCTATG	-185	-135	1.475	0.413	0.342	0.72	1
			GGTTTTC	-200	-150	1.281	0.467	0.5	0.314	4
			TAAAGAT	-175	-125	1.262	0.448	0.5	0.314	5

			GGAGGTG	-190	-140	1.135	0.479	0.342	0.314	7
			ACGTCTG	-200	-145	1.113	0.435	0.34	0.338	8
			TCAGGTA	-195	-130	1.034	0.452	0.309	0.274	10
			TGGTTTTC	-220	-150	0.972	0.49	0.426	0.056	11
			TTCTAT	-185	-135	0.914	0.417	0.295	0.202	12
			ATGGACG	-215	-140	0.875	0.424	0.181	0.27	15
			GGTTTT	-190	-105	0.868	0.431	0.384	0.053	16
			TGTGGTT	-215	-160	0.855	0.471	0.164	0.22	19
			CCATGTT	-215	-135	0.854	0.406	0.118	0.33	20
			TTCTATGG	-195	-135	0.836	0.353	0.202	0.281	21
			GGACGTTT	-190	-140	0.835	0.5	0.258	0.078	22
			GATAAGG	-175	-125	0.825	0.419	0.092	0.314	23
			ATAGTGGT	-190	-135	0.813	0.371	0.268	0.173	24
			AGTTCCA	-185	-135	0.79	0.398	0.078	0.314	27
			ATAAAT	-210	-160	0.777	0.365	0.138	0.274	28
			TAAAGATC	-175	-125	0.765	0.429	0.258	0.078	30
			GTGGTT	-195	-140	0.765	0.422	0.168	0.176	29
			TCGATG	-210	-135	0.757	0.397	0.235	0.125	31
			GACGTT	-190	-140	0.754	0.485	0.202	0.067	32
			CTATGG	-195	-145	0.745	0.416	0.15	0.178	34
			GTTTGCTA	-200	-150	0.738	0.458	0.202	0.078	36
			ATGTCG	-205	-130	0.735	0.454	0.227	0.054	37
			ATGGAT	-215	-150	0.733	0.402	0.176	0.155	38
			GTCTTT	-265	-160	0.718	0.401	0.256	0.062	40
			ATGAAG	-180	-130	0.692	0.354	0.088	0.25	44
			TAAAGC	-175	-125	0.682	0.413	0.202	0.067	45
			CTCTGGTG	-195	-145	0.669	0.335	0.135	0.199	47
			TCCATC	-185	-135	0.657	0.401	0.146	0.11	48
			AAGGTTCT	-180	-120	0.656	0.339	0.122	0.195	49
			GGTCCATC	-180	-130	0.646	0.356	0.141	0.149	52
			GTCTGC	-195	-140	0.637	0.381	0.101	0.155	53
			CATATAAG	-175	-125	0.636	0.339	0.072	0.224	54
			AAGTCT	-220	-160	0.631	0.374	0.102	0.155	55
			CGATAA	-175	-115	0.621	0.465	0.092	0.064	58
			TAAGGCGG	-185	-125	0.62	0.352	0.115	0.153	59
			GATAAGGA	-180	-125	0.614	0.329	0.112	0.173	62
			GTTTCC	-190	-135	0.601	0.365	0.06	0.176	64
			GGAGATAC	-190	-140	0.583	0.413	0.092	0.078	66
			GGTTGTGG	-205	-155	0.58	0.41	0.092	0.078	67
			CTCAAATA	-175	-125	0.577	0.407	0.092	0.078	69
			AACCATGT	-185	-135	0.577	0.395	0.082	0.101	68
			TGGAGGTG	-190	-140	0.561	0.405	0.078	0.078	70
			ATGTGCCA	-190	-140	0.543	0.398	0.067	0.078	71
35	**TCTTCAGAGc**	-56	TCTTGAG	-80	-30	0.74	0.445	0.11	0.185	35
			ATTCAGAA	-75	-25	0.701	0.405	0.171	0.124	43
			TTCAGC	-80	-30	0.679	0.449	0.162	0.067	46
			GTCAGA	-75	-25	0.631	0.37	0.035	0.226	56
			CTTCTG	-65	-15	0.618	0.37	0.07	0.178	61
			GCTTCTGA	-80	-30	0.586	0.323	0.089	0.173	65
36	**AGGACCAG**	-40	GGACCTG	-65	-15	1.281	0.467	0.5	0.314	3
			GACAAGA	-60	-10	1.056	0.408	0.202	0.447	9
			CAGGAC	-60	-10	0.884	0.37	0.139	0.375	14
			AGAACC	-60	-10	0.864	0.363	0.151	0.349	18
			CAGAACCA	-60	-10	0.793	0.318	0.117	0.358	26
			AGGACTAG	-60	-10	0.651	0.354	0.072	0.224	50
			AGGGCCA	-65	-15	0.618	0.404	0.029	0.185	60

			GACCAA	-55	-5	0.613	0.371	0.11	0.132	63
metallothionein	**GCTATAAAc***	-103	TAAAAA	-120	-70	1.168	0.427	0.145	0.596	6
			GCTATAA	-135	-85	1.128	0.451	0.352	0.325	7
37			TATAAG	-130	-80	1.033	0.467	0.313	0.253	15
			CTAAAAAC	-120	-70	0.979	0.429	0.231	0.319	21
			ATAAAAG	-130	-80	0.967	0.412	0.169	0.386	22
			CGCTATAA	-120	-70	0.927	0.484	0.318	0.125	26
			GGTATA	-120	-70	0.811	0.452	0.26	0.098	39
			GCGCTA	-100	-50	0.795	0.459	0.202	0.134	42
			TTTAAA	-120	-70	0.783	0.437	0.072	0.274	43
			GATAAA	-130	-80	0.756	0.445	0.099	0.211	46
			CTCTAA	-115	-5	0.749	0.44	0.187	0.122	48
			CTTTAAAG	-115	-65	0.741	0.439	0.162	0.14	50
38	**CATGCGCAGg**	-143	TGTGCA	-175	-125	1.224	0.475	0.153	0.596	4
			GTGCGC	-245	-105	1.187	0.46	0.352	0.375	5
			CGTGCGCA	-245	-135	1.113	0.493	0.331	0.288	8
			GCGCAG	-175	-125	1.043	0.451	0.138	0.453	13
			CACGCG	-175	-125	0.86	0.435	0.108	0.317	32
			CACGCGGA	-175	-120	0.718	0.44	0.123	0.155	55
39	**cCGTGTGCAg***	-239	GTGCGC	-245	-105	1.187	0.46	0.352	0.375	5
			CGTGCGCA	-245	-135	1.113	0.493	0.331	0.288	8
			GTGCGCA	-245	-190	1.108	0.408	0.105	0.596	9
			GAGTGC	-225	-145	1.032	0.455	0.295	0.282	16
			CGCGTGCT	-230	-145	0.959	0.45	0.269	0.239	23
			TGTGCA	-250	-200	0.871	0.475	0.057	0.339	30
			GGTGTG	-305	-140	0.871	0.447	0.233	0.191	29
			TGTGCACC	-245	-125	0.82	0.456	0.15	0.214	36
			GTGCGCAG	-300	-195	0.819	0.439	0.11	0.27	38
			CGTATG	-225	-165	0.819	0.5	0.261	0.059	37
			GTGTAC	-250	-190	0.798	0.455	0.136	0.208	41
			CGTATGC	-250	-200	0.736	0.492	0.095	0.15	52
			AGGGTGCA	-250	-200	0.714	0.46	0.052	0.202	57
			GGCGTGTG	-295	-210	0.684	0.449	0.088	0.146	61
40	**CGTGTGCAggc***	-156	CGCGTG	-195	-145	1.447	0.446	0.332	0.67	1
			CGCGTGC	-195	-145	1.283	0.422	0.19	0.672	2
			TGTGCA	-175	-125	1.224	0.475	0.153	0.596	30
			GTGCGC	-245	-105	1.187	0.46	0.352	0.375	5
			CGTGCGCA	-245	-135	1.113	0.493	0.331	0.288	8
			GAGTGC	-225	-145	1.032	0.455	0.295	0.282	16
			GGGTGCAG	-175	-125	1.013	0.456	0.113	0.443	18
			CGCGTGCT	-230	-145	0.959	0.45	0.269	0.239	23
			GGTGTG	-305	-140	0.871	0.447	0.233	0.191	29
			TGTGCACC	-245	-125	0.82	0.456	0.15	0.214	36
			CGTATG	-225	-165	0.819	0.5	0.261	0.059	37
			GTTTGCAT	-205	-130	0.768	0.45	0.163	0.155	45
			CGAGTACA	-195	-145	0.743	0.472	0.146	0.125	49
41	**TTTGCACACG***	-142	TGCGCGCG	-215	-120	1.236	0.45	0.326	0.46	3
			TGTGCA	-175	-125	1.224	0.475	0.153	0.596	4
			TTTGCGC	-155	-105	1.077	0.429	0.2	0.449	11
			GCACCC	-145	-95	1.055	0.451	0.127	0.477	12
			TTTTGCGC	-145	-95	1.034	0.471	0.295	0.268	14
			TGCACG	-175	-125	1.023	0.458	0.135	0.43	17
			GCACCCG	-155	-105	0.987	0.393	0.082	0.512	20
			TTTGCG	-155	-105	0.948	0.463	0.233	0.253	25



			GCAGAC	-185	-130	0.864	0.461	0.137	0.266	31
			CACGCG	-175	-125	0.86	0.435	0.108	0.317	32
			TGTGCACC	-245	-125	0.82	0.456	0.15	0.214	36
			GTTTGCAT	-205	-130	0.768	0.45	0.163	0.155	45
			TCTGCACC	-145	-65	0.714	0.455	0.085	0.174	58
			TGCAGACT	-185	-130	0.704	0.433	0.089	0.183	60
			CACAGGG	-165	-115	0.667	0.388	0.044	0.235	62
42	**tGCGCCCGG***	-222	TGCGCGCG	-215	-120	1.236	0.45	0.326	0.46	3
			CGCCCG	-220	-85	1.08	0.461	0.319	0.299	10
			CGCCCGGG	-220	-120	0.997	0.418	0.19	0.389	19
			GCGCTC	-245	-195	0.835	0.453	0.087	0.295	34
			GCCCAGG	-230	-180	0.751	0.388	0.068	0.295	47
			GCGCCCCG	-275	-195	0.704	0.436	0.073	0.194	59
43	**TGCACTCG***	-126	TGCGCGCG	-215	-120	1.236	0.45	0.326	0.46	3
			TGTGCA	-175	-125	1.224	0.475	0.153	0.596	4
			GCACCC	-145	-95	1.055	0.451	0.127	0.477	12
			TGCACG	-175	-125	1.023	0.458	0.135	0.43	17
			GCACCCG	-155	-105	0.987	0.393	0.082	0.512	20
			CTCGGC	-205	-115	0.953	0.465	0.156	0.332	24
			CACGCG	-175	-125	0.86	0.435	0.108	0.317	32
			CGCTCGG	-130	-80	0.738	0.399	0.043	0.295	51
44	**TAACTGATAAA**	-324								
45	**TACACTCAG**	-207	CTCAGCCC	-230	-170	0.922	0.451	0.123	0.348	27
			GCTCAG	-230	-180	0.908	0.446	0.1	0.362	28
			TGCACTC	-215	-165	0.833	0.413	0.034	0.386	35
			TGCACT	-250	-180	0.782	0.459	0.089	0.235	44
	TCCCACCAA	-497								
46	**CAGGCACCT**	-284								
47	**TGCACACGG***	-374	GCACAGGG	-440	-390	0.723	0.43	0.042	0.251	54
			GCACAGG	-425	-370	0.635	0.38	0.025	0.23	63
48	**tGTACATTGTga**	-129	CATTGTGC	-255	-100	0.735	0.441	0.218	0.077	53
			GGGTACA	-165	-100	0.623	0.426	0.082	0.115	64
49	**GCTTTAAAA**	-114	TAAAAA	-120	-70	1.168	0.427	0.145	0.596	6
			GCTATAA	-135	-85	1.128	0.451	0.352	0.325	7
			ATAAAAG	-130	-80	0.967	0.412	0.169	0.386	22
			CGCTATAA	-120	-70	0.927	0.484	0.318	0.125	26
			AAAAGC	-125	-75	0.849	0.441	0.113	0.295	33
			GGGCTTT	-140	-90	0.803	0.419	0.149	0.235	40
			TTTAAA	-120	-70	0.783	0.437	0.072	0.274	43
			CTCTAA	-115	-5	0.749	0.44	0.187	0.122	48
			CTTTAAAG	-115	-65	0.741	0.439	0.162	0.14	50
			GGTTAAAA	-130	-80	0.715	0.42	0.156	0.14	56

Supplementary Figure 2. Comparison of LocalMotif's predictions with published TFBS annotations in long upstream regulatory sequences. The species whose sequence annotations are derived from the literature is highlighted in boldface and the reference is provided alongside. Each published TFBS is shown with its matching LocalMotif prediction, and the matching subsequence within the TFBS is highlighted in red color. Only the top 25 predictions made by LocalMotif were considered. "NP" indicates that none of the top 25 LocalMotif predictions match the TFBS.

Gene	From literature			LocalMotif Predictions		
	Pos	Binding site	Transcription Factor	Predicted Motif	Position interval	Rank
CRHR1 **homo sapiens [Parham et al. (2004)]** mus musculus rattus norvegicus gallus gallus	-44	AG**GACCCGG**GC	SP1	GACCGG	[-55,-35]	2
	-56	TGG**GATGTCC**	NF-KAPPA	GATGTCG	[-70,-50]	8
	-282	**GGGGAGG**TG	SP1	GGGGGGG	[-290,-270]	21
	-305	GG**CGAGGA**GCGGC	SP1	CGAGGA	[-325,-300]	5
	-320	**GGGGCGG**GGA	EGR-1/EGR-2	GGGGGGG	[-325,-305]	6
	-355	**GAGGGG**GAGGAAG	SP1	GGAGGGG	[-370,-345]	7
	-374	GGGGAGC**GGAGGGG**	SP1	GGAGGGG	[-370,-345]	7
	-401	GG**GGCGAG**GCGCGGAGG	SP1	GGCGAG	[-410,-360]	18
	-417	GCTG**GGAGGG**	SP1	GGAGCG	[-475,-335]	12
	-445	GGGGAGGGAA	SP1	NP	-	-
	-458	CG**GGCCGG**GGG	SP1	GACCGG	[-475,-425]	20
	-477	G**GCGGCG**GGACA	SP1	GCGGCG	[-480,-445]	15
	-560	CTCCCCGGGCT**GCGGCG**G	AP2	GCGGCC	[-555,-505]	11
	-720	GGA**CCGCCCTG**TTCC	SP1/NF-KB/EBP-1	TCGCCCTA	[-705,-685]	17
	-923	A**TGAATAA**GG	PIT-1A	TGAATGA	[-950,-930]	1
	-962	CAGTTTGTAA	PR	NP	-	-
	-997	CCAGCCTCTTG	SP1	NP	-	-
	-1027	GGGGCTCCCAGG	AP-2/SP1	NP	-	-
	-1053	**GGGCAC**CGCCG	SP1	GCGCAC	[-1105,-1040]	22
	-1158	CCTCCCCA**CGCCCT**GCCCGCGGGC	SP1/ETF/AP-2A	CGCCGT	[-1180,-1095]	14
	-1177	CCGGGAGGGGGCGC	SP1/ETF/KROX-20	NP	-	-
	-1295	TCTGT**TCATCT**	GATA1	TTATCT	[-1335,-1285]	19
	-1484	AGGGCAGGTG	RXR	NP	-	-
	-1497	GGGCACAGGG	RXR	NP	-	-
	-1532	AGAGGGCAGGAGGGAGGAG	SP1	NP	-	-
	-1702	CTGTGAGCTGG	ER	NP	-	-
	-1720	GGCCCAGCCCTC	SP1	NP	-	-
	-1735	CCCAGGCCCCTTT	SP1	NP	-	-
	-1754	CCCCTCCCCA	SP1	NP	-	-
	-1841	TTTTGCAAGACT	SP1/NF-1	NP	-	-
	-1931	GCT**TAGCATG**T	OCT1	TAGTATG	[-1930,-1845]	24
	-2050	**AGTTTAT**ACAGCTTGTAAG	GR	AAGTTTAT	[-2070,-2050]	16
	-2089	ATAG**ATGAGA**	GATA3	ATTAGA	[-2100,-2080]	10
	-2290	AGGGTGGGACC	SP1	NP	-	-
	-2368	CTCTCCTCT	SP1	NP	-	-
	-2412	CTTGGCTGGG	NF-1	NP	-	-
	-2430	CCAGGGAGGGA	YY1	NP	-	-
HHEX homo sapiens **mus musculus [Myint et al. (1999)]** rattus norvegicus gallus gallus sus scrofa	-64	C**AAATAAAT**	TATA_BOX	ACATAAAT	[-85,-65]	4
	-110	GCC**CCACCCC**GCGG	SP1/AP-2	CCACACC	[-125,-105]	12
	-148	**GGCCG**CAGGGC	AP-2	GGGCCG	[-150,-125]	8
	-172	G**GCGAAT**CT	CCAAT_BOX	GCGAATC	[-175,-150]	2
	-185	AGT**GGGGGG**CGGA	MZF1/SP1	GGGGGG	[-200,-180]	1
	-196	GCG**CGGGGG**	AP-2	GGGGGG	[-200,-180]	7
	-214	GCC**GGTGGG**GGCGGATC	AP-2/MZF1/SP1	GGTCGG	[-215,-185]	11
	-228	G**GGGCGGG**AG	SP1	GCGGGC	[-230,-185]	19
	-297	CTGGG**GGCGCC**	SP1	GGCGCC	[-305,-225]	10
	-370	**TCCGCG**CCCCGCGGCG	SP1/AP-2	TCGGCG	[-365,-345]	9
	-459	GC**CGGCGG**CC	AP-2	CGCCGG	[-465,-445]	6
	-530	T**CCCCCG**TT	MZF1	CCACCG	[-575,-500]	17
	-574	**GCCAACG**GCT	AP-2	GCCAACG	[-575,-500]	18
HK2 homo sapiens **mus musculus [Heikkinen et al. (2000)]** rattus norvegicus	-68	**GGGCG**GCC	SP1	TGGGCG	[-75,-55]	14
	-81	A**CGTCACT**G	CREB	CGTCAAT	[-90,-40]	2
	-98	**AGCCAAT**GAG	CAAT	AGCCAAT	[-140,-85]	7
	-140	**CCGCG**GGCGG	AP-2/SP1	CCCGCG	[-125,-105]	1
	-154	T**GATTGG**CT	AP-1/NF-1	GATCGG	[-175,-145]	18
	-178	**GCCGCG**CCCG	SP1	CCCGCG	[-235,-185]	6
	-205	**CCCGCC**GCC	SP1	CCCGCG	[-235,-185]	10
	-233	C**TCCGCC**CCT	SP1	TCCGCT	[-240,-220]	3
	-333	G**CGCCCC**CCACCC	SP1/PUF	CGCCCG	[-365,-195]	20
	-370	TTTCCAGTC	C/EBP	NP	-	-

gallus gallus danio rerio	-429	CTCACATG	USF	CTCTCAT	[-480,-420]	19
	-439	GGGCAGTGG	SP1	NP	-	-
	-525	ACTTTTGTAAT	C/EBP	NP	-	-
	-592	TCCCCAAT	AP-2	NP	-	-
	-631	TGACCTTGGG	PPAR	NP	-	-
LTF homo sapiens **mus musculus [Liu, Y. H. and Teng (1991)]** rattus norvegicus gallus gallus canis familiaris						
	106	GGCGGGAGGT	SP1	GCGCGAG	[85,105]	3
	-32	AGATAAAGGG	TATA	ATAAAG	[-80,-30]	7
	-73	GGGCAATAGGG	CAAT	GGCAATAG	[-110,-55]	2
	-102	GGGCAATGGG	CAAT	GGCAATAG	[-110,-55]	2
	-157	TTCCTCCTTC	putative	NP	-	-
	-339	GGTCAAGGTAAC	ERE	AAGGTAAC	[-350,-330]	10
	-381	GAGGAAGGGG	putative	GCAAGG	[-495,0]	12
	-514	TGGACCCCAC	AP-2	GACCGCA	[-540,-515]	17
	-533	GGCGGGTTT	SP1	GGCGGGTT	[-550,-530]	16
	-617	TTCCTCGCT	putative	NP	-	-
	-651	AAAAGGAGC	putative	AAAGGA	[-695,-670]	1
	-667	GAGGAAGGAA	putative	AGGAAA	[-715,-650]	19
	-690	GAGGAAGAGGAA	putative	AGGAAA	[-715,-650]	19
	-907	CCGCCCGGC	SP1	ACCGCCCG	[-925,-905]	21
PCMT1 **homo sapiens [DeVry et al. (1996)]** mus musculus rattus norvegicus gallus gallus						
	54	GCTCCGAGTGT	MED-1	GCTCCGA	[20,40]	6
	-28	GGGCGGTGAC	SP1	GCGGTGA	[-85,-25]	9
	-111	GCCGCGGGGGA	AP-2	GCGGGG	[-145,-100]	14
	-134	GCGGCGTCACA	ARE	GCGGCG	[-135,-55]	11
	-172	CCCCGCCCTCGGCCC	ETF	CCCCCC	[-175,-150]	1
	-205	GCCACAGGGGCGGGCGG	AP-2/SP1	GGGGCGG	[-225,-170]	2
	-249	CTGACTCAGCC	AP-1	CGCTGACT	[-270,-250]	14
	-270	CACGCAGCAGC	XRE	CGCAGGA	[-290,-270]	7
	-313	GCCGCAGGGC	AP-2	GCCGGA	[-325,-290]	12
	-361	TGACCGGAGA	ERE	GACCGC	[-400,-300]	17
	-387	CCCCGCCATCCCGCC	ETF/SP1	CCGCCG	[-415,-395]	3
	-435	TGACCCAGCGA	ERE	GACCCG	[-455,-405]	5
	-615	GCCAGAGGCCG	AP-2	NP	-	-
	-633	TGACGTGCTT	CREB	GACGTGC	[-650,-630]	21
	-719	CCCCAACCCCCACCCC	ETF	NP	-	-
	-1235	GGTCAGGAGA	ERE	GATCAG	[-1255,-1230]	8
	-1248	GGGCGGATC	SP1	NP	-	-
	-1457	GGTCACATA	ERE	NP	-	-
	-1645	TGCGTGCCTG	XRE	TGCGTG	[-1645,-1595]	15
	-1706	GGTCAACAT	ERE	AGGTCA	[-1755,-1705]	19
	-1729	GGTCAGGAGT	ERE	AGGTCA	[-1755,-1705]	19
SLC29A1 **homo sapiens [Abdulla and Coe (2007)]** mus musculus rattus norvegicus gallus gallus canis familiaris						
	-99	TGGCAGGGCC	SP1	TCAGGGC	[-100,-80]	11
	-186	GGGTGCAGAGG	GATA-1	GGGGCA	[-240,-190]	25
	-213	GGTCAAGTTGAGT	ERE	GTCAAGTT	[-230,-210]	4
	-263	GGCAGGGGGG	SP1	GGCAGG	[-290,-270]	2
	-303	GTGGCAGCGGC	SP1	GCAGCC	[-305,-255]	9
	-325	CGGCTGCTGG	SP1	CGGCGGC	[-325,-305]	5
	-340	GCTTATAAACT	TATA_BOX	TTATAAAC	[-340,-320]	1
	-356	ACCCTCCTGTT	SP1	CTCCTC	[-380,-350]	19
	-373	CTTCCCTCCTGC	SP1	TTCCCT	[-400,-380]	15
	-401	TCCCTCCCTCCCATC	putative	TTCCCT	[-400,-380]	15
	-419	GCCTTCTTTGA	SP1	TTTCTT	[-425,-405]	10
	-446	TCCCTGACCCC	SRF	TCCCTC	[-510,-445]	7
	-470	TCCCTCCTTCCC	SP1	TCCCTC	[-510,-445]	7
	-548	GCAGCTGCTG	MYOD	NP	-	-
	-592	CCTGGGGAGC	AP-2	TGGGGC	[-605,-585]	17
	-635	GGCTCCCCAG	AP-2	NP	-	-
	-743	AGCCCCTGGG	SP1	NP	-	-
	-883	GGGAAGCGGA	SP1	GAAACGG	[-900,-880]	3
	-924	TCCACCCCTCC	SP1	NP	-	-
	-998	TGCCAGGGGG	SP1	NP	-	-
	-1048	TCTGCCTGGCT	MYOGENIN	NP	-	-
	-1096	AGGCCTGGGC	SP1	NP	-	-

Abdulla, P., and Coe, I.R. (2007). *Nucleosides, nucleotides & nucleic acids*, **26**(1), 99-110.

DeVry, C.G., Tsai, W., and Clarke, S. (1996). *Archives of biochemistry and biophysics*, **335**(2), 321-332.

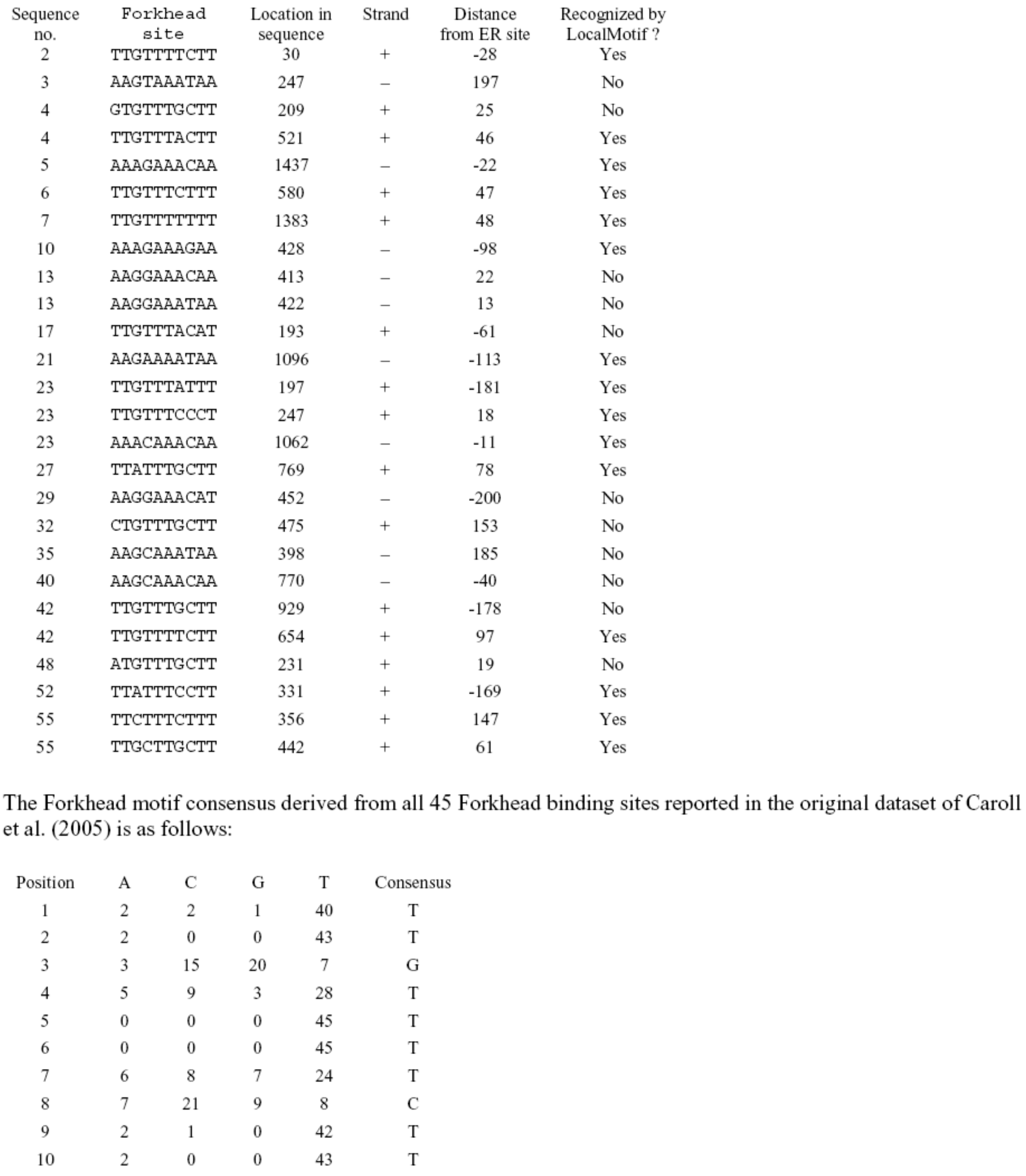
Heikkinen, S., Suppola, S., Malkki, M., Deeb, S.S., Janne, J., and Laakso, M. (2000). *Mammalian genome*, **11**(2), 91-96.

Liu, Y.H., and Teng, C.T. (1991). *The Journal of biological chemistry*, **266**(32), 21880-21885.

Myint, Z., Inazu, T., Tanaka, T., Yamada, K., Keng, V.W., Inoue, Y., Kuriyama, M., and Noguchi, T. (1999). *Journal of biochemistry*, **125**(4), 795-802.

Parham, K.L., Zervou, S., Karteris, E., Catalano, R.D., Old, R.W., and Hillhouse, E.W. (2004). *Endocrinology*, **145**(8), 3971-3983.

Supplementary Figure 3. Validation of Forkhead motif consensus identified by LocalMotif. All Forkhead binding sites present within 200 bp distance of a known ER full or half binding site are listed with their locations in the original dataset of Caroll et al. (2005). Binding sites that contribute to Forkhead consensus reported by LocalMotif are marked.



Sequence no.	Forkhead site	Location in sequence	Strand	Distance from ER site	Recognized by LocalMotif ?
2	TTGTTTTCTT	30	+	-28	Yes
3	AAGTAAATAA	247	–	197	No
4	GTGTTTGCTT	209	+	25	No
4	TTGTTTACTT	521	+	46	Yes
5	AAAGAAACAA	1437	–	-22	Yes
6	TTGTTTCTTT	580	+	47	Yes
7	TTGTTTTTTT	1383	+	48	Yes
10	AAAGAAAGAA	428	–	-98	Yes
13	AAGGAAACAA	413	–	22	No
13	AAGGAAATAA	422	–	13	No
17	TTGTTTACAT	193	+	-61	No
21	AAGAAAATAA	1096	–	-113	Yes
23	TTGTTTATTT	197	+	-181	Yes
23	TTGTTTCCCT	247	+	18	Yes
23	AAACAAACAA	1062	–	-11	Yes
27	TTATTTGCTT	769	+	78	Yes
29	AAGGAAACAT	452	–	-200	No
32	CTGTTTGCTT	475	+	153	No
35	AAGCAAATAA	398	–	185	No
40	AAGCAAACAA	770	–	-40	No
42	TTGTTTGCTT	929	+	-178	No
42	TTGTTTTCTT	654	+	97	Yes
48	ATGTTTGCTT	231	+	19	No
52	TTATTTCCTT	331	+	-169	Yes
55	TTCTTTCTTT	356	+	147	Yes
55	TTGCTTGCTT	442	+	61	Yes

The Forkhead motif consensus derived from all 45 Forkhead binding sites reported in the original dataset of Caroll et al. (2005) is as follows:

Position	A	C	G	T	Consensus
1	2	2	1	40	T
2	2	0	0	43	T
3	3	15	20	7	G
4	5	9	3	28	T
5	0	0	0	45	T
6	0	0	0	45	T
7	6	8	7	24	T
8	7	21	9	8	C
9	2	1	0	42	T
10	2	0	0	43	T

Supplementary Figure 4. Complete list of 813 CRMs predicted by Modulexplorer.

Location	Score	Target gene(s)
chr2L:156500-157500	0.928	CG3436
chr2L:293500-294500	0.923	U2AF38
chr2L:379500-380500	0.944	AL
chr2L:380000-381000	0.932	AL
chr2L:501000-502000	0.927	CBT;USH
chr2L:501500-502500	0.932	CBT;USH
chr2L:532500-533500	0.925	USH
chr2L:533000-534000	0.948	USH
chr2L:533500-534500	0.937	USH
chr2L:552000-553000	0.927	ETS21C
chr2L:589500-590500	0.977	GSC
chr2L:590000-591000	0.951	GSC
chr2L:597000-598000	0.924	GSC;CG13689
chr2L:599000-600000	0.933	GSC;CG13689
chr2L:599500-600500	0.925	GSC;CG13689
chr2L:805500-806500	0.936	PKG21D;CG31658
chr2L:1076000-1077000	0.932	S
chr2L:1076500-1077500	0.933	S
chr2L:1177500-1178500	0.939	CG4896
chr2L:1178000-1179000	0.924	CG4896
chr2L:1305500-1306500	0.922	ROBO3;A5
chr2L:1670000-1671000	0.932	CG31666
chr2L:1670500-1671500	0.949	CG31666
chr2L:2439500-2440500	0.922	DPP
chr2L:2628500-2629500	0.921	CG15395;CG9962
chr2L:2635500-2636500	0.922	CG15395;CG9962
chr2L:3543000-3544000	0.921	DRM
chr2L:3583500-3584500	0.921	SOB;ODD
chr2L:3612500-3613500	0.932	ODD;DOT
chr2L:3860000-3861000	0.922	SLP2;CG3964
chr2L:3860500-3861500	0.929	SLP2;CG3964
chr2L:3913000-3914000	0.926	FRED;CG31773;
chr2L:4038000-4039000	0.920	ED
chr2L:4306500-4307500	0.936	TUTL
chr2L:4366000-4367000	0.921	TRAF1
chr2L:4710000-4711000	0.924	CG34351
chr2L:5029000-5030000	0.926	CG25C
chr2L:5253500-5254500	0.924	TKV
chr2L:5301000-5302000	0.921	VRI
chr2L:5427500-5428500	0.926	H15;CG31647
chr2L:6204000-6205000	0.939	CG34380
chr2L:6204500-6205500	0.926	CG34380
chr2L:6531000-6532000	0.920	EYA
chr2L:6534000-6535000	0.924	EYA
chr2L:6914500-6915500	0.925	XL6
chr2L:6970500-6971500	0.920	SNRNP70K
chr2L:7010000-7011000	0.923	SP1070;CG13776
chr2L:7035000-7036000	0.923	CG11266
chr2L:7082000-7083000	0.930	PVF2
chr2L:7294000-7295000	0.948	CG31909;WG
chr2L:7298500-7299500	0.921	CG31909;WG
chr2L:7300000-7301000	0.920	CG31909;WG
chr2L:7313000-7314000	0.927	WG
chr2L:7319500-7320500	0.930	WG;WNT6
chr2L:7323500-7324500	0.924	WG;WNT6
chr2L:7333000-7334000	0.926	WG;WNT6
chr2L:7333500-7334500	0.971	WG;WNT6
chr2L:7355000-7356000	0.923	WNT6;WNT10
chr2L:7360000-7361000	0.924	WNT6;WNT10
chr2L:7539500-7540500	0.927	RAPGAP1
chr2L:7854500-7855500	0.934	CG14535;LECTIN-28C
chr2L:8247500-8248500	0.930	CG8086
chr2L:8248000-8249000	0.934	CG8086
chr2L:8809500-8810500	0.933	CG9468;SOXN
chr2L:9182500-9183500	0.923	TAI
chr2L:9592000-9593000	0.930	CG3841;CG4382
chr2L:9592500-9593500	0.976	CG3841;CG4382
chr2L:10647500-10648500	0.927	CG17097
chr2L:11049000-11050000	0.933	SAMUEL
chr2L:11049500-11050500	0.927	SAMUEL
chr2L:11162000-11163000	0.921	CA-BETA
chr2L:11182500-11183500	0.927	CG4788;AB
chr2L:11445500-11446500	0.921	SALM
chr2L:11986500-11987500	0.934	CG6686
chr2L:11987000-11988000	0.920	CG6686
chr2L:12587000-12588000	0.920	NUB
chr2L:12666000-12667000	0.921	PDM2
chr2L:12681500-12682500	0.964	PDM2
chr2L:12682000-12683000	0.931	PDM2
chr2L:13227000-13228000	0.920	CG16813
chr2L:13865500-13866500	0.957	CENG1A
chr2L:13871000-13872000	0.934	CENG1A
chr2L:13885500-13886500	0.926	CENG1A
chr2L:13889000-13890000	0.924	CENG1A
chr2L:14113000-14114000	0.925	CG31769;CG17341
chr2L:14113500-14114500	0.923	CG31769;CG17341
chr2L:14120500-14121500	0.926	CG31769;CG17341
chr2L:14125500-14126500	0.925	CG31769;CG17341
chr2L:14126000-14127000	0.926	CG31769;CG17341
chr2L:14133000-14134000	0.934	CG31769;CG17341
chr2L:14133500-14134500	0.922	CG31769;CG17341
chr2L:14154000-14155000	0.938	CG31769;CG17341
chr2L:14425000-14426000	0.950	CPR35B
chr2L:14425500-14426500	0.928	CPR35B;CG15283
chr2L:14489000-14490000	0.938	CG15283;NOC
chr2L:14531000-14532000	0.921	CG33648;CG4218
chr2L:14580500-14581500	0.924	CG3473;OSP
chr2L:15240500-15241500	0.922	CG3994
chr2L:15313500-15314500	0.924	CG15262;NHT
chr2L:15418500-15419500	0.921	CG18482;WOR
chr2L:15425000-15426000	0.932	WOR
chr2L:15425500-15426500	0.936	WOR
chr2L:15461000-15462000	0.925	CG4161;SNA
chr2L:15483500-15484500	0.956	SNA;TIM17B2
chr2L:15558500-15559500	0.920	CG15256
chr2L:15734500-15735500	0.930	CYCE
chr2L:15735000-15736000	0.951	CYCE
chr2L:16369500-16370500	0.927	JHAMT;CG5888
chr2L:16481000-16482000	0.921	DAC
chr2L:16830500-16831500	0.924	CG31739
chr2L:17227000-17228000	0.938	BEAT-IIIC
chr2L:18500500-18501500	0.920	PERD;CG10178
chr2L:18598000-18599000	0.927	AMOS;CG10413
chr2L:18778500-18779500	0.947	HAM
chr2L:19623000-19624000	0.920	LAR
chr2L:19937500-19938500	0.926	SICK
chr2L:20402000-20403000	0.923	CG10947;CG10949
chr2L:20483500-20484500	0.934	CG2493;MIR-1
chr2L:20484000-20485000	0.936	CG2493;MIR-1
chr2L:20526500-20527500	0.933	CG34007;CG15477
chr2L:20768500-20769500	0.952	DIA;CAD
chr2L:21889500-21890500	0.921	TSH;CG11629
chr2L:22659000-22660000	0.923	CG40006;CG41120
chr2L:22660500-22661500	0.935	CG40006;CG41120
chr2L:22661000-22662000	0.935	CG40006;CG41120
chr2L:22683500-22684500	0.927	CG41120;CG40439
chr2R:184500-185500	0.931	CG17665;GPRK1
chr2R:1704000-1705000	0.937	OR42B;DPR12
chr2R:1782000-1783000	0.924	DPR12;CG12551
chr2R:2162000-2163000	0.927	PLD;JING
chr2R:2162500-2163500	0.922	PLD;JING

chr2R:2454500-2455500	0.921	JING
chr2R:2455000-2456000	0.929	JING
chr2R:2460500-2461500	0.939	JING
chr2R:2973500-2974500	0.924	ESN
chr2R:3027000-3028000	0.927	CYP9B2;SPN43AA
chr2R:3050000-3051000	0.943	PK
chr2R:3106000-3107000	0.929	PK
chr2R:3520000-3521000	0.922	CG30492
chr2R:3879500-3880500	0.932	NUP44A;CG11196
chr2R:3880000-3881000	0.924	NUP44A;CG11196
chr2R:3922500-3923500	0.931	OPTIX
chr2R:3937500-3938500	0.923	OPTIX;CG12769
chr2R:4134000-4135000	0.920	PNUT;CG14760
chr2R:4464000-4465000	0.920	CG14748
chr2R:4724000-4725000	0.928	SNS
chr2R:5068000-5069000	0.931	LTD
chr2R:5068500-5069500	0.934	LTD
chr2R:5526000-5527000	0.921	MMP2
chr2R:5589000-5590000	0.924	UBA1;CG30002
chr2R:5964500-5965500	0.939	CG1371;EGR
chr2R:6036000-6037000	0.924	KCNQ
chr2R:6146500-6147500	0.927	CG12911
chr2R:6424500-6425500	0.924	LOLA
chr2R:6425000-6426000	0.934	LOLA
chr2R:6434000-6435000	0.924	LOLA;PSQ
chr2R:6434500-6435500	0.927	LOLA;PSQ
chr2R:6435500-6436500	0.921	LOLA;PSQ
chr2R:6539000-6540000	0.925	CG12934;STAN
chr2R:6987500-6988500	0.920	LUNA
chr2R:7285500-7286500	0.921	CG30022;CG7759
chr2R:7400500-7401500	0.925	INV;CG30034
chr2R:7445000-7446000	0.923	EN;TOU
chr2R:8275500-8276500	0.924	CPR49AC
chr2R:8276000-8277000	0.935	CPR49AC
chr2R:8337500-8338500	0.923	DYB
chr2R:8347500-8348500	0.930	LAC
chr2R:8589000-8590000	0.924	CG8776
chr2R:8898500-8899500	0.939	SU(Z)2;CG13323
chr2R:9226500-9227500	0.926	MIR-184S;CG17048
chr2R:9450500-9451500	0.925	CG13334;CG13335
chr2R:9451000-9452000	0.929	CG13334;CG13335
chr2R:9457500-9458500	0.927	CG13334;CG13335
chr2R:9505000-9506000	0.930	CG12464;FAS
chr2R:9714000-9715000	0.927	CG6329
chr2R:10064000-10065000	0.934	CG
chr2R:10155000-10156000	0.925	CG8547
chr2R:10318500-10319500	0.932	PHYL
chr2R:10319000-10320000	0.937	PHYL
chr2R:10341000-10342000	0.923	CG17389;CG17390
chr2R:10642500-10643500	0.922	CG10151
chr2R:10678500-10679500	0.926	KN
chr2R:10687500-10688500	0.926	KN
chr2R:11420500-11421500	0.929	KHC-73;CG30471
chr2R:11552000-11553000	0.929	FUS
chr2R:12042500-12043500	0.925	EXT2;CG10734
chr2R:12043000-12044000	0.926	EXT2;CG10734
chr2R:12164000-12165000	0.922	CG33017
chr2R:12256500-12257500	0.921	CG15711;CG33960
chr2R:12983500-12984500	0.924	GSTS1
chr2R:13149500-13150500	0.923	CG12699;MBL
chr2R:13222000-13223000	0.942	MBL
chr2R:13596000-13597000	0.924	CG6424
chr2R:13596500-13597500	0.942	CG6424
chr2R:13604000-13605000	0.941	CG6424
chr2R:13702500-13703500	0.929	GRH;CG30111;
chr2R:13703000-13704000	0.945	GRH;CG30111;
chr2R:14099500-14100500	0.932	CG30114;FJ
chr2R:14228500-14229500	0.922	TANGO8;IM23
chr2R:14867000-14868000	0.921	CG18605;CG15105
chr2R:15072500-15073500	0.921	CG7097
chr2R:15128000-15129000	0.931	CORA
chr2R:15142500-15143500	0.955	CG33453;CG7229
chr2R:15143000-15144000	0.940	CG33453;CG7229
chr2R:15144500-15145500	0.951	CG33453;CG7229
chr2R:15145000-15146000	0.971	CG33453;CG7229
chr2R:15145500-15146500	0.921	CG33453;CG7229
chr2R:15152500-15153500	0.936	CG7229;RIB
chr2R:15549000-15550000	0.932	MIR-3;MIR-309;CG11018
chr2R:15681500-15682500	0.930	OBP56G;OBP56H
chr2R:15682000-15683000	0.930	OBP56G;OBP56H
chr2R:15783000-15784000	0.922	CG13872;CG30447
chr2R:16828000-16829000	0.934	RX;ACT57B
chr2R:16848000-16849000	0.921	HBN
chr2R:17063500-17064500	0.937	PU
chr2R:17249500-17250500	0.922	CV-2
chr2R:17432500-17433500	0.930	EGFR
chr2R:17540500-17541500	0.935	CG10082
chr2R:17566000-17567000	0.932	CG30263
chr2R:17566500-17567500	0.935	CG30263
chr2R:17766000-17767000	0.937	FILI
chr2R:17930500-17931500	0.921	PPD5;CG13500
chr2R:18119000-18120000	0.920	OATP58DC;DVE
chr2R:18136000-18137000	0.929	DVE
chr2R:18146500-18147500	0.927	DVE
chr2R:18147000-18148000	0.920	DVE
chr2R:18258000-18259000	0.922	CG11206
chr2R:18394500-18395500	0.942	PX
chr2R:18400000-18401000	0.925	PX
chr2R:18630500-18631500	0.929	CG3536
chr2R:18799000-18800000	0.921	NAHODA
chr2R:18922500-18923500	0.926	CG9895;CG34209
chr2R:19078000-19079000	0.920	CG34371
chr2R:19086500-19087500	0.926	CG34371
chr2R:19094000-19095000	0.928	CG13539;CG3162
chr2R:19104500-19105500	0.922	CG3162;CG3092
chr2R:19212500-19213500	0.932	CG9861
chr2R:19221500-19222500	0.928	CG30413;CG3502
chr2R:19654500-19655500	0.921	CG9850
chr2R:19685000-19686000	0.926	CG9850
chr2R:19758500-19759500	0.929	KEN
chr2R:19759000-19760000	0.937	KEN
chr2R:20050500-20051500	0.921	CG3328
chr2R:20130500-20131500	0.920	NORD
chr2R:20225500-20226500	0.949	SLBO;BS
chr2R:20226000-20227000	0.934	SLBO;BS
chr2R:20227000-20228000	0.931	SLBO;BS
chr2R:20311500-20312500	0.924	EYC
chr2R:20413000-20414000	0.925	CG4612;CG30169;
chr2R:20681500-20682500	0.920	ETH
chr2R:20936500-20937500	0.934	GSB-N
chr3L:26500-27500	0.931	MTHL8;LSP1GAMMA
chr3L:209500-210500	0.951	CG7028
chr3L:383500-384500	0.936	TRH
chr3L:384000-385000	0.920	TRH
chr3L:413000-414000	0.923	CG13885;CG13891
chr3L:413500-414500	0.923	CG13885;CG13891
chr3L:450000-451000	0.928	KLAR
chr3L:592000-593000	0.947	CG17181;MED14
chr3L:607500-608500	0.947	REG-2;BAN
chr3L:613500-614500	0.921	REG-2;BAN
chr3L:645000-646000	1.000	BAN;CG12030
chr3L:645500-646500	0.948	BAN;CG12030
chr3L:668000-669000	0.921	REV1;CG17129
chr3L:697000-698000	0.921	CG13894
chr3L:713000-714000	0.928	CG13896;CG13897
chr3L:725000-726000	0.925	CG13896;CG13897
chr3L:725500-726500	0.945	CG13896;CG13897
chr3L:746000-747000	0.933	CG13897;EMC
chr3L:1239500-1240500	0.929	CG9194
chr3L:1240000-1241000	0.921	CG9194
chr3L:1311000-1312000	0.943	CG2469;CG9186;
chr3L:1311500-1312500	0.930	CG2469;CG9186;
chr3L:1370000-1371000	0.923	PTP61F;RU;
chr3L:1441500-1442500	0.923	SA-2;RHO
chr3L:1445000-1446000	0.926	SA-2;RHO
chr3L:1445500-1446500	0.927	SA-2;RHO
chr3L:1609500-1610500	0.928	CG18170;CG33791;

chr3L:1683000-1684000	0.934	CG13930;CG12011
chr3L:1910000-1911000	0.929	CG1887
chr3L:2115000-2116000	0.927	SLS
chr3L:2137500-2138500	0.923	ZORMIN
chr3L:2553000-2554000	0.925	SPN
chr3L:2762000-2763000	0.921	MRTF
chr3L:2995500-2996500	0.935	CG2113
chr3L:3851500-3852500	0.929	AWH
chr3L:4168000-4169000	0.926	MAS;ERO1L
chr3L:4409000-4410000	0.924	CG11347
chr3L:5381000-5382000	0.922	CG10633;CG34391
chr3L:5454000-5455000	0.941	CG34391;CG4835
chr3L:5454500-5455500	0.930	CG34391;CG4835
chr3L:5616500-5617500	0.927	EAF6;BLIMP-1
chr3L:6109000-6110000	0.921	ETS65A
chr3L:6282000-6283000	0.921	IMPL3;OR65A
chr3L:6352500-6353500	0.935	CG13300;CG32398
chr3L:6375000-6376000	0.934	CG32398;CG14910
chr3L:6375500-6376500	0.936	CG32398;CG14910
chr3L:6437500-6438500	0.926	CG32398;CG14910
chr3L:6697000-6698000	0.921	SP1173;CG8519
chr3L:6697500-6698500	0.925	SP1173;CG8519
chr3L:6728000-6729000	0.920	CG32394
chr3L:6729000-6730000	0.948	CG32394
chr3L:6729500-6730500	0.924	CG32394
chr3L:6833000-6834000	0.935	VVL;PRAT2
chr3L:6833500-6834500	0.961	VVL;PRAT2
chr3L:6834000-6835000	0.926	VVL;PRAT2
chr3L:6916500-6917500	0.927	PRAT2;CG14820
chr3L:7103500-7104500	0.923	FORM3
chr3L:7130500-7131500	0.926	MELT
chr3L:7177000-7178000	0.934	CG32387
chr3L:7607500-7608500	0.925	CG33275
chr3L:7735000-7736000	0.920	CG32373;CG7422;
chr3L:7768000-7769000	0.929	CLK
chr3L:8210000-8211000	0.922	CG8012;CG13674
chr3L:8304000-8305000	0.935	CG7185
chr3L:8995500-8996500	0.928	CG5087
chr3L:9022000-9023000	0.928	DOC2;DOC1
chr3L:9603000-9604000	0.926	CG3280
chr3L:9619000-9620000	0.920	CG3335
chr3L:10122000-10123000	0.921	CG6640;DPR10
chr3L:10274000-10275000	0.924	OR67D;CG6559
chr3L:10290500-10291500	0.927	CG6559
chr3L:10322500-10323500	0.948	MIR-276B;MIR-276AS
chr3L:10323000-10324000	0.957	MIR-276B;MIR-276AS
chr3L:10708500-10709500	0.921	NIJA;CG12523
chr3L:10784000-10785000	0.927	NIJA;CG12523
chr3L:10819500-10820500	0.930	CG12523;TNA
chr3L:10825500-10826500	0.920	CG12523;TNA
chr3L:10845500-10846500	0.924	TNA
chr3L:10946000-10947000	0.927	CG34050;CG14147
chr3L:10994000-10995000	0.944	KLU
chr3L:11039000-11040000	0.931	CG6327
chr3L:11370500-11371500	0.940	CG6163;CG11726
chr3L:11566000-11567000	0.926	CG33490
chr3L:11864500-11865500	0.931	CG5906;CG14128
chr3L:11918000-11919000	0.921	CG5718;BYN
chr3L:11920000-11921000	0.923	CG5718;BYN
chr3L:12076000-12077000	0.922	SEMA-5C;CG17154
chr3L:12177500-12178500	0.923	CAH2;CG6910
chr3L:12434500-12435500	0.921	TOE;EYG
chr3L:12454000-12455000	0.937	TOE;EYG
chr3L:12454500-12455500	0.933	TOE;EYG
chr3L:12567500-12568500	0.924	CG10632;ARA
chr3L:12608000-12609000	0.922	CAUP
chr3L:12640000-12641000	0.951	CG32111
chr3L:12644500-12645500	0.922	CG32111;MIRR
chr3L:12680500-12681500	0.923	CG32111;MIRR
chr3L:13180500-13181500	0.933	CG11281;CAPS
chr3L:13185500-13186500	0.927	CG11281;CAPS
chr3L:13371000-13372000	0.927	SNCF
chr3L:13384500-13385500	0.934	CG10191;SENS
chr3L:13401000-13402000	0.921	SENS;CG10222
chr3L:13401500-13402500	0.929	SENS;CG10222
chr3L:13487000-13488000	0.926	CG10741
chr3L:13649000-13650000	0.923	BRU-3
chr3L:13659500-13660500	0.920	BRU-3
chr3L:14132000-14133000	0.921	SOX21B;D
chr3L:14141500-14142500	0.946	SOX21B;D
chr3L:14142000-14143000	0.936	SOX21B;D
chr3L:14149500-14150500	0.920	SOX21B;D
chr3L:14176500-14177500	0.923	D;NAN
chr3L:14283500-14284500	0.942	FZ
chr3L:14443000-14444000	0.948	CG9598;BBG
chr3L:14443500-14444500	0.921	CG9598;BBG
chr3L:14960500-14961500	0.950	BOBA;TOM
chr3L:15540000-15541000	0.924	CREBA
chr3L:15692500-15693500	0.926	COMM2
chr3L:16091500-16092500	0.920	CG5389
chr3L:16188000-16189000	0.927	CG33795;CG33687
chr3L:16403500-16404500	0.922	FAX
chr3L:16466000-16467000	0.938	CG33158;ARGOS;
chr3L:16765500-16766500	0.921	NRT
chr3L:16766000-16767000	0.924	NRT
chr3L:16933000-16934000	0.921	CPR73D
chr3L:17024500-17025500	0.922	CG14059;CG7842
chr3L:17025000-17026000	0.928	CG14059;CG7842
chr3L:17026500-17027500	0.921	CG14059;CG7842
chr3L:17448500-17449500	0.922	CG18265;CG7603
chr3L:18615500-18616500	0.920	CG6897
chr3L:18616000-18617000	0.926	CG6897
chr3L:18842500-18843500	0.936	CG32027;MIR-315
chr3L:18843000-18844000	0.966	CG32027;MIR-315
chr3L:19033500-19034500	0.921	NKD
chr3L:19050000-19051000	0.938	ACP76A;CG3797
chr3L:19050500-19051500	0.937	ACP76A;CG3797
chr3L:19166000-19167000	0.921	FZ2;CG33647
chr3L:19195000-19196000	0.925	FZ2;CG33647
chr3L:19224000-19225000	0.924	FZ2;CG33647
chr3L:19525000-19526000	0.931	CPR76BC;CPR76BD
chr3L:19649500-19650500	0.927	TEY
chr3L:19675500-19676500	0.922	TEY;CG8765
chr3L:19700000-19701000	0.922	SERP
chr3L:19722000-19723000	0.922	GYC76C
chr3L:20622500-20623500	0.928	KNRL;CG13251
chr3L:20627500-20628500	0.921	KNRL;CG13251
chr3L:20628000-20629000	0.929	KNRL;CG13251
chr3L:20909500-20910500	0.922	CG11458;FNG
chr3L:21591000-21592000	0.920	AP-2
chr3L:21679500-21680500	0.924	CG11248;CG32447
chr3L:21702500-21703500	0.948	CG32447;CG11247
chr3L:21703000-21704000	0.932	CG32447;CG11247
chr3L:22190500-22191500	0.925	OLF413
chr3L:22832500-22833500	0.932	CG11226;CG32462
chr3L:24483000-24484000	0.936	CG41458;CG40057
chr3R:288000-289000	0.927	CG41465
chr3R:602500-603500	0.928	ATMS;CG14655;
chr3R:674500-675500	0.927	CG14659;OPA
chr3R:679500-680500	0.936	OPA
chr3R:685000-686000	0.928	OPA
chr3R:705500-706500	0.929	LAF;GNF1
chr3R:710500-711500	0.934	LAF;GNF1
chr3R:935000-936000	0.927	CORTO;CG12007
chr3R:935500-936500	0.923	CORTO;CG12007
chr3R:1409500-1410500	0.920	CG2926
chr3R:1439000-1440000	0.926	ATU
chr3R:1545500-1546500	0.921	CAS;CG1239
chr3R:1548000-1549000	0.932	CAS;CG1239
chr3R:1548500-1549500	0.939	CAS;CG1239
chr3R:1554500-1555500	0.924	CAS;CG1239
chr3R:2803000-2804000	0.934	ANTP
chr3R:2803500-2804500	0.920	ANTP
chr3R:2820500-2821500	0.927	ANTP
chr3R:3084000-3085000	0.923	GLD
chr3R:3721000-3722000	0.936	MRPS9;CG31284;
chr3R:3721500-3722500	0.925	MRPS9;CG31284;

chr3R:3936000-3937000	0.931	PUC
chr3R:3936500-3937500	0.930	PUC
chr3R:3961000-3962000	0.921	CG7891;GRN
chr3R:3989500-3990500	0.940	GRN
chr3R:3990000-3991000	0.928	GRN
chr3R:4049000-4050000	0.922	CG18249;DNAPOL-IOTA
chr3R:4049500-4050500	0.941	CG18249;DNAPOL-IOTA
chr3R:4507500-4508500	0.985	CG33325;HB
chr3R:4508000-4509000	0.934	CG33325;HB
chr3R:4510000-4511000	0.924	CG33325;HB
chr3R:4514000-4515000	0.936	CG33325;HB
chr3R:4618000-4619000	0.920	PIF1B;PIF1A;CG11776
chr3R:4687500-4688500	0.920	PYD;CG9731;
chr3R:5860500-5861500	0.948	MICAL
chr3R:6292000-6293000	0.938	CYP12E1;HTH
chr3R:6299000-6300000	0.933	CYP12E1;HTH
chr3R:6299500-6300500	0.924	CYP12E1;HTH
chr3R:6321500-6322500	0.927	CYP12E1;HTH
chr3R:6383500-6384500	0.921	HTH
chr3R:6390500-6391500	0.940	HTH
chr3R:6391000-6392000	0.921	HTH
chr3R:6393000-6394000	0.936	HTH
chr3R:6393500-6394500	0.923	HTH
chr3R:6405500-6406500	0.926	HTH
chr3R:6447500-6448500	0.924	HTH
chr3R:6494500-6495500	0.923	CG34304;CG6465
chr3R:6495000-6496000	0.925	CG34304;CG6465
chr3R:6519000-6520000	0.920	CG14681;SKELETOR;
chr3R:6862000-6863000	0.923	CG34114
chr3R:7031000-7032000	0.929	NOCTURNIN
chr3R:7105500-7106500	0.923	CG31386
chr3R:7130500-7131500	0.920	CG31386
chr3R:7131000-7132000	0.928	CG31386
chr3R:7136000-7137000	0.920	CG31386;KP78B
chr3R:7136500-7137500	0.945	CG31386;KP78B
chr3R:7137000-7138000	0.923	CG31386;KP78B
chr3R:7180500-7181500	0.922	KP78A;PROS
chr3R:8101000-8102000	0.927	SVP
chr3R:8113500-8114500	0.946	SVP
chr3R:8114000-8115000	0.926	SVP
chr3R:8117500-8118500	0.929	SVP
chr3R:9040500-9041500	0.923	CG8483;CG8476
chr3R:9293500-9294500	0.932	CG17025;CG12538
chr3R:9636500-9637500	0.924	DIP-B;PRI
chr3R:9683500-9684500	0.924	CG14362;E5
chr3R:9705000-9706000	0.936	E5;EMS
chr3R:9712500-9713500	0.936	E5;EMS
chr3R:9713000-9714000	0.922	E5;EMS
chr3R:9736500-9737500	0.920	EMS;ART9
chr3R:9767000-9768000	0.924	ART9
chr3R:9767500-9768500	0.934	ART9
chr3R:9822500-9823500	0.926	RDX
chr3R:10412500-10413500	0.932	STUMPS
chr3R:11468500-11469500	0.927	CG5302;PXB
chr3R:11469000-11470000	0.921	CG5302;PXB
chr3R:11473000-11474000	0.922	CG5302;PXB
chr3R:11475000-11476000	0.927	CG5302;PXB
chr3R:11495000-11496000	0.945	PXB
chr3R:11811500-11812500	0.925	SRP
chr3R:11938500-11939500	0.928	AKT1;SB
chr3R:11977500-11978500	0.926	FER2;CG6006
chr3R:12348000-12349000	0.930	CG11769;CG14888
chr3R:12430000-12431000	0.921	CG10345
chr3R:12606500-12607500	0.933	TRE-1;GLUT3
chr3R:12618000-12619000	0.926	GLUT3;ABD-A
chr3R:12671500-12672500	0.925	ABD-A;IAB-4
chr3R:12690000-12691000	0.922	MIR-IAB-4-3P;CG10349
chr3R:12690500-12691500	0.944	MIR-IAB-4-3P;CG10349
chr3R:12717000-12718000	0.923	CG10349;ABD-B
chr3R:12717500-12718500	0.943	CG10349;ABD-B
chr3R:12727500-12728500	0.930	CG10349;ABD-B
chr3R:12763000-12764000	0.921	ABD-B
chr3R:12769500-12770500	0.928	ABD-B
chr3R:12770000-12771000	0.922	ABD-B
chr3R:12778500-12779500	0.920	ABD-B
chr3R:12982500-12983500	0.921	CG4090
chr3R:13377000-13378000	0.927	PRX5037;HMX
chr3R:13890500-13891500	0.922	HTL;CG14317
chr3R:15107500-15108500	0.920	CG14280;DL
chr3R:15155500-15156500	0.938	DL;INO80
chr3R:15171500-15172500	0.923	DL;INO80
chr3R:15716000-15717000	0.924	NINAE;CG4733
chr3R:16303500-16304500	0.921	CG34118
chr3R:16625500-16626500	0.933	CG17838
chr3R:16888500-16889500	0.928	CORTACTIN
chr3R:16889000-16890000	0.938	CORTACTIN
chr3R:17077000-17078000	0.922	ETHR
chr3R:17240000-17241000	0.937	LBL
chr3R:17242000-17243000	0.928	LBL
chr3R:17261500-17262500	0.925	LBL;LBE
chr3R:17284000-17285000	0.930	LBE;CG7922
chr3R:17291500-17292500	0.927	LBE;CG7922
chr3R:17292000-17293000	0.935	LBE;CG7922
chr3R:17292500-17293500	0.929	LBE;CG7922
chr3R:17329000-17330000	0.922	C15
chr3R:17329500-17330500	0.923	C15
chr3R:17379500-17380500	0.923	SLOU
chr3R:17470000-17471000	0.923	E2F
chr3R:17774000-17775000	0.920	CG17843;EIP93F
chr3R:18053500-18054500	0.928	CG31163
chr3R:18955500-18956500	0.922	HH
chr3R:19942500-19943500	0.926	CG13607;CG6129
chr3R:20409500-20410500	0.923	CG13625
chr3R:20657500-20658500	0.927	CG31357;CG13636
chr3R:20672000-20673000	0.932	ESP
chr3R:21000500-21001500	0.923	CG34110
chr3R:21016000-21017000	0.949	CG34110;DAN;
chr3R:21088500-21089500	0.924	CG11858;AATS-GLN;
chr3R:21855500-21856500	0.935	HLHM5;M6
chr3R:22641500-22642500	0.923	TL
chr3R:22941500-22942500	0.931	CG6066
chr3R:23096000-23097000	0.924	L(3)MBT
chr3R:23651000-23652000	0.929	CG34353
chr3R:23651500-23652500	0.925	CG34353
chr3R:24684500-24685500	0.934	CG31048;INX3;
chr3R:24712500-24713500	0.920	CG33203
chr3R:25080500-25081500	0.926	STG
chr3R:25081000-25082000	0.924	STG
chr3R:25119500-25120500	0.931	STG;CG14506
chr3R:25120000-25121000	0.940	STG;CG14506
chr3R:25122500-25123500	0.923	CG14506
chr3R:25125500-25126500	0.921	CG14506;CNX99A
chr3R:25129500-25130500	0.927	CG14506;CNX99A
chr3R:25130500-25131500	0.940	CG14506;CNX99A
chr3R:25131000-25132000	0.928	CG14506;CNX99A
chr3R:25132000-25133000	0.924	CG14506;CNX99A
chr3R:25132500-25133500	0.942	CG14506;CNX99A
chr3R:25133000-25134000	0.925	CG14506;CNX99A
chr3R:25400000-25401000	0.922	DR;CG7567
chr3R:25409000-25410000	0.931	DR;CG7567
chr3R:25411500-25412500	0.923	DR;CG7567
chr3R:25736500-25737500	0.933	CG31036
chr3R:25874500-25875500	0.921	CG15528
chr3R:25875000-25876000	0.921	CG15528
chr3R:25999000-26000000	0.922	MLC2
chr3R:26017500-26018500	0.921	CG9747;CG15531
chr3R:26049500-26050500	0.923	CECC;CG9737
chr3R:26101500-26102500	0.926	CG18404;HDC
chr3R:26151000-26152000	0.921	HDC
chr3R:26159500-26160500	0.929	HDC
chr3R:26160000-26161000	0.935	HDC
chr3R:26189500-26190500	0.923	CG34300;FER1HCH
chr3R:26240000-26241000	0.921	FER2LCH;CG2217
chr3R:26491500-26492500	0.922	CG34433;CG1342
chr3R:26544500-26545500	0.920	CG31010;CG1340
chr3R:26701000-26702000	0.928	CG15546
chr3R:26736000-26737000	0.936	CG33483;PTX1

chr3R:26746000-26747000	0.955	PTX1
chr3R:26746500-26747500	0.965	PTX1
chr3R:26770000-26771000	0.927	CG15550;CG15548
chr3R:26796500-26797500	0.925	5-HT7
chr3R:26890000-26891000	0.943	DCO;SOX100B
chr3R:27018000-27019000	0.931	CG31004;BNK;
chr3R:27018500-27019500	0.920	CG31004;BNK;
chr3R:27318000-27319000	0.925	CG11333;CG12063
chr3R:27321000-27322000	0.925	CG11333;CG12063
chr3R:27345000-27346000	0.922	CG12063;CG1499
chr3R:27424000-27425000	0.925	CYCG
chr3R:27508500-27509500	0.930	CG11550;CG34046
chr3R:27509000-27510000	0.931	CG11550;CG34046
chr3R:27527500-27528500	0.924	CG11550;CG34046
chr3R:27556000-27557000	0.923	TTK
chr3R:27556500-27557500	0.926	TTK
chrX:1155000-1156000	0.924	A3-3
chrX:1155500-1156500	0.941	A3-3
chrX:1177000-1178000	0.937	A3-3;CG32812
chrX:1276000-1277000	0.938	CG32813
chrX:1276500-1277500	0.925	CG32813
chrX:1277500-1278500	0.925	CG32813
chrX:1296000-1297000	0.921	CG11448;FUTSCH
chrX:1360500-1361500	0.923	CG14777
chrX:1433500-1434500	0.925	CG32810;CG14796
chrX:1438000-1439000	0.937	CG32810;CG14796
chrX:1438500-1439500	0.939	CG32810;CG14796
chrX:1486500-1487500	0.921	CG14796;BR
chrX:1501000-1502000	0.925	CG14796;BR
chrX:1505000-1506000	0.948	BR
chrX:2027500-2028500	0.933	PH-P
chrX:2280000-2281000	0.930	CG14045
chrX:2280500-2281500	0.956	CG14045
chrX:2285000-2286000	0.922	CG14045;CG12496
chrX:2285500-2286500	0.938	CG14045;CG12496
chrX:2286500-2287500	0.923	CG14045;CG12496
chrX:2292000-2293000	0.938	CG14045;CG12496
chrX:2292500-2293500	0.931	CG14045;CG12496
chrX:2306000-2307000	0.940	CG12496;CG32797
chrX:2306500-2307500	0.934	CG12496;CG32797
chrX:2310000-2311000	0.929	CG12496;CG32797
chrX:2310500-2311500	0.926	CG12496;CG32797
chrX:2358500-2359500	0.931	BOI
chrX:2359000-2360000	0.942	BOI
chrX:2360000-2361000	0.922	BOI
chrX:2360500-2361500	0.947	BOI
chrX:2361000-2362000	0.932	BOI
chrX:2361500-2362500	0.929	BOI
chrX:2438500-2439500	0.940	TROL
chrX:2439000-2440000	0.920	TROL
chrX:2823500-2824500	0.931	CG32793
chrX:2844000-2845000	0.933	CG3603;RST
chrX:2915000-2916000	0.928	RST;CG4116
chrX:2915500-2916500	0.921	RST;CG4116
chrX:2936000-2937000	0.928	RST;CG4116
chrX:3030000-3031000	0.938	N
chrX:3030500-3031500	0.921	N
chrX:3238500-3239500	0.921	CG10793;DM
chrX:3291500-3292500	0.921	CG12535;CG14269
chrX:3528500-3529500	0.921	ALSTR
chrX:3737000-3738000	0.927	EC
chrX:3739000-3740000	0.925	EC
chrX:3955000-3956000	0.924	CG6414;CG32790
chrX:3967500-3968500	0.929	CG6414;CG32790
chrX:4024000-4025000	0.946	CG4857;GLCAT-I
chrX:4090500-4091500	0.924	FAS2
chrX:4285500-4286500	0.922	CG32773;BI
chrX:4340500-4341500	0.925	BI
chrX:4350500-4351500	0.927	BI
chrX:4385000-4386000	0.925	BI;CG12685
chrX:4388500-4389500	0.921	BI;CG12685
chrX:4417000-4418000	0.922	CG12685;CG3556
chrX:4455500-4456500	0.930	CG3546;CG12684
chrX:4456000-4457000	0.931	CG3546;CG12684
chrX:4461500-4462500	0.923	CG3546;CG12684
chrX:4462000-4463000	0.944	CG3546;CG12684
chrX:4463000-4464000	0.927	CG3546;CG12684
chrX:4463500-4464500	0.935	CG3546;CG12684
chrX:4533500-4534500	0.922	CG3081
chrX:4877000-4878000	0.921	SIP3;CG12680
chrX:4909000-4910000	0.922	CG12680;OVO
chrX:4953500-4954500	0.927	OVO
chrX:5430500-5431500	0.921	CG33980
chrX:5440500-5441500	0.936	CG33980
chrX:5478000-5479000	0.922	CG33980;CG4136
chrX:5647500-5648500	0.921	CG15772;LIN-52;
chrX:5691500-5692500	0.921	CG12239;CPR5C
chrX:5730000-5731000	0.926	CG15765
chrX:5904000-5905000	0.925	MAB-2
chrX:5916500-5917500	0.932	MAB-2;RUX
chrX:5922500-5923500	0.941	MAB-2;RUX
chrX:5923500-5924500	0.932	MAB-2;RUX
chrX:5924000-5925000	0.943	MAB-2;RUX
chrX:5952500-5953500	0.924	CG5937;CG5921
chrX:5983000-5984000	0.921	MIPP2;RAPTOR;
chrX:6039000-6040000	0.927	CA-ALPHA1T
chrX:6050500-6051500	0.927	CA-ALPHA1T;CG32750
chrX:6634000-6635000	0.930	CG3168
chrX:6638500-6639500	0.928	RPL17;CG14439
chrX:7044000-7045000	0.932	CR32730;CG9650
chrX:7064500-7065500	0.948	CR32730;CG9650
chrX:7069000-7070000	0.929	CR32730;CG9650
chrX:7092000-7093000	0.936	CG9650
chrX:7098500-7099500	0.922	CG9650
chrX:7117500-7118500	0.923	CG9650
chrX:7118000-7119000	0.932	CG9650
chrX:7129500-7130500	0.933	CG9650
chrX:7149500-7150500	0.923	CG9650;CG1958
chrX:7153000-7154000	0.925	CG9650;CG1958
chrX:7160000-7161000	0.940	CG1958
chrX:7167000-7168000	0.922	CG1958;CG1677
chrX:7167500-7168500	0.940	CG1958;CG1677
chrX:7168000-7169000	0.929	CG1958;CG1677
chrX:7206000-7207000	0.921	BRK;ATG5
chrX:7463000-7464000	0.923	CG32720;CG11369
chrX:7463500-7464500	0.926	CG32720;CG11369
chrX:7489500-7490500	0.933	CG12689;CT
chrX:7490000-7491000	0.934	CG12689;CT
chrX:7509500-7510500	0.943	CT
chrX:7549000-7550000	0.921	CT
chrX:7785500-7786500	0.933	NEK2
chrX:7799000-7800000	0.920	DPR14
chrX:8160000-8161000	0.922	CG1632
chrX:8510500-8511500	0.922	CAF1-180;OC
chrX:8659500-8660500	0.922	LIM1
chrX:8723500-8724500	0.920	CG32710;CG12075
chrX:8724000-8725000	0.924	CG32710;CG12075
chrX:10046000-10047000	0.932	CG34104;CG12645
chrX:10206000-10207000	0.945	ALPHA-MAN-I
chrX:10578000-10579000	0.921	X11LBETA
chrX:10720500-10721500	0.922	IMP;SBR
chrX:10723000-10724000	0.937	IMP;SBR
chrX:10723500-10724500	0.934	IMP;SBR
chrX:10826000-10827000	0.926	CG2145;MYO10A
chrX:11370500-11371500	0.936	CYP4G15
chrX:11371000-11372000	0.966	CYP4G15
chrX:11830000-11831000	0.930	CAC
chrX:12047000-12048000	0.920	CG2750;CG1924
chrX:12119000-12120000	0.924	CG12720;TEN-A
chrX:12119500-12120500	0.929	CG12720;TEN-A
chrX:12130000-12131000	0.925	TEN-A
chrX:12155000-12156000	0.936	TEN-A
chrX:12200000-12201000	0.929	TEN-A
chrX:13243500-13244500	0.926	CG15747;CG10617
chrX:13250500-13251500	0.922	CG15747;CG10617
chrX:13713500-13714500	0.928	CLIC
chrX:13957000-13958000	0.929	STE12DOR;CG32614

chrX:14059000-14060000	0.920	CG12479;CG12480
chrX:14498500-14499500	0.924	NETA
chrX:14544500-14545500	0.930	NETA
chrX:14884500-14885500	0.924	EAG
chrX:14885000-14886000	0.927	EAG
chrX:14966500-14967500	0.921	LSD-2
chrX:15250000-15251000	0.940	ACJ6
chrX:15504000-15505000	0.930	SOG
chrX:15526000-15527000	0.921	SOG;CG8117
chrX:15540500-15541500	0.931	CG8117;CG8119
chrX:15558000-15559000	0.928	CG8119;CG15646
chrX:15563000-15564000	0.924	CG15646
chrX:15565500-15566500	0.930	CG15646;CG12708
chrX:15566000-15567000	0.925	CG15646;CG12708
chrX:15575500-15576500	0.935	CG12708;CG15599
chrX:16020000-16021000	0.951	DISCO-R;DISCO
chrX:16027000-16028000	0.937	DISCO-R;DISCO
chrX:16199500-16200500	0.924	CG3679
chrX:16205500-16206500	0.924	CG9915;CG33251;
chrX:16206000-16207000	0.921	CG9915;CG33251;
chrX:16966500-16967500	0.936	CG8949
chrX:17244500-17245500	0.943	B-H2;B-H1
chrX:17301500-17302500	0.922	B-H1;CG8611
chrX:17302000-17303000	0.936	B-H1;CG8611
chrX:17440000-17441000	0.945	CG12432;PPK23
chrX:17650500-17651500	0.923	CG15816;UNC-4
chrX:17651000-17652000	0.932	CG15816;UNC-4
chrX:17662000-17663000	0.940	CG15816;UNC-4
chrX:17683000-17684000	0.929	UNC-4;ODSH
chrX:17702500-17703500	0.922	ODSH
chrX:17705500-17706500	0.924	ODSH
chrX:17919000-17920000	0.931	SH;CG6867;
chrX:18129000-18130000	0.925	CG33639;UPD2
chrX:18151000-18152000	0.926	UPD2;CG15059
chrX:18151500-18152500	0.925	UPD2;CG15059
chrX:18164000-18165000	0.930	CG15059;CG15057
chrX:18164500-18165500	0.940	CG15059;CG15057
chrX:18175500-18176500	0.922	UPD3
chrX:18176000-18177000	0.932	UPD3
chrX:18178500-18179500	0.921	UPD3
chrX:18182500-18183500	0.923	UPD3;OS
chrX:18198000-18199000	0.938	UPD3;OS
chrX:18202500-18203500	0.923	OS
chrX:18203000-18204000	0.942	OS
chrX:18206500-18207500	0.925	OS;CG6023
chrX:18207000-18208000	0.929	OS;CG6023
chrX:18397500-18398500	0.931	WNT5
chrX:18433500-18434500	0.923	BX
chrX:18453500-18454500	0.928	BX
chrX:18484000-18485000	0.920	BX;CG15040
chrX:18538000-18539000	0.925	L(1)G0003
chrX:18595000-18596000	0.927	CYP18A1;CCKLR-17D1
chrX:18608500-18609500	0.922	CCKLR-17D1
chrX:18785000-18786000	0.962	CG7358
chrX:18785500-18786500	0.953	CG7358
chrX:18786000-18787000	0.938	CG7358
chrX:18839500-18840500	0.932	CG32541
chrX:18867500-18868500	0.934	CG32541
chrX:18895000-18896000	0.923	CG32541
chrX:18895500-18896500	0.923	CG32541
chrX:18936500-18937500	0.933	CG32541;CG34329
chrX:19064000-19065000	0.924	INX5
chrX:19064500-19065500	0.922	INX5
chrX:19114500-19115500	0.923	CG7884
chrX:19702000-19703000	0.935	CG12701;SKPD
chrX:19702500-19703500	0.920	CG12701;SKPD
chrX:19873500-19874500	0.928	NEP3;CG17003
chrX:20108500-20109500	0.921	SW;OBST-A
chrX:20413500-20414500	0.929	CG15455
chrX:20422000-20423000	0.927	CG15455
chrX:20424500-20425500	0.923	CG15455
chrX:20472500-20473500	0.930	CG34145
chrX:20487000-20488000	0.923	CG34145;CYP6V1
chrX:20516500-20517500	0.921	CG34145;CYP6V1
chrX:20522500-20523500	0.925	CG34145;CYP6V1
chrX:20523000-20524000	0.936	CG34145;CYP6V1
chrX:20548000-20549000	0.947	HYDRA;RUN
chrX:20584000-20585000	0.926	RUN;CG1324
chrX:20584500-20585500	0.946	RUN;CG1324
chrX:20595500-20596500	0.923	RUN;CG1324
chrX:20596000-20597000	0.926	RUN;CG1324
chrX:20597000-20598000	0.924	RUN;CG1324
chrX:20597500-20598500	0.922	RUN;CG1324
chrX:20605000-20606000	0.921	CG1324
chrX:20661000-20662000	0.932	SHAKB
chrX:20706000-20707000	0.921	CG15450;CG1314
chrX:20706500-20707500	0.924	CG15450;CG1314
chrX:20715500-20716500	0.920	CG15450;CG1314
chrX:20716000-20717000	0.925	CG15450;CG1314
chrX:21200000-21201000	0.921	CG1486;TTY
chrX:22396600-22397600	0.921	STNA;STNB;

REFERENCES

Abeel, T., Saeys, Y., Bonnet, E., Rouze, P., and Van de Peer, Y. (2008). "Generic eukaryotic core promoter prediction using structural features of DNA." *Genome Res*, 18(2), 310-323.

Abnizova, I., te Boekhorst, R., Walter, K., and Gilks, W.R. (2005). "Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the Drosophila genome: the fluffy-tail test." *BMC Bioinformatics*, 6, 109.

Aerts, S., Van Loo, P., Thijs, G., Moreau, Y., and De Moor, B. (2003). "Computational detection of cis -regulatory modules." *Bioinformatics*, **19** Suppl 2, ii5-14.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). "Basic local alignment search tool." *J Mol Biol*, **215**(3), 403-410.

Arbeitman, M.N., Furlong, E.E., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., and White, K.P. (2002). "Gene expression during the life cycle of Drosophila melanogaster." *Science*, **297**(5590), 2270-2275.

Arnone, M.I., and Davidson, E.H. (1997). "The hardwiring of development: organization and function of genomic regulatory systems." *Development (Cambridge, England)*, **124**(10), 1851-1864.

Audic, S., and Claverie, J.M. (1997). "Detection of eukaryotic promoters using Markov transition matrices." *Comput Chem*, **21**(4), 223-227.

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

Bailey, T.L., and Noble, W.S. (2003). "Searching for statistically significant regulatory modules." *Bioinformatics (Oxford, England)*, **19 Suppl 2**, II16-II25.

Bajic, V.B., Brent, M.R., Brown, R.H., Frankish, A., Harrow, J., Ohler, U., Solovyev, V.V., and Tan, S.L. (2006). "Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment." Genome biology, 7 Suppl 1, S3 1-13.

Bajic, V.B., Choudhary, V., and Hock, C.K. (2004). "Content analysis of the core promoter region of human genes." *In Silico Biol*, **4**(2), 109-125.

Bajic, V.B., Tan, S.L., Suzuki, Y., and Sugano, S. (2004). "Promoter prediction analysis on the whole human genome." *Nat Biotechnol*, **22**(11), 1467-1473.

Bajic, V.B., and Seah, S.H. (2003). "Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units." *Genome Res*, **13**(8), 1923-1929.

Bajic, V.B., Seah, S.H., Chong, A., Krishnan, S.P., Koh, J.L., and Brusic, V. (2003). "Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates." *J Mol Graph Model*, **21**(5), 323-332.

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). "Modeling dependencies in protein-dna binding sites." in *Proc. of the 7th RECOMB Conference*.

Barber, T.D., Barber, M.C., Cloutier, T.E., and Friedman, T.B. (1999). "PAX3 gene structure, alternative splicing and evolution." *Gene*, **237**(2), 311-319.

Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., Posch, S., and Grosse, I. (2005). “Identification of transcription factor binding sites with variable-order Bayesian networks.” *Bioinformatics (Oxford, England)*, **21**(11), 2657-2666.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L. (2002). “GenBank.” *Nucleic Acids Res*, **30**(1), 17-20.

Bergman, C.M., Carlson, J.W., and Celniker, S.E. (2005). “Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster.” *Bioinformatics (Oxford, England)*, **21**(8), 1747-1749.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. (2002). “Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.” *Proceedings of the National Academy of Sciences of the United States of America*, **99**(2), 757-762.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. (2004). “Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura.” *Genome Biol*, **5**(9), R61.

Blanchette, M., Schwikowski, B., and Tompa, M. (2002). “Algorithms for phylogenetic footprinting.” *J Comput Biol*, **9**(2), 211-223.

Borghese, L., Fletcher, G., Mathieu, J., Atzberger, A., Eades, W.C., Cagan, R.L., and Rorth, P. (2006). “Systematic analysis of the transcriptional switch inducing migration of border cells.” *Dev Cell*, **10**(4), 497-508.

Borkowski, O.M., Brown, N.H., and Bate, M. (1995). “Anterior-posterior subdivision and the diversification of the mesoderm in Drosophila.” *Development*, **121**(12), 4183-4193.

Brenowitz, M., Senear, D.F., Shea, M.A., and Ackers, G.K. (1986). “Quantitative DNase footprint titration: a method for studying protein-DNA interactions.” *Methods Enzymol*, **130**, 132-181.

Brocchieri, L., and Karlin, S. (1998). “A symmetric-iterated multiple alignment of protein sequences.” *J Mol Biol*, **276**(1), 249-264.

Bucher, P. (1990). “Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences.” *J Mol Biol*, **212**(4), 563-578.

Buhler, J., and Tompa, M. (2002). “Finding motifs using random projections.” *J Comput Biol*, **9**(2), 225-242.

Bussemaker, H.J., Li, H., and Siggia, E.D. (2000). “Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis.” *Proc Natl Acad Sci U S A*, **97**(18), 10096-10100.

Butler, M.J., Jacobsen, T.L., Cain, D.M., Jarman, M.G., Hubank, M., Whittle, J.R., Phillips, R., and Simcox, A. (2003). “Discovery of genes with highly restricted expression patterns in the Drosophila wing disc using DNA oligonucleotide microarrays.” *Development*, **130**(4), 659-670.

Carlin, B.P., and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall, Florida.

Carroll, J.S., Liu, X.S., Brodsky, A.S., Li, W., Meyer, C.A., Szary, A.J., Eeckhoute, J., Shao, W., Hestermann, E.V., Geistlinger, T.R., Fox, E.A., Silver, P.A., and Brown, M. (2005). "Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1." *Cell*, **122**(1), 33-43.

Chan, B.Y., and Kibler, D. (2005). "Using hexamers to predict cis-regulatory motifs in Drosophila." *BMC Bioinformatics*, **6**, 262.

Chen, Q.K., Hertz, G.Z., and Stormo, G.D. (1997). "PromFD 1.0: a computer program that predicts eukaryotic pol II promoters using strings and IMD matrices." *Comput Appl Biosci*, **13**(1), 29-35.

Chin, F.Y.L., Leung, H.C.M., Yiu, S.M., Lam, T.W., Rosenfeld, R., Tsang, W.W., Smith, D.K., Jiang, Y. (2004). "Finding motifs for insufficient number of sequences with strong binding to transcription factor." *Proc. RECOMB 2004*, 125-132.

Collins, J.E., Goward, M.E., Cole, C.G., Smink, L.J., Huckle, E.J., Knowles, S., Bye, J.M., Beare, D.M., and Dunham, I. (2003). "Reevaluating human gene annotation: a second-generation analysis of chromosome 22." *Genome Res*, **13**(1), 27-36.

Crowley, E.M., Roeder, K., and Bina, M. (1997). "A statistical model for locating regulatory regions in genomic DNA." *J Mol Biol*, **268**(1), 8-14.

Davuluri, R.V., Grosse, I., and Zhang, M.Q. (2001). "Computational identification of promoters and first exons in the human genome." *Nat Genet*, **29**(4), 412-417.

D'Haeseleer, P. (2006a). "How does DNA sequence motif discovery work?" *Nat Biotechnol*, **24**(8), 959-961.

D'Haeseleer, P. (2006b). "What are DNA sequence motifs?" *Nat Biotechnol*, **24**(4), 423-425.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Rust, A.G., Pan, Z., Schilstra, M.J., Clarke, P.J., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L., and Bolouri, H. (2002). "A genomic regulatory network for development." *Science*, **295**(5560), 1669-1678.

Day, W.H., and McMorris, F.R. (1992). "Critical comparison of consensus methods for molecular sequences." *Nucleic Acids Res*, **20**(5), 1093-1099.

de Celis, J.F., Llimargas, M., and Casanova, J. (1995). "Ventral veinless, the gene encoding the Cf1a transcription factor, links positional information and cell differentiation during embryonic and imaginal development in Drosophila melanogaster." *Development*, **121**(10), 3405-3416.

Domingos, P., and Pazzani, M. (1996). "Beyond independence: conditions for the optimality of the simple Bayesian classifier." *Int Conf Machine Learning, Bari, Italy*, 105-112.

Down, T.A., and Hubbard, T.J. (2002). "Computational detection and location of transcription start sites in mammalian genomic DNA." *Genome Res*, **12**(3), 458-461.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge University Press.

Encode (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project." *Science*, **306**(5696), 636-640.

Eskin, E., and Pevzner, P.A. (2002). "Finding composite regulatory patterns in DNA sequences." *Bioinformatics*, **18 Suppl 1**, S354-363.

Ettwiller, L., Paten, B., Ramialison, M., Birney, E., and Wittbrodt, J. (2007). “Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation.” *Nat Methods*, **4**(7), 563-565.

Euskirchen, G., and Snyder, M. (2004). “A plethora of sites.” *Nat Genet*, **36**(4), 325-326.

Favorov, A.V., Gelfand, M.S., Gerasimova, A.V., Ravcheev, D.A., Mironov, A.A., and Makeev, V.J. (2005). “A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length.” *Bioinformatics* (Oxford, England), **21**(10), 2240-2245.

Fickett, J.W., and Hatzigeorgiou, A.G. (1997). “Eukaryotic promoter recognition.” *Genome Res*, **7**(9), 861-878.

Fratkin, E., Naughton, B.T., Brutlag, D.L., and Batzoglou, S. (2006). “MotifCut: regulatory motifs finding with maximum density subgraphs.” *Bioinformatics*, **22**(14), e150-157.

Frazier, M., Thomassen, D., Patrinos, A., Johnson, G., Oliver, C.E., and Uberbacher, E. (2003). “Stepping up the pace of discovery: the genomes to life program.” *Proc IEEE Comput Soc Bioinform Conf*, **2**, 2-9.

Frech, K., Danescu-Mayer, J., and Werner, T. (1997). “A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter.” *J Mol Biol*, **270**(5), 674-687.

Friberg, M., von Rohr, P., and Gonnet, G. (2005). “Scoring functions for transcription factor binding site prediction.” *BMC bioinformatics*, **6**, 84.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). “Bayesian network classifiers.” *Machine Learning*, **29**, 131-163.

Frith, M.C., Hansen, U., and Weng, Z. (2001). “Detection of cis-element clusters in higher eukaryotic DNA.” *Bioinformatics (Oxford, England)*, **17**(10), 878-889.

Furlong, E.E., Andersen, E.C., Null, B., White, K.P., and Scott, M.P. (2001). “Patterns of gene expression during Drosophila mesoderm development.” *Science*, **293**(5535), 1629-1633.

Furlong, E.E. (2004). “Integrating transcriptional and signalling networks during muscle development.” *Curr Opin Genet Dev*, **14**(4), 343-350.

Gallo, S.M., Li, L., Hu, Z., and Halfon, M.S. (2006). “REDfly: a Regulatory Element Database for Drosophila.” Bioinformatics (Oxford, England), 22(3), 381-383.

Ganguly, A., Jiang, J., and Ip, Y.T. (2005). “Drosophila WntD is a target and an inhibitor of the Dorsal/Twist/Snail network in the gastrulating embryo.” *Development*, **132**(15), 3419-3429.

Guigo, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., Castelo, R., Eyras, E., Ucla, C., Gingeras, T.R., Harrow, J., Hubbard, T., Lewis, S.E., and Reese, M.G. (2006). “EGASP: the human ENCODE Genome Annotation Assessment Project.” *Genome Biol*, **7 Suppl 1**, S2 1-31.

Gupta, M., and Liu, J.S. (2005). “De novo cis-regulatory module elicitation for eukaryotic genomes.” *Proceedings of the National Academy of Sciences of the United States of America*, **102**(20), 7079-7084.

Hannenhalli, S., and Levy, S. (2001). “Promoter prediction in the human genome.” *Bioinformatics*, **17 Suppl 1**, S90-96.

Harley, C.B., and Reynolds, R.P. (1987). “Analysis of E. coli promoter sequences.” *Nucleic Acids Res*, **15**(5), 2343-2361.

Hertz, G.Z., Hartzell, G.W., 3rd, and Stormo, G.D. (1990). “Identification of consensus patterns in unaligned DNA sequences known to be functionally related.” *Comput Appl Biosci*, **6**(2), 81-92.

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

Hirose, F., Ohshima, N., Shiraki, M., Inoue, Y.H., Taguchi, O., Nishi, Y., Matsukage, A., and Yamaguchi, M. (2001). “Ectopic expression of DREF induces DNA synthesis, apoptosis, and unusual morphogenesis in the Drosophila eye imaginal disc: possible interaction with Polycomb and trithorax group proteins.” *Mol Cell Biol*, **21**(21), 7231-7242.

Hutchinson, G.B. (1996). “The prediction of vertebrate promoter regions using differential hexamer frequency analysis.” *Comput Appl Biosci*, **12**(5), 391-398.

Jensen, F.V. (2001). *Bayesian Networks and Decision Graphs*, Springer Verlag, New York.

Kaplan, T., Friedman, N., and Margalit, H. (2005). “Ab initio prediction of transcription factor targets using structural knowledge.” *PLoS Comput Biol*, **1**(1), e1.

Keich, U., and Pevzner, P.A. (2002a). “Finding motifs in the twilight zone.” *Bioinformatics*, **18**(10), 1374-1381.

Keich, U., and Pevzner, P.A. (2002b). “Subtle motifs: defining the limits of motif finding algorithms.” *Bioinformatics*, **18**(10), 1382-1390.

Kel-Margoulis, O.V., Kel, A.E., Reuter, I., Deineko, I.V., and Wingender, E. (2002). “TRANSCompel: a database on composite regulatory elements in eukaryotic genes.” *Nucleic Acids Res*, **30**(1), 332-334.

Klingenhoff, A., Frech, K., Quandt, K., and Werner, T. (1999). “Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity.” *Bioinformatics*, **15**(3), 180-186.

Kondrakhin, Y.V., Kel, A.E., Kolchanov, N.A., Romashchenko, A.G., and Milanesi, L. (1995). “Eukaryotic promoter recognition by binding sites for transcription factors.” *Comput Appl Biosci*, **11**(5), 477-488.

Kusch, T., and Reuter, R. (1999). “Functions for Drosophila brachyenteron and forkhead in mesoderm specification and cell signalling.” *Development*, **126**(18), 3991-4003.

Latchman, D.S. (2003). *Eukaryotic transcription factors*, Academic Press, London.

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

Lecuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T.R., Tomancak, P., and Krause, H.M. (2007). “Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function.” *Cell*, **131**(1), 174-187.

Li, N., and Tompa, M. (2006). “Analysis of computational approaches for motif discovery.” *Algorithms for molecular biology*, **1**, 8.

Li, L., Zhu, Q., He, X., Sinha, S., and Halfon, M.S. (2007). “Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses.” *Genome Biol*, **8**(6), R101.

Lifanov, A.P., Makeev, V.J., Nazina, A.G., and Papatsenko, D.A. (2003). “Homotypic regulatory clusters in Drosophila.” *Genome research*, **13**(4), 579-588.

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.

Macina, R.A., Barr, F.G., Galili, N., and Riethman, H.C. (1995). “Genomic organization of the human PAX3 gene: DNA sequence analysis of the region disrupted in alveolar rhabdomyosarcoma.” *Genomics*, **26**(1), 1-8.

Mahony, S., and Benos, P.V. (2007). “STAMP: a web tool for exploring DNA-binding motif similarities.” *Nucleic Acids Res*, **35**(Web Server issue), W253-258.

Mandel-Gutfreund, Y., Baron, A., and Margalit, H. (2001). “A structure-based approach for prediction of protein binding sites in gene upstream regions.” *Pac Symp Biocomput*, 139-150.

Mann, R.S., and Morata, G. (2000). “The developmental and molecular biology of genes that subdivide the body of Drosophila.” *Annu Rev Cell Dev Biol*, **16**, 243-271.

Marchal, K., Thijs, G., De Keersmaecker, S., Monsieurs, P., De Moor, B., and Vanderleyden, J. (2003). “Genome-specific higher-order background models to improve motif detection.” *Trends Microbiol*, **11**(2), 61-66.

Markstein, M., Markstein, P., Markstein, V., and Levine, M.S. (2002). “Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo.” *Proceedings of the National Academy of Sciences of the United States of America*, **99**(2), 763-768.

Markstein, M., Zinzen, R., Markstein, P., Yee, K.P., Erives, A., Stathopoulos, A., and Levine, M. (2004). “A regulatory code for neurogenic gene expression in the Drosophila embryo.” *Development*, **131**(10), 2387-2394.

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

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., and Jacq, B. (2004). “GOToolBox: functional analysis of gene datasets based on Gene Ontology.” *Genome Biol*, **5**(12), R101.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). “TRANSFAC: transcriptional regulation, from patterns to profiles.” *Nucleic Acids Res*, **31**(1), 374-378.

Molina, C., and Grotewold, E. (2005). “Genome wide analysis of Arabidopsis core promoters.” *BMC Genomics*, **6**(1), 25.

Morata, G. (2001). “How Drosophila appendages develop.” *Nature reviews*, **2**(2), 89-97.

Narang, V., Sung, W.K., and Mittal, A. (2005). “Computational modeling of oligonucleotide positional densities for human promoter prediction.” *Artificial Intelligence in Medicine*, **35**(1-2), 107-119.

Narang, V., Sung, W.K., and Mittal, A. (2006). “Bayesian network modeling of transcription factor binding sites.” in: *Bayesian Network Technologies: Applications and Graphical Models*, A. Mittal and A. Kassim, eds., Idea Group Publishing, Pennsylvania, USA.

Narang, V., Sung, W.K., and Mittal, A. “LocalMotif - an in silico tool for detecting localized motifs in regulatory sequences.” *18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2006), Washington D.C.,USA, November 13-15, 2006.*, 791-799.

Narang, V., Sung, W.K., and Mittal, A. (2006). “Computational annotation of transcription factor binding sites in D. melanogaster developmental genes.” *Genome Informatics*, **17**(2), 14-24.

Narang, V., Sung, W.K., and Mittal, A. (2007). “Localized motif discovery in metazoan regulatory sequences.” *Under submission*.

Narang, V., Sung, W.K., and Mittal, A. (2008). “Probabilistic graphical modeling of cis-regulatory codes govering Drosophila development.” *Under submission*.

Neuwald, A.F., Liu, J.S., Lipman, D.J., and Lawrence, C.E. (1997). “Extracting protein alignment models from the sequence database.” *Nucleic Acids Res*, **25**(9), 1665-1677.

Ochoa-Espinosa, A., Yucel, G., Kaplan, L., Pare, A., Pura, N., Oberstein, A., Papatsenko, D., and Small, S. (2005). “The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila.” *Proc Natl Acad Sci U S A*, **102**(14), 4960-4965.

Ohler, U., Harbeck, S., Niemann, H., Noth, E., and Reese, M.G. (1999). “Interpolated markov chains for eukaryotic promoter recognition.” *Bioinformatics*, **15**(5), 362-369.

Ohler, U., Liao, G.C., Niemann, H., and Rubin, G.M. (2002). “Computational analysis of core promoters in the Drosophila genome.” *Genome Biol*, **3**(12), RESEARCH0087.

Okladnova, O., Syagailo, Y.V., Tranitz, M., Riederer, P., Stober, G., Mossner, R., and Lesch, K.P. (1999). “Functional characterization of the human PAX3 gene regulatory region.” *Genomics*, **57**(1), 110-119.

Pavesi, G., Mauri, G., and Pesole, G. (2001). “An algorithm for finding signals of unknown length in DNA sequences.” *Bioinformatics*, **17 Suppl 1**, S207-214.

Pedersen, A.G., Baldi, P., Chauvin, Y., and Brunak, S. (1999). “The biology of eukaryotic promoter prediction--a review.” *Comput Chem*, **23**(3-4), 191-207.

Pevzner, P.A., Borodovsky, M., and Mironov, A.A. (1989). “Linguistics of nucleotide sequences I: the significance of deviations from the mean statistical characteristics and prediction of the frequencies of occurrence of words.” *J Biomol Struct Dyn*, **6**(5), 1013-1026.

Pilot, F., Philippe, J.M., Lemmers, C., Chauvin, J.P., and Lecuit, T. (2006). “Developmental control of nuclear morphogenesis and anchoring by charleston, identified in a functional genomic screen of Drosophila cellularisation.” *Development*, **133**(4), 711-723.

Prestridge, D.S. (1995). “Predicting Pol II promoter sequences using transcription factor binding sites.” *J Mol Biol*, **249**(5), 923-932.

Prokop, A., Bray, S., Harrison, E., and Technau, G.M. (1998). “Homeotic regulation of segment-specific differences in neuroblast numbers and proliferation in the Drosophila central nervous system.” *Mech Dev*, **74**(1-2), 99-110.

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E.D. (2002). “Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo.” *BMC bioinformatics*, **3**, 30.

Rice, J.A. (1995). *Mathematical Statistics and Data Analysis*, Duxbury Press.

Roepcke, S., Zhi, D., Vingron, M., and Arndt, P.F. (2006). "Identification of highly specific localized sequence motifs in human ribosomal protein gene promoters." *Gene*, 365, 48-56.

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P., and van de Peer, Y. (2003). "Computational approaches to identify promoters and cis-regulatory elements in plant genomes." *Plant physiology*, 132(3), 1162-1176.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." *Nat Biotechnol*, **16**(10), 939-945.

Rushlow, C., Colosimo, P.F., Lin, M.C., Xu, M., and Kirov, N. (2001). "Transcriptional regulation of the Drosophila gene zen by competing Smad and Brinker inputs." *Genes Dev*, 15(3), 340-351.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." *Nucleic Acids Res*, **32**(Database issue), D91-94.

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

Schmid, C.D., Praz, V., Delorenzi, M., Perier, R., and Bucher, P. (2004). "The Eukaryotic Promoter Database EPD: the impact of in silico primer extension." *Nucleic Acids Res*, **32**(Database issue), D82-85.

Schroeder, M.D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E.D., and Gaul, U. (2004). "Transcriptional control in the segmentation gene network of Drosophila." *PLoS biology*, **2**(9), E271.

Segal, E., and Sharan, R. (2005). "A discriminative model for identifying spatial cis-regulatory modules." *Journal of computational biology*, **12**(6), 822-834.

Sharan, R., Ben-Hur, A., Loots, G.G., and Ovcharenko, I. (2004). "CREME: Cis-Regulatory Module Explorer for the human genome." *Nucleic Acids Res*, **32**(Web Server issue), W253-256.

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.

Sinha, S., and Tompa, M. (2003). "YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation." *Nucleic Acids Res*, **31**(13), 3586-3588.

Sinha, S., van Nimwegen, E., and Siggia, E.D. (2003). "A probabilistic method to detect regulatory modules." *Bioinformatics (Oxford, England)*, **19 Suppl 1**, i292-301.

Skeath, J.B., and Thor, S. (2003). "Genetic control of Drosophila nerve cord development." *Curr Opin Neurobiol*, **13**(1), 8-15.

Smale, S.T., and Kadonaga, J.T. (2003). "The RNA polymerase II core promoter." *Annu Rev Biochem*, **72**, 449-479.

Sokal, R.R., and Michener, C.D. (1958). "A statistical method for evaluating systematic relationships." *Univ. Kansas Sci. Bull.*, **38**, 1409-1438.

Song, X., Wong, M.D., Kawase, E., Xi, R., Ding, B.C., McCarthy, J.J., and Xie, T. (2004). "Bmp signals from niche cells directly repress transcription of a differentiation-promoting gene, bag of marbles, in germline stem cells in the Drosophila ovary." *Development*, **131**(6), 1353-1364.

Staden, R. (1989). "Methods for discovering novel motifs in nucleic acid sequences." *Comput Appl Biosci*, **5**(4), 293-298.

Stein, L. (2001). "Genome annotation: from sequence to biology." *Nature reviews*, **2**(7), 493-503.

Stormo, G.D., Schneider, T.D., and Gold, L.M. (1982). "Characterization of translational initiation sites in E. coli." *Nucleic Acids Res*, **10**(9), 2971-2996.

Stormo, G.D. (2000). "DNA binding sites: representation and discovery." *Bioinformatics*, **16**(1), 16-23.

Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., Suyama, A., Sakaki, Y., Morishita, S., Okubo, K., and Sugano, S. (2001). "Identification and characterization of the potential promoter regions of 1031 kinds of human genes." *Genome Res*, **11**(5), 677-684.

Tharakaraman, K., Marino-Ramirez, L., Sheetlin, S., Landsman, D., and Spouge, J.L. (2005). "Alignments anchored on genomic landmarks can aid in the identification of regulatory elements." *Bioinformatics (Oxford, England)*, **21 Suppl 1**, i440-448.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2001). "A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling." *Bioinformatics*, **17**(12), 1113-1122.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2002). "A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes." *J Comput Biol*, **9**(2), 447-464.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acids Res*, **22**(22), 4673-4680.

Tomancak, P., Berman, B.P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., Celniker, S.E., and Rubin, G.M. (2007). "Global analysis of patterns of gene expression during Drosophila embryogenesis." *Genome Biol*, **8**(7), R145.

Tompa, M. (1999). "An exact method for finding short motifs in sequences, with application to the ribosome binding site problem." *Proc Int Conf Intell Syst Mol Biol*, 262-271.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., 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.

Ukkonen, E. (1995). "On-line construction of suffix trees." *Algorithmica*, **14**(3), 249-260.

Uren, P., Cameron-Jones, M., and Sale, A. (2006). "Promoter prediction using physico-chemical properties of DNA." *Lect Notes Comput Sci*, **4216**: 21–31.

van Helden, J., Andre, B., and Collado-Vides, J. (1998). "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies." *J Mol Biol*, **281**(5), 827-842.

van Helden, J., Rios, A.F., and Collado-Vides, J. (2000). "Discovering regulatory elements in non-coding sequences by analysis of spaced dyads." *Nucleic Acids Res*, **28**(8), 1808-1818.

Veraksa, A., Kennison, J., and McGinnis, W. (2002). "DEAF-1 function is essential for the early embryonic development of Drosophila." Genesis, 33(2), 67-76.

Verbeek, J.J., Vlassis, N., and Kraose, B. (2003). "Efficient greedy learning of Gaussian mixture models." *Neural Computation*, **15**, 469-485.

Vomlel, J. (2002). "Exploiting Functional Dependence in Bayesian Network Inference." *Proceedings of The 18th Conference on Uncertainty in Artificial Intelligence (UAI 2002)*, University of Alberta, Edmonton, Canada, 528-535.

Vomlel, J. (2006). "Noisy-or classifier." *International Journal of Intelligent Systems*, **21**(3), 381-398.

Wang, J., Han, J., and Pei, J. (2003). "CLOSET+: searching for the best strategies for mining frequent closed itemsets." *Proc. 9th ACM SIGKDD*, *ACM*, Washington, D.C., 236-245.

Wasserman, W.W., and Fickett, J.W. (1998). "Identification of regulatory regions which confer muscle-specific gene expression." *Journal of molecular biology*, **278**(1), 167-181.

Wasserman, W.W., and Krivan, W. (2003). "In silico identification of metazoan transcriptional regulatory regions." *Naturwissenschaften*, **90**(4), 156-166.

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

Werner, T. (1999). "Models for prediction and recognition of eukaryotic promoters." *Mamm Genome*, **10**(2), 168-175.

Werner, T. (2003). "The state of the art of mammalian promoter recognition." *Brief Bioinform*, **4**(1), 22-30.

Wijaya, E., Rajaraman, K., Yiu, S.M., and Sung, W.K. (2007). "Detection of generic spaced motifs using submotif pattern mining." *Bioinformatics*, **23**(12), 1476-1485.

Wilson, R.J., Goodman, J.L., and Strelets, V.B. (2008). "FlyBase: integration and improvements to query tools." *Nucleic Acids Res*, **36**(Database issue), D588-593.

Workman, C.T., and Stormo, G.D. (2000). "ANN-Spec: a method for discovering transcription factor binding sites with improved specificity." *Pac Symp Biocomput*, 467-478.

Xing, E.P., Wu, W., Jordan, M.I., and Karp, R.M. (2004). "Logos: a modular bayesian model for de novo motif detection." *J Bioinform Comput Biol*, **2**(1), 127-154.

Zhang, C.C., Muller, J., Hoch, M., Jackle, H., and Bienz, M. (1991). "Target sequences for hunchback in a control region conferring Ultrabithorax expression boundaries." *Development* (Cambridge, England), **113**(4), 1171-1179.

Zhang, M.Q. (2002). "Computational methods for promotor recognition" in *Current topics in computational molecular biology*, T. Jiang, Y. Xu, and M.Q. Zhang, eds., MIT Press, Cambridge, Massachusetts, 249-268.