

Prediction and Analysis of Nucleosome Positioning in Genomic Sequences

Samiul Hasan

A dissertation submitted to the University of Cambridge for the
degree of Doctor of Philosophy

April 2003

Wolfson College, University of Cambridge

and

The Wellcome Trust Sanger Institute,

Hinxton, Cambridgeshire

Abstract

A nucleosome is the resultant structure formed when 1.6 left-handed turns of DNA (~146 bp) are wound around a basic complex of histone proteins (the histone octamer). Nucleosomes occur naturally and ubiquitously in all eukaryotic genomes; the histone proteins themselves are highly conserved in eukaryotes. Experimental evidence suggests that specific DNA sequences may exhibit high or low nucleosome-forming tendencies compared to random DNA. This could mean that nucleosomes, whose positions are influenced by the underlying DNA sequence, can in turn govern the accessibility of regulatory DNA sequences such as transcription initiation and replication origin sites. This forms the need to search for evidence of nucleosome positioning and consequently build models to predict and investigate such locations.

One theory suggests that DNA sequences, which are intrinsically “curved”, can position nucleosomes. In a previous study, using “cyclical” hidden-markov models, it had been suggested that a 10 periodic occurrence of the [VWG] motif could have such an effect and could help nucleosomes to be positioned in human exons. This work was extended in this thesis. 60% of human genomic sequences were seen to be covered in apparently weak 9-10 bp periodic patches of [CWG]. [CWG]-dense regions were seen to alternate with regions which were rich in [W] motifs in human. However, the pattern was not the same in mouse.

Another theory suggests that highly flexible or highly rigid DNA sequences may favour or disfavour nucleosome formation respectively. The locations of such patterns were investigated in human sequences using the wavelet technique. This approach identified confined periodic patterns (in the range of 80-200 bp) of rigidity in human genomic sequences; the patterns themselves were, however, mainly consequences of alu repeat-clustering. However, the same analysis in the mouse genome indicated that such a mechanism for positioning nucleosomes was not conserved and therefore unlikely.

A different approach to model nucleosomes was to train weighted DNA matrices from experimentally-mapped nucleosome datasets. This technique gave some encouraging results (one model showing 100% accuracy at 40% coverage), but was restricted by the limited size of the datasets.

Overall the conclusion is that there is some evidence for sequence specific nucleosome positioning, but that more experimental data is needed to build and evaluate practical and predictive computational models.

Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

The work in this thesis is not substantially the same as any I have submitted for a degree or diploma or other qualification at any other University.

Samiul Hasan,

14/07/2003

Acknowledgements

I would firstly like to thank my supervisor, Dr Tim Hubbard, for giving me the opportunity to carry out my PhD research at the Sanger Institute. The nucleosome prediction problem was challenging and I definitely needed all the help I got. I am grateful for all his invaluable advice and for all that I have learnt as a result. I am also indebted to Thomas Down and Matthew Pocock for their courteous advice and help, especially for solving a number of difficult programming issues in my project. Thanks also to Aroul Selvam Ramadass, Raphael Leplae and Yen-Hua Huang for their generous help and for frequent discussions relating to this project.

A special thanks to Dr. Akilesh Eswaran for helping me with computer-related problems on numerous occasions. I am also grateful to Jason Dowling, Diego Di Bernardo, Arnaud Kerhornou, Arundhati S Ghosh, Mahesh Menon, Ashish Tuteja, Maushumi Guha, Giota Mitrou, Andreas Heger, Kevin Howe, the Arrelanos, Visalakshi Kannan, Antonella Ferrecchia, Gabriella Rustici, Ranjeeva Ranasingh, Waqas Awan, Jayne Vallance and to the many other friends I made during my time at Cambridge for their friendship and courtesy and for helping me to enjoy my time here even more. Finally, I would also like to express my gratitude to my family in particular to my parents, my brother, my uncles Tariq and Safiul, Dr Karim, Fariha, Sofia and Mowli for their kind words and acts of encouragement during this time.

Ambiguity Codes for DNA as specified by the Convention of the International Union of Pure and Applied Chemistry¹

IUPAC Code	Meaning	Complement
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
N	G or A or T or C	N

¹ Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* 13, 3021-30.

Prediction and Analysis of Nucleosome Positioning in Genomic Sequences

1 A GENERAL INTRODUCTION TO NUCLEOSOMES AND NUCLEOSOME POSITIONING	1-1
1.1 NUCLEOSOMES: THE BUILDING BLOCKS OF CHROMATIN	1-2
1.2 DNA-PROTEIN INTERACTIONS IN THE NUCLEOSOME CORE PARTICLE	1-5
1.3 THE CONCEPT OF NUCLEOSOME POSITIONING	1-7
1.4 AN INTRODUCTION TO NUCLEOSOME ROTATIONAL POSITIONING	1-8
1.4.1 Intrinsic DNA curvature: Bending based on 10-phased [A] tracts	1-8
1.4.2 Intrinsic DNA curvature and the initial assessment of nucleosome rotational positioning	1-10
1.4.3 Further evidence to support nucleosome rotational positioning	1-12
1.4.4 Nucleosome rotational positioning and DNA regulatory regions	1-13
1.5 AN INTRODUCTION TO NUCLEOSOME TRANSLATIONAL POSITIONING	1-16
1.5.1 DNA sequences that repel nucleosome formation	1-16
1.5.2 DNA sequences that favour nucleosome formation	1-19
1.5.3 Nucleosome translational positioning and DNA regulatory regions	1-20
1.6 REGIONS OF PHASED NUCLEOSOMES	1-22
1.7 STRENGTH OF NUCLEOSOME POSITIONING SEQUENCES IN VIVO	1-23
1.8 EXPERIMENTALLY MAPPED NUCLEOSOME DATASETS	1-24
1.8.1 Database of chicken core DNA sequences	1-24
1.8.2 Nucleosome database from mapping studies on various species	1-25
1.9 COMPUTATIONAL APPROACHES TO UNDERSTANDING NUCLEOSOME POSITIONING IN OTHER LABORATORIES	1-29
1.9.1 Using DNA structural parameters to predict nucleosome positioning	1-29
1.9.2 [AA/TT] rotational positioning pattern obtained using multiple sequence alignment	1-30
1.9.3 10-periodic [VWG] pattern obtained using hidden markov models	1-31
1.9.4 RECON: A nucleosome prediction model based on dinucleotide relative abundance distance	1-33
1.10 SUMMARY OF AIMS	1-34
1.10.1 The scope for studying nucleosome positioning	1-34
1.10.2 Aims and benefits of predicting nucleosome positioning	1-35

1.11	<i>APPROACHES PROPOSED FOR MODELLING NUCLEOSOME POSITIONING</i>	1-38
1.11.1	Potential for studying 10 bp-phased motifs.....	1-38
1.11.2	Potential for studying nucleosome translational positioning	1-39
1.11.3	Using DNA weight matrices to model the existing nucleosome datasets.....	1-39
2	GENERAL INTRODUCTION TO COMPUTATIONAL METHODS USED IN THIS THESIS	2-41
2.1	<i>THE APPLICATION OF BAYESIAN METHODS IN SEQUENCE ANALYSIS</i>	2-42
2.2	<i>HIDDEN MARKOV MODEL THEORY</i>	2-44
2.2.1	A general introduction to hidden markov models.....	2-44
2.2.2	Predicting the most likely path of a HMM through a sequence using the Viterbi algorithm.....	2-47
2.2.3	Training a HMM using the Baum Welch algorithm	2-48
2.3	<i>THE USE OF FLEXIBILITY SEQUENCES</i>	2-51
2.3.1	An Introduction to flexibility sequences.....	2-51
2.3.2	Flexibility emission alphabet for using with HMMs	2-53
2.4	<i>A BASIC INTRODUCTION TO WAVELETS</i>	2-54
2.4.1	An introduction to wavelets	2-54
2.4.2	The multiresolution property of wavelets	2-57
3	CYCLICAL HIDDEN MARKOV MODEL ANALYSIS TO FIND SIGNALS INVOLVED IN NUCLEOSOME ROTATIONAL POSITIONING	3-58
3.1	<i>INTRODUCTION</i>	3-59
3.2	<i>METHODS</i>	3-62
3.2.1	Construction of different kinds of wheel architecture.....	3-63
3.2.2	Parameter setups in preparation for cyclical HMM training	3-65
3.2.3	Model training.....	3-66
3.2.4	Construction of emission alphabets other than DNA.....	3-67
3.2.5	Datasets of training sequences.....	3-68
3.2.6	Viterbi labelling analysis.....	3-69
3.2.7	Analysis of a model's "wheel"-labelling pattern.....	3-69
3.2.8	Labelling analysis of chicken nucleosome sequences and chicken genomic sequences.....	3-70
3.2.9	Estimation of frequently "wheel-state"-labelled features.....	3-70
3.2.10	Visualisation of predictions against genomic annotations	3-71
3.3	<i>RESULTS AND DISCUSSION</i>	3-72

3.3.1	Model-training experiences using different kinds of cyclical HMM architectures....	3-72
3.3.2	Experiences of using non-DNA emission alphabets with cyclical HMMs	3-78
3.3.3	An initial test to investigate if the B&B model had learnt codon bias	3-80
3.3.4	The [VWG] motif in retrospect and the distinction of two apparent motifs learnt in F3 human models	3-83
3.3.5	F3 model training results from Archaea and the 2 nucleosome datasets.....	3-99
3.3.6	Labelling analysis of chicken nucleosome sequences and chicken genomic sequences	3-102
3.3.7	Analysis of periodicity of the two opposing motifs	3-105
3.3.8	Labelling density of [CWG]-learnt models.....	3-110
3.4	CONCLUSION	3-114

4 PERIODIC FLEXIBILITY PATTERNS IN DNA: A SCAN FOR SIGNALS INVOLVED IN NUCLEOSOME TRANSLATIONAL POSITIONING **4-115**

4.1 INTRODUCTION

4.2 METHODS

4.2.1	Construction of flexibility sequences.....	4-117
4.2.2	Wavelet transform of whole chromosomal flexibility sequences.....	4-117
4.2.3	Thresholding by wavelet co-efficient strengths	4-119
4.2.4	Probability distribution of periodic flexibility patterns	4-119
4.2.5	Estimation of genomic features frequently associated with periodic flexibility patterns	4-120
4.2.6	Alignment of flexibility sequences	4-120

4.3 RESULTS AND DISCUSSION

4.3.1	Differences in wavelet spectra between eukaryotic and prokaryotic flexibility sequences.....	4-121
4.3.2	Whole chromosomal flexibility landscape in higher eukaryotic chromosomes	4-124
4.3.3	Genomic features frequently associated with strongly periodic flexibility patterns .	4-127
4.3.4	Why Alu repeats were frequently associated with periodic flexibility patterns.....	4-130
4.3.5	Conservation of periodic flexibility patterns in eukaryotic genomes	4-133
4.3.6	Re-examination of the hypothesis of nucleosome translational positioning with respect to Alu repeats	4-135

4.4 CONCLUSION

5	MODELLING DNA SEQUENCE MOTIFS FROM KNOWN NUCLEOSOME DATASETS	5-140
5.1	INTRODUCTION	5-141
5.1.1	The <i>Eponine</i> Tool	5-141
5.2	METHODS	5-144
5.2.1	Selection of positive and negative datasets	5-144
5.2.2	Estimation of a model's predictive power	5-145
5.2.3	A modified approach to find rotational positioning motifs	5-146
5.2.4	Model prediction using <i>Eponine</i>	5-146
5.2.5	Principal components analysis of trinucleotide background distributions	5-147
5.3	RESULTS AND DISCUSSION	5-148
5.3.1	Unanchored training results	5-148
5.3.2	Anchored training results using randomized chicken nucleosome sequences as negative data	5-150
5.3.3	Could the background trinucleotide distribution in different genomes affect nucleosome positioning?	5-153
5.3.4	Anchored training results using background chicken genomic DNA as negative data	5-156
5.4	CONCLUSION	5-161
6	SUMMARY	6-162
6.1	A DIFFICULT AREA TO RESEARCH	6-163
6.1.1	The sensitivity of different methods used to detect nucleosome positioning	6-164
6.1.2	Properties of the [CWG] motif	6-165
6.1.3	Possible influence of repeats in nucleosome positioning	6-165
6.1.4	Concluding remarks	6-166
7	REFERENCES	7-167
8	APPENDIX	8-176
A.	MULTIPLE SEQUENCE ALIGNMENTS OF EXPERIMENTALLY-MAPPED NUCLEOSOME DATASETS	8-176
B.	CYCLICAL HIDDEN MARKOV MODELS TRAINED FROM VARIOUS TYPES OF SEQUENCES	8-177

List of Figures

Figure 1.1: A hierarchical view of chromatin structure. Reproduced figure (Hartl & Jones, 1998).....	1-2
Figure 1.2: Top-level view of a nucleosome. Cylinders indicate alpha-helices; white hooks represent arginine/lysine tails. Reproduced figure (Rhodes, 1997)).	1-3
Figure 1.3: Organism sources of Levitsky et al's nucleosome sequence dataset (Levitsky <i>et al.</i> , 1999).	1-26
Figure 1.4: Length distribution of sequences in Levitsky <i>et al's</i> nucleosome database.....	1-27
Figure 1.5: Simple counting of [AA]-spacing in the 2 experimentally-mapped nucleosome datasets (Section 1.8).	1-31
Figure 2.1: A 2-state hidden markov model which emits symbols from the DNA alphabet. Boxes represent states and arrows represent transitions. The emission and transition distributions for State A are shown in red; State B's corresponding distributions are shown in blue.	2-45
Figure 2.2: 2 DNA sequences which are likely to receive a high score and a weak score respectively with the model of Figure 2.1. The locations of [W] regions are underlined. 2-	46
Figure 2.3: (a) A possible path through the HMM which could have emitted (b) the corresponding DNA sequence.....	2-47
Figure 2.4: Histogram of DNA flexibility values (Packer <i>et al.</i> , 2000b).....	2-52
Figure 2.5: The concept of translation and scale in wavelet terminology. This figure is a slightly modified version of a figure from Robi Polikar's 'Introduction to Wavelets' online tutorial (Polikar, 2000).	2-55
Figure 2.6: The multiresolution property of wavelets. The x and y axes represent increasing values along the DNA sequence co-ordinates and frequency values respectively.	2-57
Figure 3.1: The original 10-state cyclical hidden markov model (HMM) trained from exon sequences (Baldi <i>et al.</i> , 1996). The motif [VWG] was observed in states 8, 9 and 10. . 3-	60
Figure 3.2: Different cyclical HMM architectures: (a) F1, (b) F2 and (c) F3.	3-63
Figure 3.3: An example of 'Viterbi-labelling' a DNA sequence (top row) with a 10-state cyclical HMM. In the example Viterbi path (second row), the regions labelled '0123456789' demarcate corresponding locations in the DNA sequence where the wheel of the cyclical HMM has been used. 'n' is assigned to regions where the 'Null' state has been used.	3-69
Figure 3.4: Models learnt using different architectures of 10-state cyclical HMMs. Each column in the figure represents a state in the HMM. States within the wheel are indexed from 0 to the number of the last state in the wheel. "n" represents the Null state. The two rows represent the probability distributions of the emission and transition spectra respectively. The height of the respective characters represent their information content in the distribution. Shown are (a) F1 model learnt from exon sequences, (b) F2 model learnt from intron sequences and (c) F3 model learnt from repeat-masked intron sequences.....	3-72

Figure 3.5: An F3 model, whose emission parameters have been crudely reproduced from the B&B model. The transition parameters were all fixed to the same value since the original parameters were not available.....	3-77
Figure 3.6: 10 state cyclical HMMs learnt using alphabets other than 1 st order DNA: (a) F2 dinucleotide alphabet model learnt from intron sequences. Here, the emission spectrum is represented as the probability of observing a letter in position j given the position of a primary letter in $j-1$ (the row header represents the primary letter). (b) F3 flexibility alphabet model learnt from exon sequences.....	3-79
Figure 3.7: Frequency of distances between a state, within a wheel, back to itself in the state paths of two 10-state cyclical HMMs. The models used were (a) a crudely-reproduced B&B model illustrated in Figure 3.5 and (b) an F2 model illustrated in Figure 3.4(b).	3-82
Figure 3.8: 2 apparent motifs observed in F3 models: (a) [CWG] motif observed in <i>States</i> 234 and (b) [W] motif observed in <i>State</i> 3. The 2 examples shown are 11 state cyclical models; however, the same motifs were also observed in cyclical models of wheel size range 6 – 12 states (Appendix B).	3-84
Figure 3.9: Examples of Viterbi labelling a 13MB contig of human chromosome 22 using various F3 models.....	3-86
Figure 3.10: Comparison of model to model labelling. An F3 model, which had learnt a [CWG] motif (Model ID <i>interMask1_c12</i> in Appendix B), was used to label a 2.5MB sequence of human chromosome 22. The labelling of this was compared to the labelling of other models, of different wheel sizes, whose apparent motifs were either [CWG] or [W] respectively.....	3-89
Figure 3.11: Boxplots showing percentage of genome sequence labelled as wheel states by models which learnt apparent [CWG] or [W] motifs respectively.	3-89
Figure 3.12: The 23 most frequent trinucleotides in the background distributions of (a) human and (b) mouse.....	3-91
Figure 3.13: Fasta sequences of an Alu sequence (frequently labelled by cyclical [CWG] models) and a Charlie sequence (frequently labelled by cyclical [W] models). Sequences obtained from RepBase (Smit & Green, 1997).....	3-97
Figure 3.14: Histogram of lengths of cycle-labelled regions using F3 models. (a), (b) show data for human and mouse genomic sequences respectively; these were labelled with a [CWG]-learnt model (Model ID: <i>intronM1_c10</i> (Appendix B)). (c), (d) show data for human and mouse genomic sequences respectively, which were labelled with a [W]-learnt model (Model ID: <i>intronM2_c11</i> (Appendix B)). The balloons show features which were frequently associated with the corresponding peaks (the values shown are the ratio of the observed to expected frequencies).....	3-99
Figure 3.15: Viterbi alignments of chicken sequences, with 10-state F3 models which were trained from the chicken nucleosome datasets. (A) Alignments of 6 sets of jack-knifed test sequences (10 sequences per set). The ends of the sequences were padded in grey to represent the results in 150 bp windows. (B) Alignment of randomly-selected 146 bp chicken genomic fragments with a model trained from the chicken nucleosome dataset.	3-103
Figure 3.16: Boxplots of forward scores of test sequences labelled with F3 models of different wheel sizes.	3-107
Figure 3.17: Analysis of motif periodicity using a simple counting procedure: (a) [CWG] motif and (b) [WWW] motif.....	3-109

- Figure 3.18: (a) Plot of a [CWG] motif-learned F3 model's labelling density vs. density of the [CWG] motif itself (window size: 100 Kbp). These are shown alongside exon and Alu densities in a 5MB contig of human chromosome 22. (b) Correlation co-efficients of these densities. 3-111
- Figure 3.19: Density plots of [CWG] repeats in a human genomic sequence shown at different resolutions. 'w' is the window parameter and 'd' the threshold density of [CWG] within the window. The top density plot is a 'moving average' representation. The red and black boxes below represent non-overlapping 200 bp windows having >0.33 and >0.29 [CWG] densities respectively. 3-112
- Figure 4.1: Overlapping windowing scheme for removing edge effects in 'wavelet transform' analysis windows. 4-118
- Figure 4.2: Continuous wavelet transform spectra compared between eukaryotic DNA flexibility sequences and a sample prokaryotic DNA flexibility sequence. The figures were obtained by performing the wavelet transform on randomly chosen 100,000 bp segments of the following sequences: (a) a clone from human chromosome 22 (Ensembl ID: AC004019.20.1.260409), (b) *Saccharomyces cerevisiae* chromosome XV (Genbank ID: NC_001147) and (c) the *Mycobacterium tuberculosis* genome (Genbank ID: AE000516). The units on the z-axis were measured in decibels (dB); the colour gradients shown are based on a contour map of 48 colours ascending from red to blue. Red represents 0 or <0 dB intensity and dark blue represents the strongest observed intensities around 31 dB. 4-122
- Figure 4.3: Continuous wavelet transforms of 3 large eukaryotic DNA contigs. These 2D plots were obtained by thresholding the wavelet co-efficients at 28 dB and plotting only those regions which were above this threshold. These results were obtained from transforming (a) 63 MB of human chromosome 20, (b) the q arm of chromosome 22 (32 MB) and (c) a 30 MB syntenic region between human chromosome 20 (sequence co-ordinates 29.4 MB to 62.9 MB) and mouse chromosome 2 (sequence co-ordinates 172.1 MB to 202.3 MB). 4-124
- Figure 4.4: Probability distribution of observing periodic flexibility patterns in the range 50–1000 bp in 3 different eukaryotic chromosomes. The results were obtained from (a) human chromosome 20, (b) human chromosome 22 and (c) mouse chromosome 19. . 4-126
- Figure 4.5: Features frequently associated with periodic flexibility patterns in (a) human chromosome 20, (b) human chromosome 22, and (c) mouse chromosome 19. Values greater than 1.0 indicate that a feature was more frequently associated with flexibility patterns than expected. The reverse is true for values less than 1.0. 4-128
- Figure 4.6: Flexibility alignment of 300 bp sequences of A) regions exhibiting 100–200 bp periodicity in flexibility (wavelet co-efficients >28 dB) and B) randomly selected human DNA sequences. Red and green colours represent strong rigidity and strong flexibility respectively. RepeatMasker analysis showed that the sequences in A) were mostly Alu repeats. 4-129
- Figure 4.7: Zoomed view of periodic flexibility patterns (80–120 bp) having wavelet co-efficient strengths >28 dB. 3 different resolutions are shown; in each case, the locally detected periodic flexibility is shown as a red bar. Positive and negative strand Alu repeats are shown as blue bars. 4-131
- Figure 4.8: Correlation of B repeat density and gene density in a region of mouse chromosome 2. 4-134
- Figure 5.1: ROC curves of unanchored models learnt from Levitsky *et al*'s data (Table 5.2). The test set contains 25 sequences from the original dataset (positive set) and 25 sequences obtained from randomizing the original dataset (negative set). 5-149

Figure 5.2: ROC curves of anchored models learnt from the chicken nucleosome dataset (Figure 5.3(h),(j)): (a) tested against a jack-knifed negative set of randomized chicken nucleosome DNA and (b) tested against a negative set of background chicken genomic DNA.....	5-150
Figure 5.3: Anchored models learnt using the chicken nucleosome dataset as a positive set and a randomized version of the same dataset as a negative set. Models (a)-(h) were learnt in different cycles from <i>training run a</i> and models (i)-(j) were learnt in different cycles from <i>training run b</i> . The inverted blue triangle represents the “anchor point”.	5-152
Figure 5.4: Principal components analysis of the background trinucleotide distributions of different genomes and the 2 nucleosome datasets.....	5-153
Figure 5.5: Background trinucleotide composition in descending order in (a) the human genome and (b) the Levitsky nucleosome data.....	5-155
Figure 5.6: (a) An anchored model learnt using the chicken nucleosome dataset as a positive set and background chicken genomic DNA as a negative set. (b) ROC curve of the same model using a jack-knife test. ROC curves are shown for this test set as well as the reverse-complements of the same test set.....	5-157
Figure 5.7: Locations of high-scoring BLAST segment pairs between the GGB locus in chicken and in mouse.	5-160
Figure 5.8: Prediction using model <i>chickBack_d5000</i> (Figure 5.6(a)) on the chicken GGB locus and homologous regions in mouse. The sequence co-ordinate axis represents the mouse sequence.....	5-160

List of Tables

Table 1.1: Accuracy of different nucleosome mapping methods (Ioshikhes & Trifonov, 1993).	1-27
Table 3.1: Transition parameters used to initialize F1 models.....	3-67
Table 3.2: Various training sequences and their respective sizes. For human exon, intron and intergenic sequences, random samples of size range 500 – 5000 bp were taken. .	3- 69
Table 3.3: Viterbi-labelling patterns, of a 10 state cyclical HMM, which were used to assess the wheel's labelling tendency. The characters, in the second column, represent the following states: "State 0", "State W" (any wheel state) and "State 9".....	3-70
Table 3.4: Analysis of skipping and looping behaviour of various F3 models (Models shown in Appendix B).	3-76
Table 3.5: Reproducibility of Viterbi labelling using different F3 models and estimation of features enriched in predictions. The results in the table are sorted by the apparent motif learnt in the model (the motifs were visually approximated). Motifs which looked partly like either [CWG] or [W] are referred to as 'intermediate'. Key for motif column: .	3- 93
Table 3.6: Features observed to frequently have high densities of [CWG] repeats. A window size of 200 bp and cutoff threshold of 35% [CWG] density was used.....	3-112
Table 4.1: Percentages of repeat families which were associated with strongly periodic flexibility regions (wavelet co-efficients >28 dB) in descending order. These are compared to the proportion of total observed periodic flexibility (second column). The second columns do not sum to 100% as the proportion is measured across the distribution of a range of periodic patterns (for instance, the same region may be strong for both 80 bp periodic as well as 200 bp periodic patterns).	4-133
Table 5.1: Summary of classification categories used.	5-145
Table 5.2: Unanchored models learnt using Levitsky <i>et al</i> 's nucleosome dataset as a positive set and a randomized version of the same dataset as a negative set. Both models, (a) and (b) were obtained from independent runs. Negative motifs have been shaded grey and CpG motifs, which are rare in eukaryotic genomes, have been highlighted in yellow.	5-148