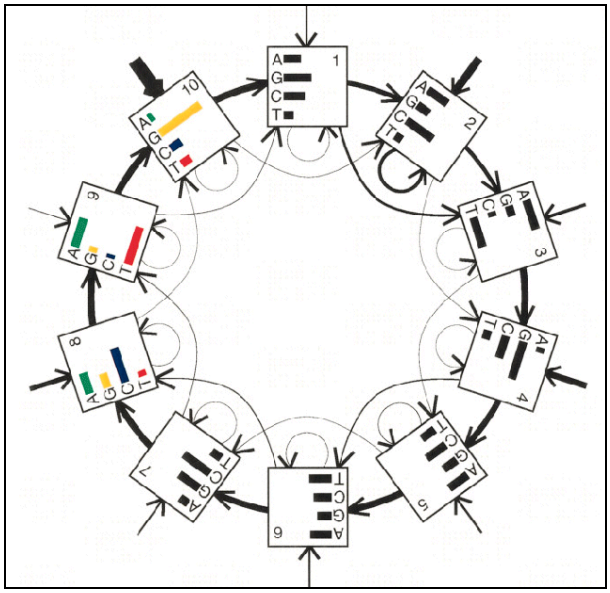
3 Cyclical Hidden Markov Model Analysis to find Signals Involved in Nucleosome Rotational Positioning

3.1 Introduction

The hypothesis for intrinsic DNA curvature is based on 10 periodic DNA motifs, which are thought to influence nucleosome rotational positioning (Sections 1.4.1, 1.4.2, 1.9.3). From the analysis of the chicken nucleosome dataset (Section 1.8.1), this was described as 10 bp-phased [AA] dinucleotides, which showed a 5 bp-phase shift from [GC] dinucleotides. For the Levitsky nucleosome dataset (Section 1.8.2), this was described as 10 bp-phased [AA] dinucleotides, which were similarly 5 bp-phase-shifted from [TT] dinucleotides. Both these proposed signals imply a 10 bp-phased "rigid" motif which could influence rotational positioning. Baldi and Brunak used a different kind of approach to find rotational positioning signals, using cyclical HMMs (Sections 1.9.3, 1.11.1). From their results, they described 10 bp-phased [VWG] motifs as a potential rotational positioning signal. The structural basis of this claim was different to the phased "rigid" motif described from analysis of the 2 nucleosome datasets. This suggests that 10 bp-phased 'flexible' motifs could influence rotational positioning. This led to the motivation to extend cyclical HMM analysis (Baldi *et al.*, 1996) to learn and predict 10 bp-phased motifs, which could potentially influence nucleosome rotational positioning.

Baldi and Brunak's cyclical HMM architecture is shown in Figure 3.1; this model is herein referred to as the B&B model. The original architecture had a series of states looped together to form a "wheel"; each state in the wheel had 3 main transitions: next, skip and loop (explained in more detail in the Methods section, 3.2.1). The [VWG] motif (States 8, 9, and 10 in Figure 3.1), was learnt strongly in exons and learnt weakly in introns and intergenic regions (Baldi *et al.*, 1996). This was an interesting finding as it suggested that exons may possess intrinsic curvature and hence be able to direct the rotational positioning of nucleosomes.

Figure 3.1: The original 10-state cyclical hidden markov model (HMM) trained from exon sequences (Baldi *et al.*, 1996). The motif [VWG] was observed in states 8, 9 and 10.



One of the first objectives of the current research was to extend the architecture of the original B&B model to model both the "wheel series of states" and an additional background state called the *Null* state. The aim of this was to learn the background distribution to any "cyclical" patterns learnt in the "wheel" part of the HMM architecture. The *Null* state was also necessary for training HMMs, which could be used as a nucleosome prediction tool. The *Biojava* programming package (Down & Pocock, 1999), which was largely being developed in-house, was used to develop the software to carry out this analysis.

One major issue that needed to be dealt with was to establish if the original signal was a consequence of codon bias (aka coding bias)[10]. This was an important distinction to make as the described 10 bp-phased [VWG] motif in the B&B model was learnt from exon training sequences. The motif itself was also a 3 state one, which could have been due to recoding of coding bias.

[10] The sequence of nucleotides, coded in triplets (codons) along the mRNA, which determines the genetic code. This determines the sequence of amino acids in protein synthesis. Different organisms use different frequencies of codons in their genetic code leading to codon bias.

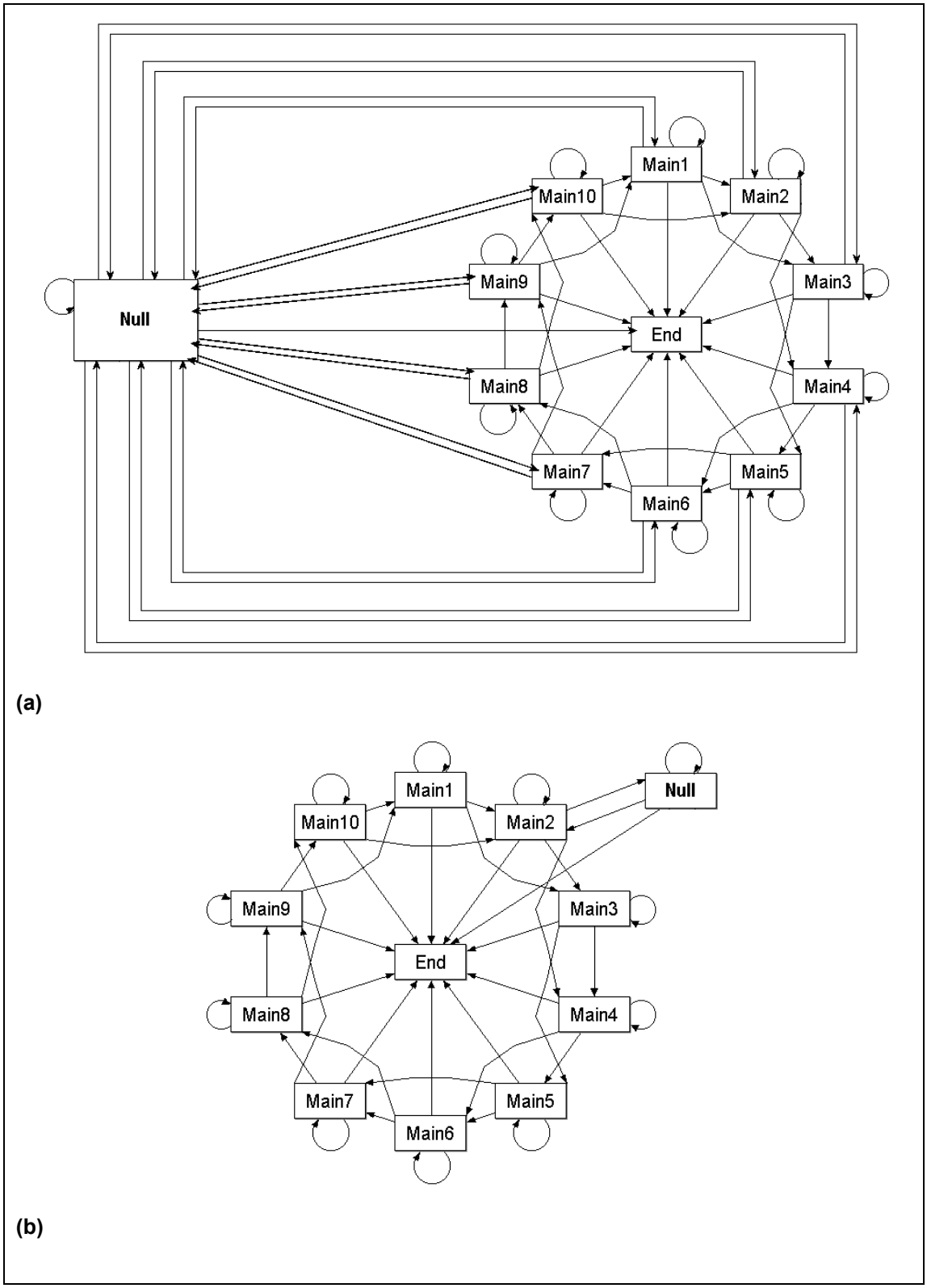
To model the physical aspect of rotational positioning more directly, a flexibility-emission alphabet was also developed to model DNA sequences as flexibility sequences (Section 2.3.2, page 2-53).

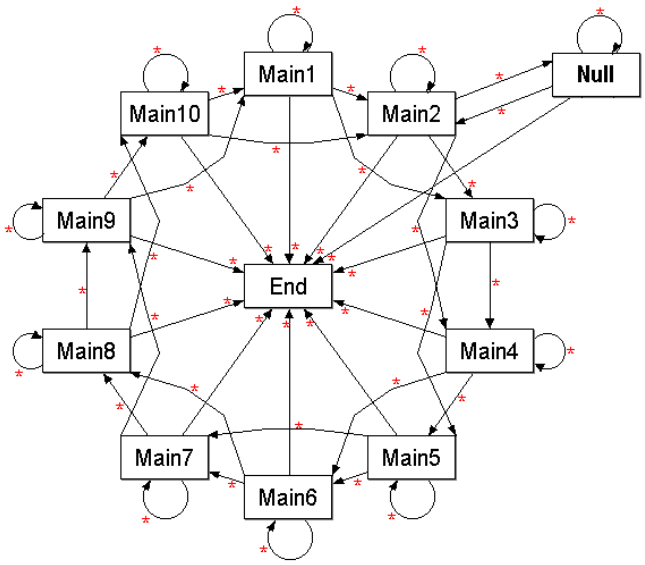
3.2 Methods

The main techniques used in this chapter involved HMM training and prediction. HMMs are introduced more generally in the introduction chapter of this thesis (Sections 2.2.1-2.2.3). This section will outline the construction, training and prediction procedure for a general architecture of HMMs, the cyclical HMM architecture. The software packages described were written using the *Biojava HMM toolkit*, which was developed by Matthew Pocock (Pocock MR *et al.*, 2000).

3.2.1 Construction of different kinds of wheel architecture

Figure 3.2: Different cyclical HMM architectures: (a) F1, (b) F2 and (c) F3.





(c)

The cyclical HMM architecture that was used for analysis in this chapter eventually resulted from a series of design refinements (Figure 3.2(a)-(c)). In Figure 3.2(a)-(c), boxes represent states in the HMM and arrows represent transition paths connecting these states. The boxes labelled *Main* are emission states which are looped together to form the wheel part of the architecture. In each of Figure 3.2(a)-(c), 10-state wheels are shown. The symbols which are emitted are from the DNA alphabet of 4 symbols: "a,c,g,t". All the *Main* states have at least 4 transition paths:

- *'next'* for going to the next state,
- *'loop'* for going back to itself,
- *'skip'* for skipping past the next state in the wheel and
- *'end'* for ending from the model

The only state which is not shown in Figure 3.2(a)-(c) is the *Start* state, which has transitions to all the emission states.

The architectures shown in Figure 3.2(a)-(c) can be described as follows:

(a) F1 cyclical HMM architecture

The initial model architecture that was developed, F1, had the greatest degree of freedom of all the architectures. All the *Main* states had a transition path to the *Null* state. The *Null* state also had transition paths back to each of the *Main* states.

(b) F2 cyclical HMM architecture

The F2 architecture can be considered 'moderately free' compared to the numerous additional paths of the F1 type architecture.

(c) F3 cyclical HMM architecture

The F3 type architecture looks exactly like F2. The only difference is that all the transition parameters were kept constant or 'untrainable'; transition and emission parameters are discussed subsequently in Section 3.2.3.

3.2.2 Parameter setups in preparation for cyclical HMM training

Once a cyclical HMM architecture was established, the next step was to train it from a sequence dataset. Two important parameters which had to be setup before starting the model training step were:

- **Number of states in the wheel**

The number of emission states which formed the wheel part of the architecture was kept as a variable. Most of the experiments involved training and analyzing 9 and 10 state wheel models; however, other models with wheel sizes ranging between 6-12 states were also trained (examples in Appendix B).

- **Pseudocounts**

Data-overfitting can occur when a specific symbol of an emission alphabet is under-represented in the training set; for example observing 0 counts for the symbol "a" in a particular emission state. The probability of observing a weak emission probability for "a" still needs to be modelled for the HMM to be a general one. A

solution for this was to add a certain number of 'fake' counts or pseudocounts to all counts of emission symbols observed. Most of the training sequences used (Section 3.2.5) were quite long (>500 bp); despite this, a low pseudocount number of 5 was used to prevent overfitting.

3.2.3 Model training

The model training procedure can be outlined in three steps:

1. **Model initialization**

At the first step of training, the models had to be initialized with fake numbers of counts. The emission probabilities were always initialized randomly. However, for the transition probabilities, initialization required adding counts in such a way that a continuous loop around the wheel would be preferred to using any of the skip or loop transition paths within the wheel. Table 3.1 summarizes the transition probability distributions used to initialize F1 models. A high *next* transition probability of 0.96 would ensure continuous use of the next transitions within the wheel compared to the relatively smaller 0.01 probabilities for using any of the other available transitions. For the *Null* state, the loop transition parameter back to itself was initialized to the same value as the *next* transition parameters within the wheel (0.96). For the *Null* state, a high loop probability coupled with a small probability to the wheel states (0.03) was expected to effectively model the background to any 'cyclical' emission distributions learnt in the wheel. The transition parameters for starting or ending from all emission state in the model were initialized with equal values.

Table 3.1: Transition parameters used to initialize F1 models

SOURCE STATE	TRANSITION TYPE	INITIAL PARAMETER
wheel state	Next	0.96
wheel state	Skip	0.01
wheel state	Loop	0.01
wheel state	null state	0.01
null state	Loop	0.96
null state	wheel state	0.03
all emission states	End	0.01
start	all emission states	1/[no. of emission states]

For F2 and F3 models, the initialization parameters were roughly the same as for F1 in Table 3.1. The major difference was that only one of the wheel states had a transition path to the *Null* state. This transition parameter was initialized to 0.02; all the *next* transition parameters within the wheel were set to a constant value of 0.96. For F3 models, all the transition parameters were kept constant or 'untrainable' between different training runs; only the emission probabilities could be trained.

2. Model training

All models were trained using the Baum-Welch training method (Section 2.2.3).

3. Training termination

All the models were trained until the log score difference between training runs had converged to 0.1. However, if the scores had not converged within 250 cycles, the training was forfeited and a fresh training run initiated. 1 in 20 training runs were forfeited due to this.

3.2.4 Construction of emission alphabets other than DNA

Alternative emission alphabets to the 4-symbol DNA alphabet were also used with the mentioned cyclical HMM architectures. Firstly, a flexibility alphabet was used (Section 2.3.2).

A dinucleotide DNA alphabet (16 symbols) was also used. The results of model training could then be compared with published DNA flexibility values based on dinucleotide parameters (Bolshoy *et al.*, 1991; Calladine & Drew, 1986; Packer *et al.*, 2000a; Satchwell *et al.*, 1986). To gain the dinucleotide view of a DNA sequence, 'overlapping windowed' views onto the original DNA sequence were taken. Each window was shifted by 1 bp relative to the position of the previous window. So, for example, for the DNA sequence "aagctg", the values of "aa, ag, gc, ct, tg" were ordered to form the dinucleotide sequence.

The results of model training could be visualized as in Figure 3.6(a) (page 3-79).

3.2.5 Datasets of training sequences

The sequences selected for model training included the 2 known mapped nucleosome datasets (Section 1.8), 1 archaeal sequence dataset (EMBL accession ID: *NC_003106*) and various sequences obtained from human chromosome 20 (data extracted from the *Ensembl* core database (Clamp *et al.*, 2003; Hubbard *et al.*, 2002)). These are summarised in Table 3.2. Only experimentally-confirmed human exon sequences were used for training.

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.

Sequence type	Dataset size
Levitsky nucleosome dataset (Levitsky *et al.*, 1999)	193 x ~146 bp = 28,178 bp
Chicken nucleosome dataset (Satchwell *et al.*, 1986)	177 x ~146 bp = 25,842 bp
Archaeal genome *Sulfolobus tokodaii* masked for coding sequences (EMBL accession ID: NC_003106)	360,141 bp
alu repeat sequences	500,000 bp (average Alu length = 300 bp)
Experimentally-confirmed exons	568,098 bp
Intergenic sequences	1,164,369 bp
Intergenic sequences masked for all kinds of repeats (including SINEs, LINEs, DNA transposons)	602,712 bp
Randomly sample intron sequences	629,770 bp
Intron sequences masked for all kinds of repeats (including SINEs, LINEs, DNA transposons)	687,945 bp

3.2.6 Viterbi labelling analysis

The most likely path a cyclical HMM takes through a sequence was predicted using the *Viterbi* algorithm (Section 2.2.2). A typical output from this algorithm is shown in Figure 3.3. The primary target sequences which were analysed included two contigs from human chromosome 22 (13MB and 2.5MB respectively) and a contig from mouse chromosome 19 (Data extracted from Ensembl core database, (Clamp *et al.*, 2003; Hubbard *et al.*, 2002)).

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.

```
ggcagtcttcacagtgatggtagctttctggagacagcctccaatttgctgcagtacctg

nnnnnnnnnn0123456789nnnnnnnnnnnnnnnnnnnnnnnnnnnn0123456789n
```

3.2.7 Analysis of a model's "wheel"-labelling pattern

Once the Viterbi path of a model on a test sequence was obtained, the frequencies of the model's wheel to (1) skip states (2) make a full turn, and (3) loop on its own states were calculated. These values were used as indicators to assess if the wheel was trying to match a higher or lower size wheel in the test sequence. For the example

Viterbi path of a 10 state cyclical HMM (Figure 3.3), the frequencies of the labelling patterns in Table 3.3 could indicate this.

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*".

Wheel's labelling tendency	Viterbi labelling pattern
Skip to fit a lower wheel size	$0\ W_{(<8)}\ 9$
Fit its own wheel size	$0\ W_{(8)}\ 9$
Loop to fit a higher wheel size	$0\ W_{(>8)}\ 9$

3.2.8 Labelling analysis of chicken nucleosome sequences and chicken genomic sequences

A jack-knife experiment was performed on the chicken nucleosome dataset. 10 sequences were kept as test sequences and the rest used for training. The aim was to examine what proportion of the test sequences were labelled with wheel states. Using this approach, the test sequences were clustered according to their labelling pattern. Fragments of the 2 available chicken genomic clones (Section 1.8.1) were also labelled to examine if the labelling patterns were different to the ones for the jack-knifed nucleosome test sequences.

3.2.9 Estimation of frequently "wheel-state"-labelled features

To estimate whether any known genomic features were enriched in 'wheel-state' labelled regions, the frequency of concurrently observing a wheel-labelled region and a known genomic feature type was calculated (the observed frequency). This was calculated as the total length spanned concurrently in a chromosome by both the wheel-labelling and the genome feature divided by the total length of the chromosome. The ratio between this frequency and the expected frequency of the

genomic feature and the wheel labelling[11] was calculated and ranked as in Table 3.5 (page 3-93). For the exon category, both predicted and experimentally confirmed exons were used.

3.2.10 Visualisation of predictions against genomic annotations

The Distributed Annotation System (DAS) (Dowell *et al.*, 2001) was used to visualize predictions and compare their locations with respect to annotated genomic features. This protocol allowed predictions to be uploaded to an Ensembl annotation server (Clamp *et al.*, 2003; Hubbard *et al.*, 2002) using a specific das file format. The main genomic annotations were stored in a reference server. An example of this kind of visual representation is seen in Figure 3.9, page 3-86.

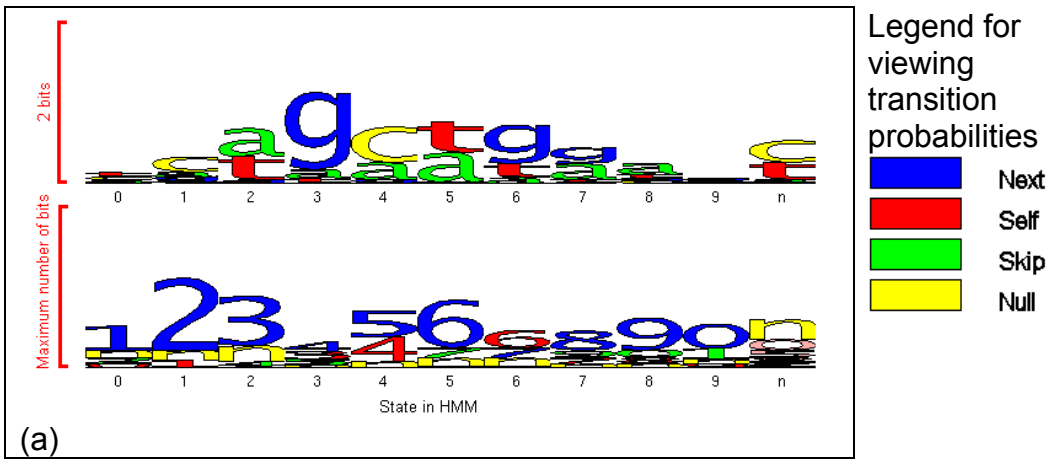
[11] The product of the wheel-labelling frequency and the frequency of the genomic feature in the chromosome

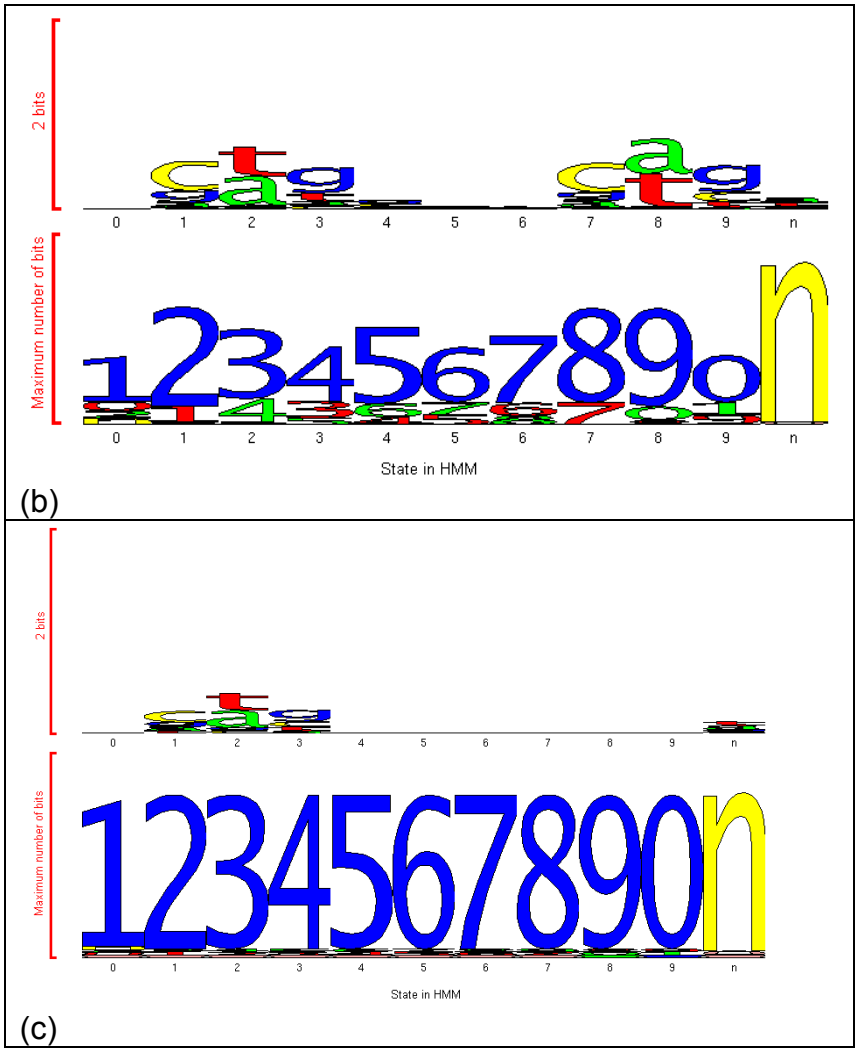
3.3 Results and Discussion

3.3.1 Model-training experiences using different kinds of cyclical HMM architectures

A number of different cyclical HMM architectures were developed and tested to learn potential rotational positioning signals. The ultimate architecture that was selected for analysis had a much more constrained transition-path component compared to the initial design. Figure 3.4(a) – (c) shows the evolution of the final architecture designated the F3 type; these examples use the DNA emission alphabet.

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.





The first kind of architecture that was developed was the "very free" F1 type. A 10-state model, which was trained from coding sequence, using this architecture, is shown in Figure 3.4(a). The motif, described by Baldi and Brunak as [VWG], was observed in this model. However, as can be seen in the example model, the motif was seen a number of times in the wheel. In Figure 3.4(a), it appears twice: firstly at *States 1,2,3* and then at *States 4,5,6* in the wheel. Between different training runs, this motif would appear more than once within the wheel but the spacing between the motifs did not remain constant. This result was most probably a consequence of the inherent freedom of the architecture: there were so many transitions possible to the *Null* state from the wheel component that the HMM did not necessarily have to use all the '*next*' transitions in the wheel states to fit a 10-periodic wheel. This extreme

freedom is exemplified in the transition distributions in Figure 3.4(a), where the information content of the *'next'* transitions was clearly not dominant over the other available transitions. Also, the transition probability to the *Null* state appeared higher for certain states compared to others (for example, *States 1,2,4,5* in Figure 3.4(a)). The inevitable downside with this approach was that a periodic signal corresponding to the wheel size of 10 states could not be modelled. Therefore, when the *Viterbi* algorithm was used to align or label a sequence with models of the F1 architecture, the state-labelling also appeared random: the labelling was not 'wheel-like' and appeared to move in and out of the wheel to the *Null* state very often. This general outcome led to the development of the next type of architecture, the F2 type.

The F2 model architecture can be described as "moderately free" (Figure 3.4(b)). The example model in Figure 3.4(b) firstly shows one important property about the [VWG] motif: this pattern could be learnt from non-coding sequence as well as from coding sequence. This example model was trained from raw intron sequences and the motif was seen in two positions: firstly, *States 1,2,3* and secondly *States 7,8,9* (Figure 3.4(b)). However, even after limiting the total number of transitions to the *Null* state from just one wheel state, the use of the transitions was still irregular as can be seen from the information content of the '*next'* probabilities: *'State 0 to State 1'* was almost half of that of *'State 1 to State 2'*. This meant that this architecture had still not been useful at modelling a period corresponding to the size of the wheel. Although labelling sequences with this model showed more 'wheel-like' behaviour compared to the F1 models, the *skip* and *loop* transitions were being used almost at the same proportions as a full turn around the wheel (Figure 3.7(b)). This observation led to a final alteration in the model architecture leading to the F3 architecture.

The F3 type architecture was consequently the tightest architecture design. This time, the transitions were made 'untrainable': these parameters remained fixed throughout training. This was expected to force the HMM to model full turns around the wheel and at the same time, learn its respective background. An example is shown in Figure 3.4(c) where the model was trained from repeat-masked intron sequences. The [VWG] motif was learnt and appeared to occur every 10 bp. The full range of trained F3 models is catalogued in Appendix B. The 10-state F3 models which showed this were trained from exon, intron, intergenic, masked intron, masked intergenic and the chicken nucleosome sequences (Appendix B). This gave an impression that the motif was a 10-periodic one but upon *Viterbi*-labelling, it was observed that the HMM would now only model full-turns around the wheel (Table 3.4). The tightening of the transition parameters may have backfired. However, analysis using this architecture continued and further analysis was performed using wheel sizes ranging between 6 and 12 states (Appendix B).

Table 3.4: Analysis of skipping and looping behaviour of various F3 models (Models shown in Appendix B).

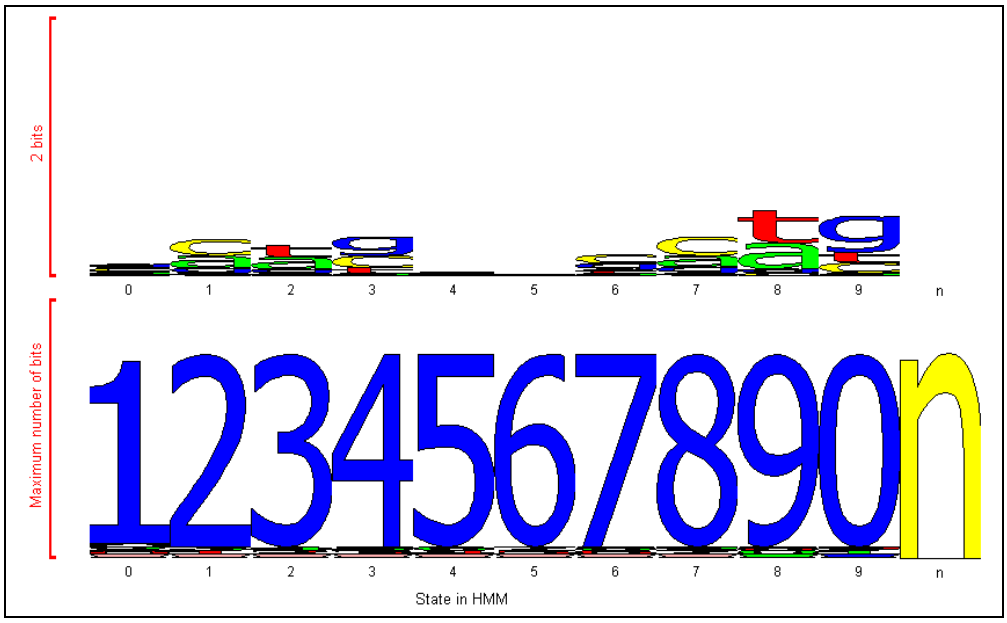
TRAINING SOURCE	STATES	SKIP	NEXT	LOOP	MOTIF
intronMasked0	6	0	2276	0	
intronMasked2		0	2283	0	
interMasked0		0	2491	0	
intronMasked1		0	2381	0	
interMasked1		0	2457	0	
interMasked2		0	2602	0	
interMasked1	7	0	2728	0	
intronMasked2		0	2199	0	
interMasked2		0	2796	0	
interMasked0		0	2458	0	
intronMasked1		0	2224	0	
intronMasked0		0	2277	0	
interMasked0	8	0	2816	0	
interMasked2		0	2392	0	
intronMasked2		0	2582	0	
interMasked1		0	2788	0	
intronMasked1		0	2575	0	
interMasked0	9	0	2588	0	
interMasked1		0	2547	0	
interMasked2		0	2587	0	
intronMasked0		0	2450	0	
intronMasked1		0	2244	0	
intronMasked2		0	2462	0	
interMasked0	10	0	2668	1	
interMasked2		0	2644	0	
intronMasked0		0	2512	1	
intronMasked1		2	2649	16	
intronMasked2		1	2476	0	
interMasked0	11	3	2881	61	[CWG][12]
interMasked1		0	2574	0	
interMasked2		0	2575	0	
intronMasked0		4	2707	44	[CWG]
intronMasked1		4	2723	42	[CWG]
intronMasked2		0	2360	1	[W]
interMasked1	12	3	2874	31	[CWG]
interMasked2		3	2874	31	[CWG]
intronMasked0		3	2666	29	[CWG]
intronMasked1		7	2687	30	[CWG]

To compare the training results from the experiments in this chapter with the B&B model, the emission parameters of the published model were crudely

[12] Why the apparent motif is indicated as [CWG] and not [VWG] in this table is noted later (Section 3.3.4, *The [VWG] motif in retrospect and the distinction of two apparent motifs learnt in F3 human models*)

reproduced to represent a corresponding F3 model (Figure 3.5). The original transition parameters were not available hence only the emission parameters could be roughly reproduced from Figure 3.1. However, a slightly strong skip transition parameter was noticed from *State 1* to *State 3* in Figure 3.1. A fallback of not having the original transition parameters was that this slightly stronger skip transition was not modelled. This could bias the reproduced B&B model to behave more like a 10-wheel model rather than modelling a weak tendency to fit a 9 wheel as the original B&B model suggests. Another alarming observation about the B&B emission parameters was made at this point: it was noticed that the motif had appeared twice in the B&B wheel: *States 1,2,3* and *7,8,9* in Figure 3.5 and *States 2,3,4* and *8,9,10* in Figure 3.1. This raised doubts about the periodicity of the [VWG] motif and prompted further investigations (Sections 3.3.3, 3.3.4 and 3.3.7).

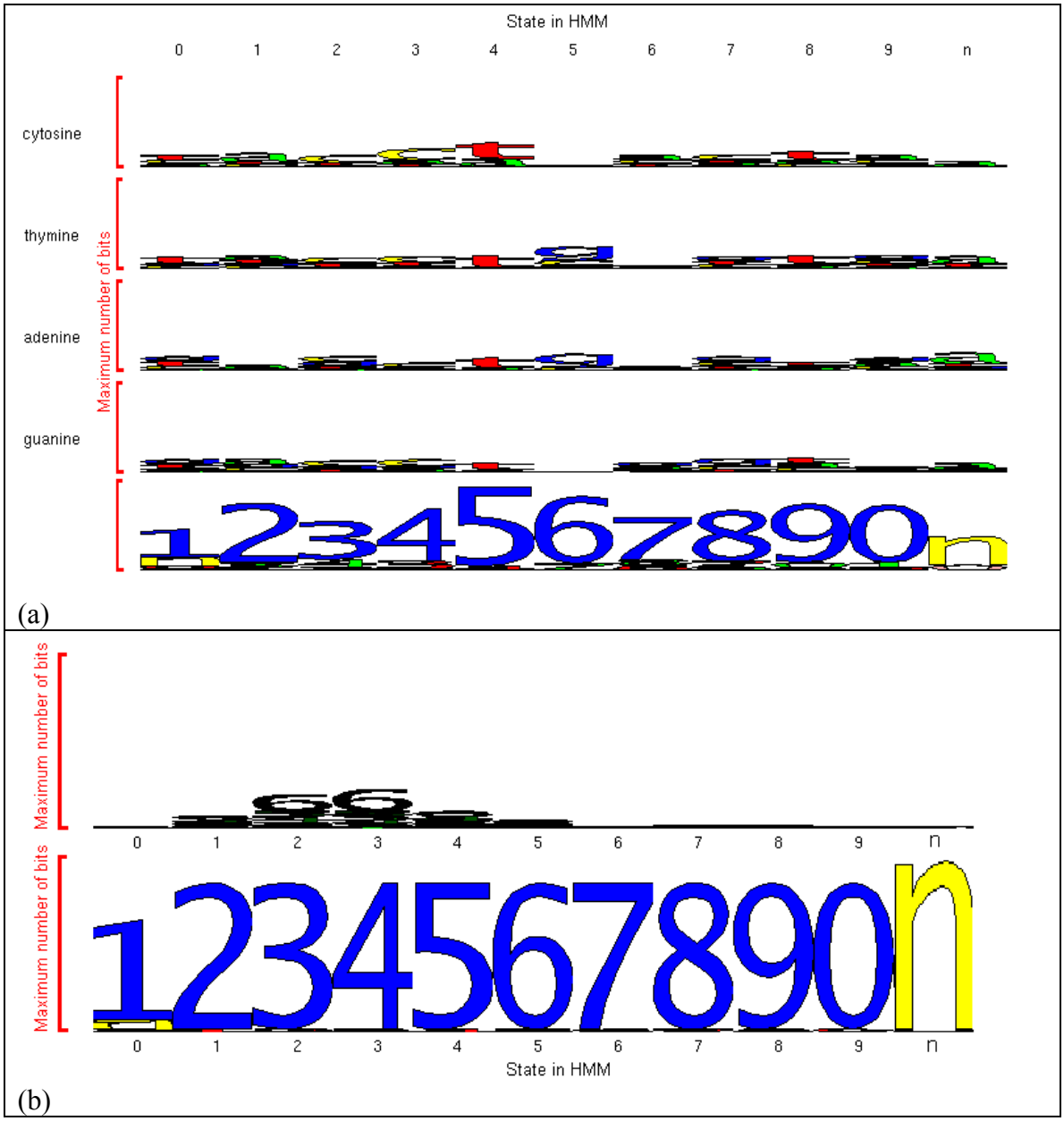
Figure 3.5: An F3 model, whose emission parameters have been crudely reproduced from the B&B model. The transition parameters were all fixed to the same value since the original parameters were not available.



3.3.2 Experiences of using non-DNA emission alphabets with cyclical HMMs

Two emission alphabets were developed in addition to the DNA alphabet for using with cyclical HMMs. The first one, which was a dinucleotide alphabet, did not yield greater information than what was already obtained using the DNA alphabet (Figure 3.6(a)). Figure 3.6(a), which shows a F2 model learnt from intron sequences, learnt the [VWG] motif in *States 3,4,5*. But this motif was seen for all 4 rows of conditional emission distributions (conditioned on observing any of the 4 symbols of *cytosine, thymine, adenine or guanine* in the previous state). If the observed motif was conditioned on only one of the symbols, the result would have been interesting and using the 2^{nd} order alphabet would have been potentially useful. The results, however, modelled the same motifs obtained using the DNA alphabet. Therefore, modelling attempts using this emission alphabet were eventually discarded.

Figure 3.6: 10 state cyclical HMMs learnt using alphabets other than 1st 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.



The other alphabet, based on flexibility, did not yield any consistent motifs between different training runs. Figure 3.6(b) is an example of an F2 model trained from coding sequences. In this case, a motif of 2 strong *'6'* symbols (representing conformational rigidity) was observed at wheel states *2* and *3*. Most other learnt models either did not have high information contents in the emission spectra or would learn motifs which were invariably different between runs on the same training data.

This lack of consistent results using the flexibility emission alphabet suggested two things:

- The flexibility conversion resulted in sequences which probably did not have any periodic patterns corresponding to the wheel sizes and
- The flexibility values of the sequence members of the [VWG] motif were not significantly different from the flexibility values of the background in the training data.

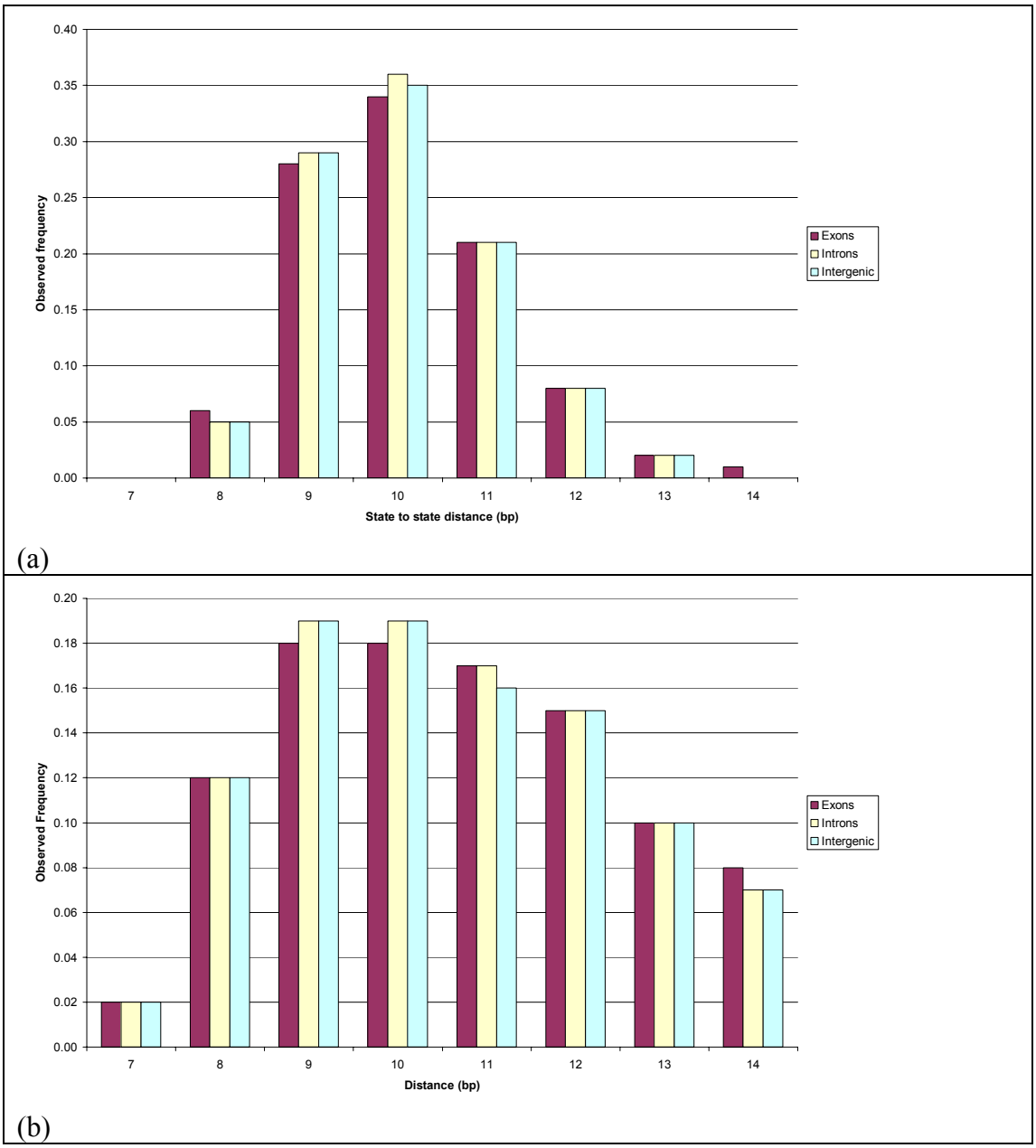
This result indicated that the structural basis for the [VWG] motif to effect nucleosome rotational positioning was perhaps not as convincing as was suggested earlier (Baldi *et al.*, 1996). However, the [VWG] motif itself was quite intriguing as it was being learnt both in coding and non-coding DNA sequences: the next step was to investigate if this motif was merely a consequence of coding bias or not.

3.3.3 An initial test to investigate if the B&B model had learnt codon bias

The fact that the [VWG] motif could be learnt in coding sequence, which itself is a relatively strong signal in genomic sequences, prompted an analysis of its periodicity. The first approach taken was to understand if the cyclical HMMs were trying to fit a 9 period rather than a 10 period. Since 9 is a modulo repeat of 3, a result of this period would suggest an effect of coding bias. To determine this, the wheel lengths of sequences labelled with a crudely-reproduced B&B model (Figure 3.5) and a 10-state F2 model trained from intron sequences (Figure 3.4(b)) were compared (Figure 3.7). An F2 model was chosen for this comparison rather than an F3 model because the frequencies of F3 models to skip and loop were marginal compared to making a full turn around the wheel (Table 3.4). In other words, an F3 model was too constrained for this comparison.

An important point about the original B&B model, which was mentioned earlier (Section 3.3.3), was that it appeared to have one *skip* transition, within the wheel, which was stronger than the other skip transitions in the wheel. This was not modelled in the F3-reproduced model as the original transition parameters were not available. This could mean that the reproduced B&B model was likely to fit a 10 state wheel more preferentially than the original B&B model. For the approximated B&B model, the wheel distance frequencies showed that the model mostly tended to make a full turn around its wheel; however, the frequency of skipping to a 9 wheel was greater than the frequency of looping to fit an 11-state wheel (Figure 3.7(a)). This observation was the same for both labelled coding sequences as well as for introns and intergenic sequences. This indicated that the model could have learnt coding signal. The fact that this skipping tendency was observed in introns and intergenic regions could perhaps be explained by the presence of un-annotated pseudogenes. Pseudogenes are short fragments of functionless coding DNA, which appear ubiquitously in genomic DNA.

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)



The wheel-labelled regions of the chosen F2 model gave a slightly different impression to the labelling of the reproduced B&B model (Figure 3.7(b)). The frequency of skipping to a 9-state wheel was the same as observing a full turn around the wheel. Once again, this behaviour was the same for coding and for non-coding DNA. The frequency of looping was once again less than the frequency of skipping.

However, compared to the B&B model, the frequency of looping was relatively closer to the frequency of making a full turn around the wheel (Figure 3.7(a)).

Fitting a 9-state wheel was, therefore, common for both the models but the 2nd F2 model had a tendency to fit higher wheel sizes as well. Based on this evidence, it could be suggested that the observation was related to coding bias. This matter was subsequently re-investigated using more direct approaches (Section 3.3.7).

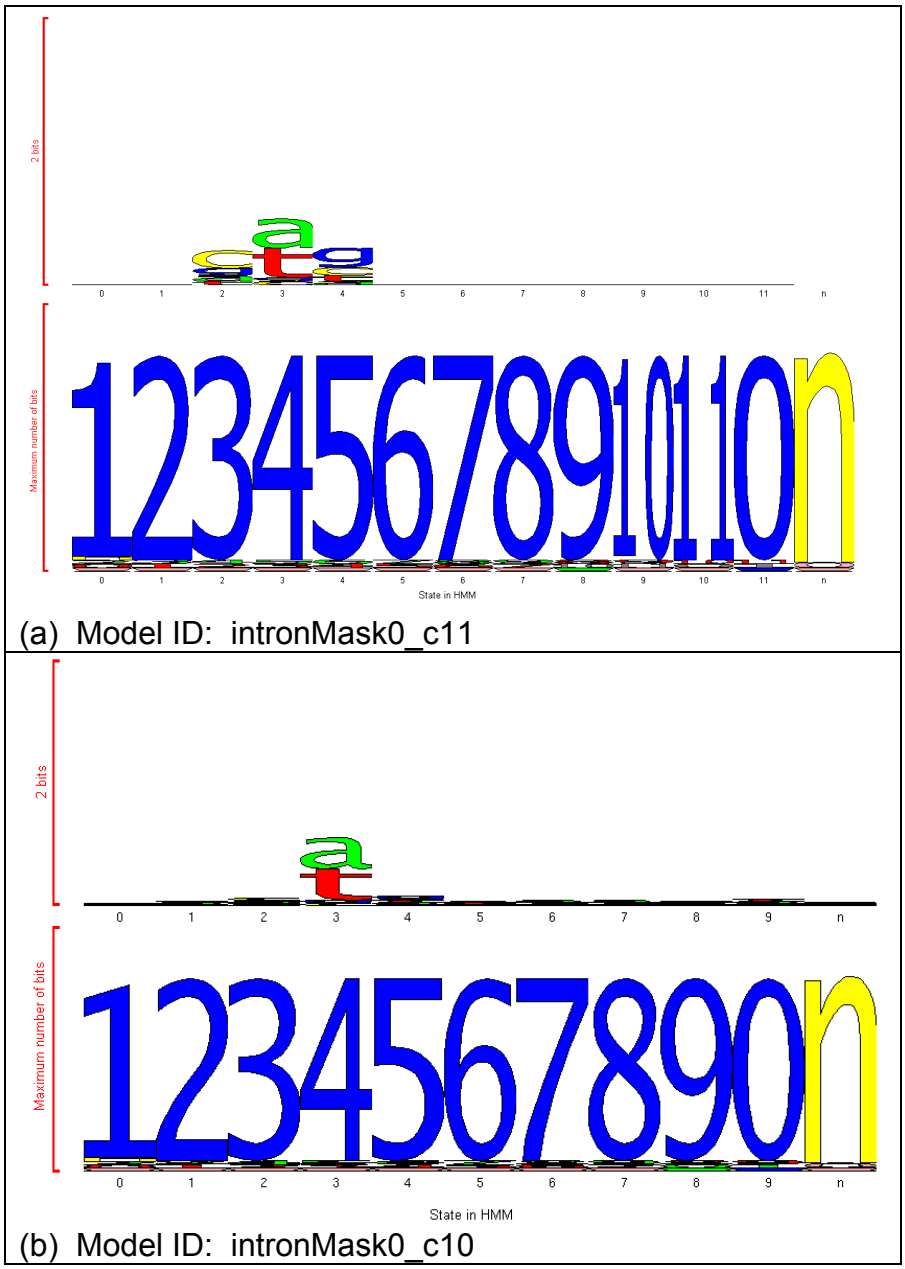
3.3.4 The [VWG] motif in retrospect and the distinction of two apparent motifs learnt in F3 human models

The cataloguing of F3 models, trained from human sequences[13], showed that most learnt either of 2 apparent motifs in the wheel: [CWG] or [W] (Figure 3.8 and Appendix B). The same training was done from mouse data, for example using repeat-masked (Smit & Green, 1997) mouse intergenic sequences (data not shown). It was observed that the models learnt the same 2 motifs that were being learnt from the human data.

With the exception of the Alu-trained models, all other models trained from human sequences learnt either of these 2 motifs within their wheel states. However, the motifs themselves were learnt for the whole wheel-size range tried, 6 – 12 states, suggesting that [CWG] and [W] occurred periodically over this entire range. An interesting property of both motifs was that they both represented the forward strand motif and its reverse complement; for example, the reverse complement of [CAG] is [CTG] and that of [A] is [T]. *Viterbi*-labelling a sequence and its reverse-complemented sequence with the same model, furthermore, showed that the models were aligning the same parts of the sequences (data not shown).

[13] The different types of human training data, that were used, were listed earlier in Table 3.2

Figure 3.8: 2 apparent motifs observed in F3 models: (a) [CWG] motif observed in *States 234* and (b) [W] motif observed in *State 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).



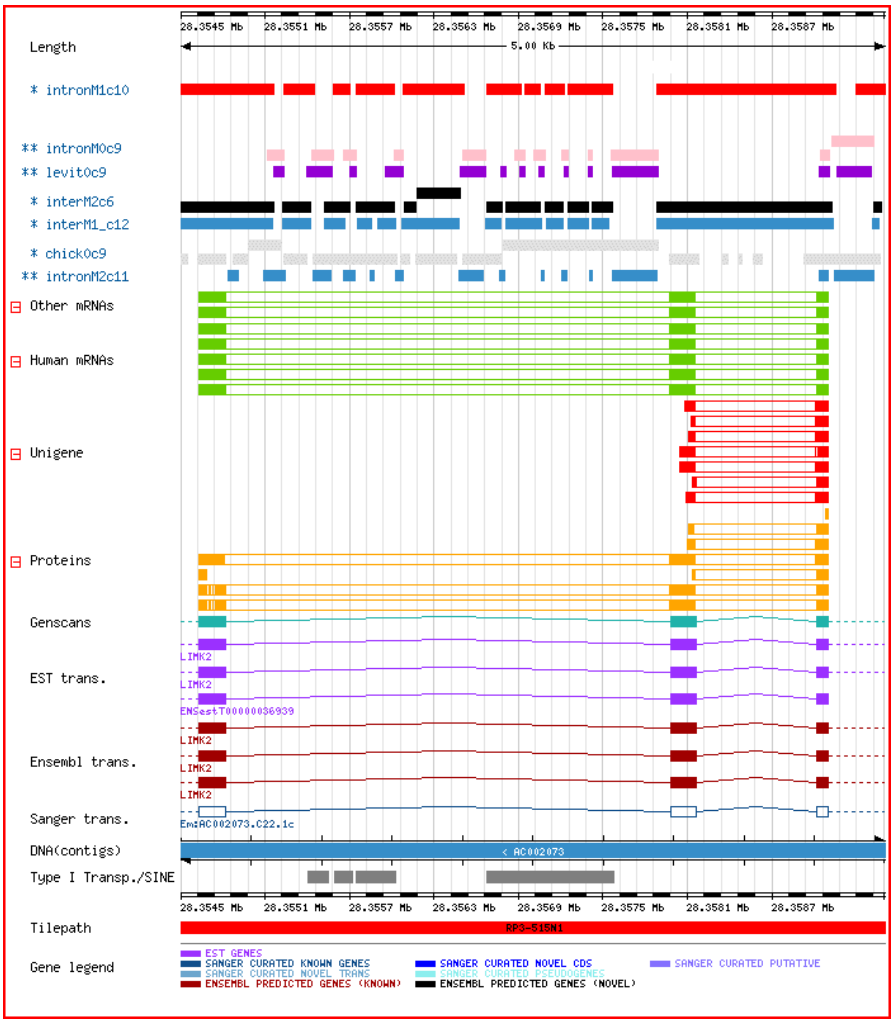
In retrospect, however, the first motif [CWG] appeared to represent the previously observed [VWG] motif (Baldi *et al.*, 1996). As seen in Figure 3.8(a) and in Appendix B, [C] always appeared to have the highest information content in the first position of this motif. This motif, is therefore, referred to as [CWG] from this point onwards. The other motif, which was being learnt, was a single strong [W] state within the wheel (Figure 3.8(b)). Although this appeared to represent a single [W]

state, this one-state motif was actually very often bounded by a very weak [C] and a very weak [G] in the bounding states (for example, model *interMask0_c10* in Appendix B). Therefore, many of these motifs were the [CWG] motifs with a much weaker [C] and [G] in the first and last positions respectively. However, the labelling properties of the 2 apparent motif-models showed that the 2 models did not behave the same way as initial impressions suggested (discussed below).

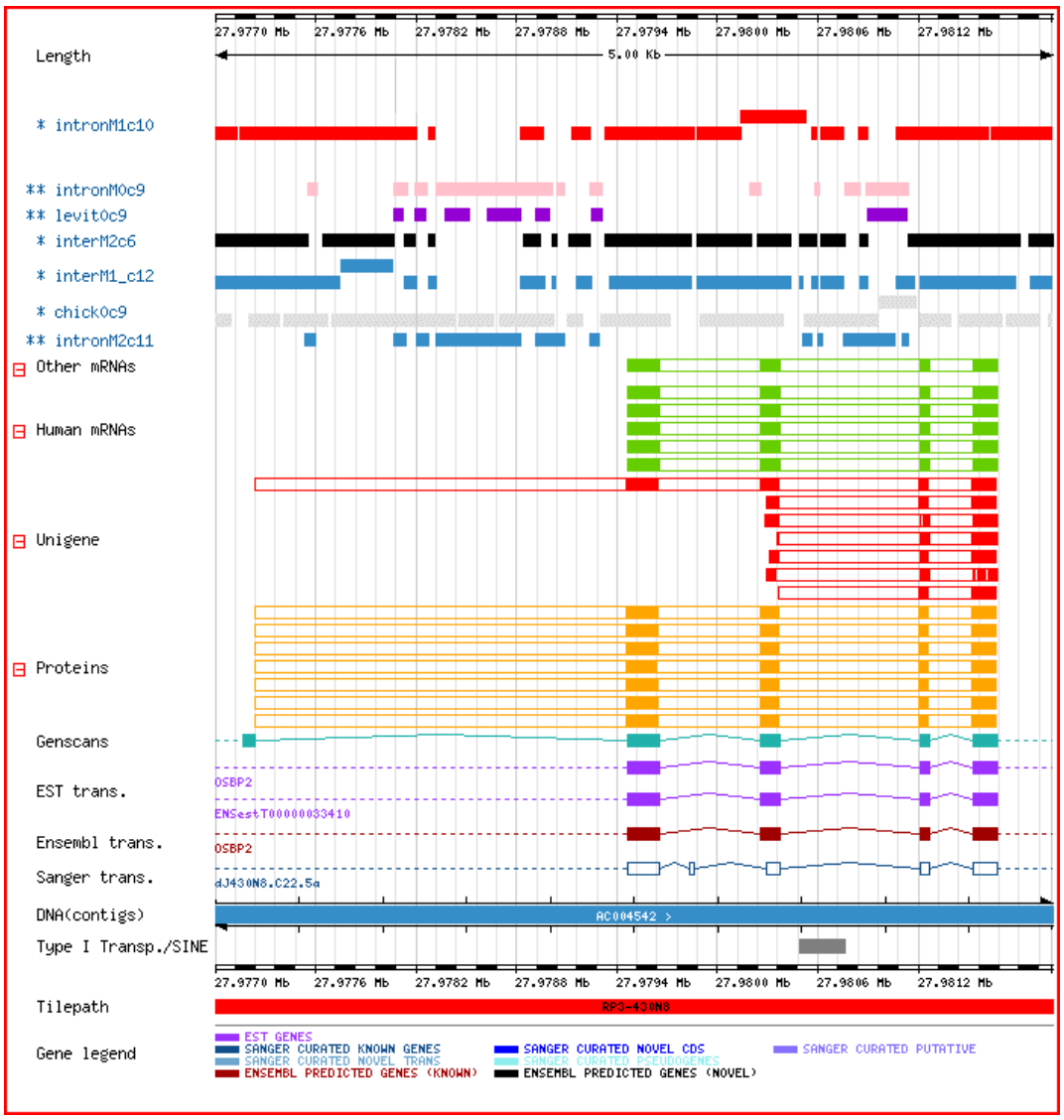
Labelling a human chromosome 22 contig with models trained from repeat-masked non-coding human sequences, showed that 2 kinds of models with complementary labelling patterns had been learnt (Figure 3.9). Figure 3.9(a)-(c) shows that there were 2 opposing labelling patterns. Of the 5 models trained from human, 3 models (*interM2_c6, intronM1_c10, interM1_c12*) labelled regions which included coding sequences (Figure 3.9(a), (b)) and SINE repeats (Figure 3.9(c)). The pattern did not appear to exclusively label coding sequences (Figure 3.9(a), (b)) but did appear to do so for the SINE repeats (Figure 3.9(c)). 2 of the other models shown (*intronM2c11, intronM0_c9*) appeared to label opposing regions labelled by the other 3 human-trained models.

Figure 3.9: Examples of Viterbi labelling a 13MB contig of human chromosome 22 using various F3 models.

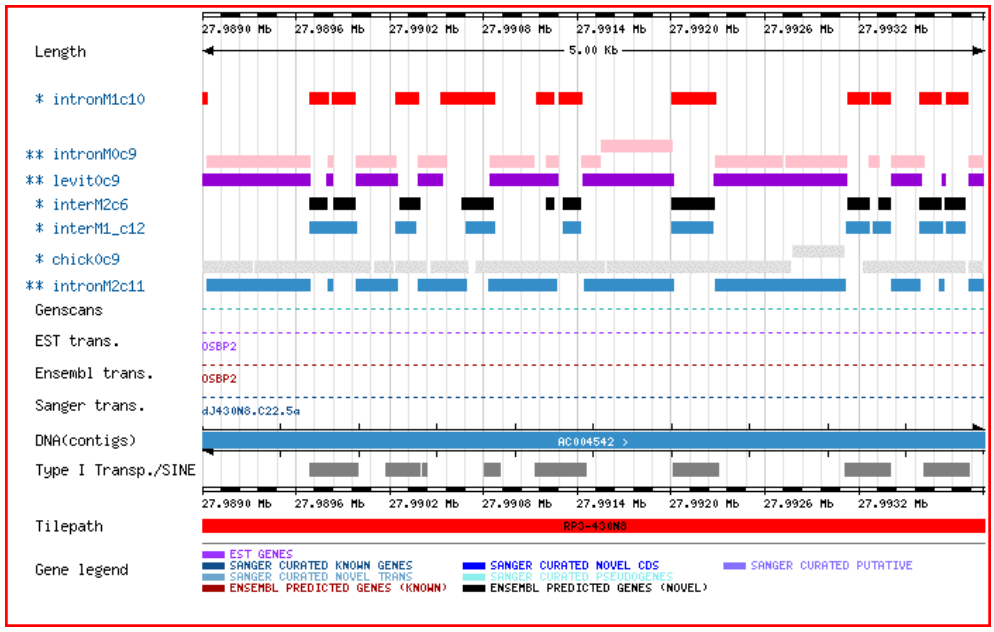
Legend: * = F3 model which learnt a [CWG] motif; ** = F3 model which learnt a [W] motif.



(a)


27.9770 Mb
27.9776 Mb
27.9782 Mb
27.9788 Mb
27.9794 Mb
27.9800 Mb
27.9806 Mb
27.9812 Mb
Length
5.00 Kb
* intronM1c10
** intronM0c9
** levit0c9
* interM2c6
* interM1_c12
* chick0c9
** intronM2c11
Other mRNAs
Human mRNAs
Unigene
Proteins
Genscans
EST trans.
OSBP2
ENSestT00000033410
Ensembl trans.
OSBP2
Sanger trans.
dJ430N8.C22.5a
DNA(contigs)
AC004542 >
Type I Transp./SINE
Tilepath
RP3-430N8
Gene legend
EST GENES
SANGER CURATED KNOWN GENES
SANGER CURATED NOVEL TRANS
ENSEMBL PREDICTED GENES (KNOWN)
SANGER CURATED NOVEL CDS
SANGER CURATED PSEUDOGENES
ENSEMBL PREDICTED GENES (NOVEL)
SANGER CURATED PUTATIVE

(b)

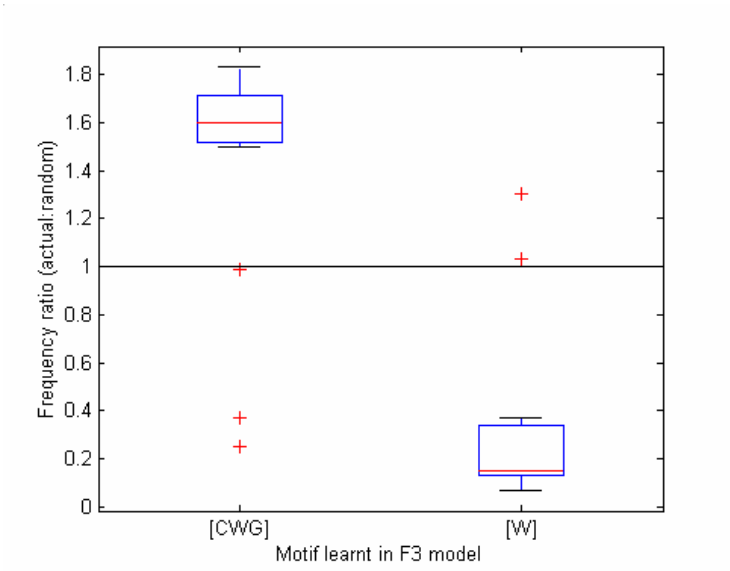

27.9890 Mb
27.9896 Mb
27.9902 Mb
27.9908 Mb
27.9914 Mb
27.9920 Mb
27.9926 Mb
27.9932 Mb
Length
5.00 Kb
* intronM1c10
** intronM0c9
** levit0c9
* interM2c6
* interM1_c12
* chick0c9
** intronM2c11
Genscans
EST trans.
OSBP2
Ensembl trans.
OSBP2
Sanger trans.
dJ430N8.C22.5a
DNA(contigs)
AC004542 >
Type I Transp./SINE
Tilepath
RP3-430N8
Gene legend
EST GENES
SANGER CURATED KNOWN GENES
SANGER CURATED NOVEL TRANS
ENSEMBL PREDICTED GENES (KNOWN)
SANGER CURATED NOVEL CDS
SANGER CURATED PSEUDOGENES
ENSEMBL PREDICTED GENES (NOVEL)
SANGER CURATED PUTATIVE

(c)

- **Labelling properties of models depended on motif learnt in the wheel**

The labelling of a human chromosome 22 contig with a 12-wheel state [CWG]-learnt model was compared with other [CWG]-learnt models of different wheel sizes (Figure 3.10). It was observed that they were mostly aligning the same parts of the test sequence. The frequency of labelling parts of the test sequence with models of different wheel sizes, but which learnt [CWG], appeared to be 1.6x greater than expected. On the other hand, comparing the alignments of models, which learnt the [W] motif, with the alignment of the same [CWG] model showed that they were aligning different parts of the test sequence (aligning the same parts 0.2x less frequently than expected). The partitioned style of labelling, therefore, depended on the motif learnt in the model and not the number of states in the wheel. A separate analysis was done to see if models, which learnt the same motif but were of different wheel sizes, were compensating to align the motif they had learnt in the same positions in the labelled sequence (results not shown). This showed that there was no such compensation. Furthermore, the skipping and labelling frequencies of the F3 models were themselves very low compared to the frequency of making a full turn around the wheel (Table 3.4, page 3-76).

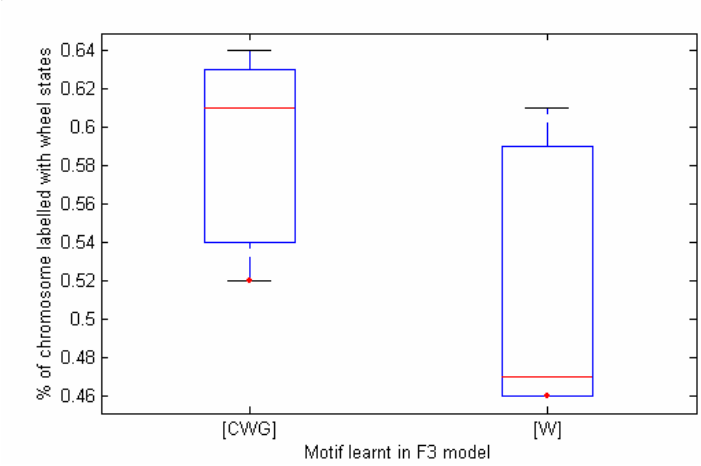
Figure 3.10: Comparison of model to model labelling. An F3 model, which had learnt a [CWG] motif (Model ID *interMask1_c12* in Appendix B), was used to label a 2.5MB sequence of human chromosome 22. The labelling of this was compared to the labelling of other models, of different wheel sizes, whose apparent motifs were either [CWG] or [W] respectively.



- **Percentage of test sequences labelled by [CWG] or [W]-learnt models**

On average, in human, 60% of the test chromosome 22 contig was labelled as wheel states by [CWG] models and 52% by [W] models (Figure 3.11); therefore, there was likely to be some overlap (~8%) between the 2 mostly opposing labelling patterns.

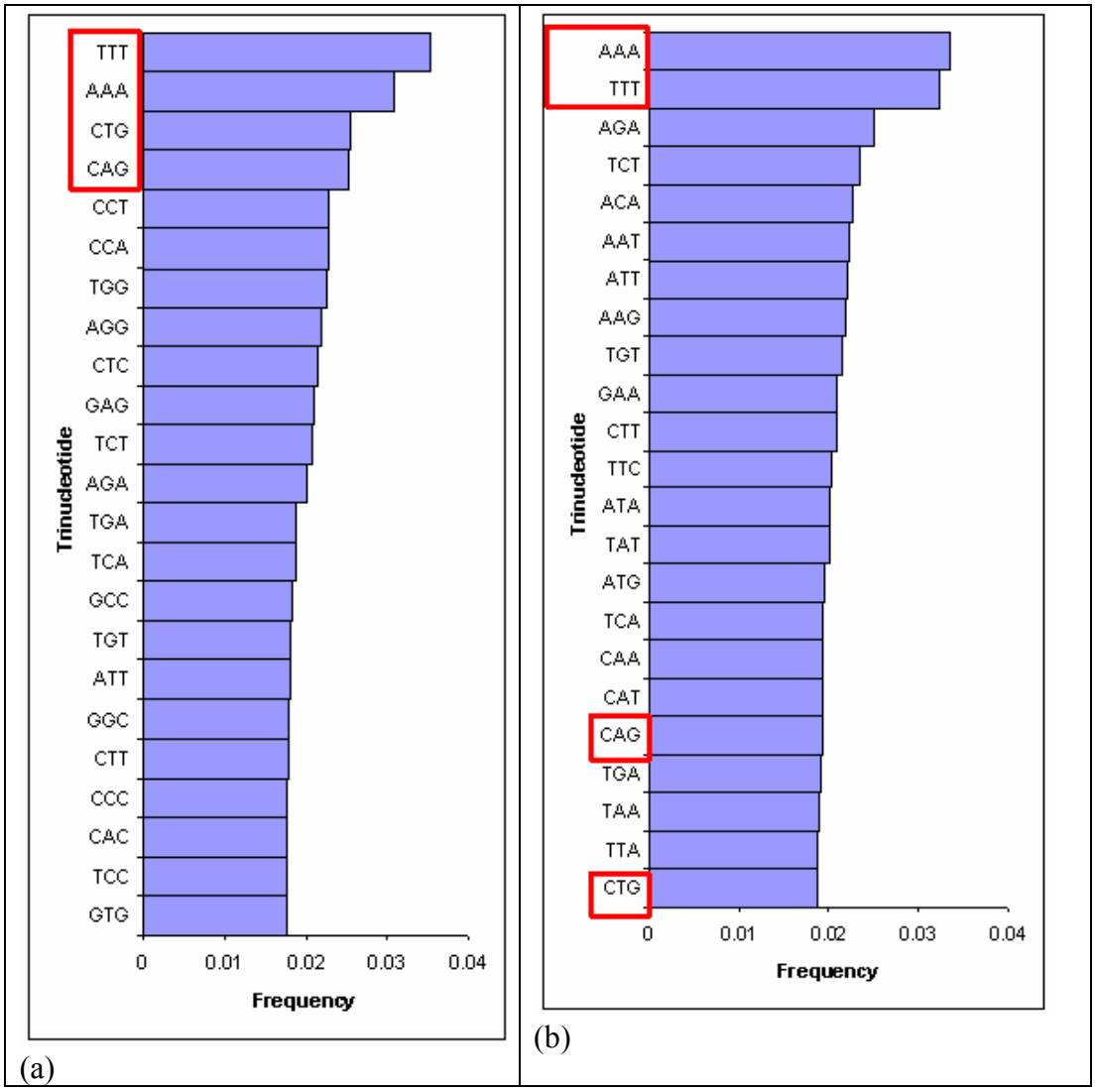
Figure 3.11: Boxplots showing percentage of genome sequence labelled as wheel states by models which learnt apparent [CWG] or [W] motifs respectively.



However, for comparison, a mouse contig of equal length was also aligned. In this case, the average density of wheel-state labelling by [CWG] and [W]-learnt

models were 33% (standard deviation: 0.22) and 81% respectively (standard deviation: 0.05) (data shown independently in Table 3.5, page 3-93). Thus, the wheel-state labelling density was significantly different for the same models in mouse and in human. A reason for this could have been the background trinucleotide density in human and mouse (Figure 3.12). Figure 3.12(a) indicates that [CWG] and [WWW] are the most frequent trinucleotides in human (motifs boxed in red). In the mouse background trinucleotide distribution, [WWW] followed by [AGA] and [TCT] are the most frequent trinucleotides (Figure 3.12(b)). Thus, the 81% wheel-state labelling by [W]-learnt models could be biased by the high content of [WWW] in the mouse genomic background. Although the labelling could have been biased by the high density of [WWW] motifs in mouse, the two motifs [CWG] and [W] were consistently learnt from repeat-masked mouse genomic DNA (data not shown). Therefore, although the labelling could possibly have been biased by the genomic trinucleotide background, the training did not appear to depend on the most frequent trinucleotides in the genomic background of the training data.

Figure 3.12: The 23 most frequent trinucleotides in the background distributions of (a) human and (b) mouse.



- **Classes of features grouped by the wheel-state labelling of the 2 motif-models**

The locations of known genomic features in the test sequences were compared to the locations of wheel state modelling by the different models. This was done for both human and mouse (Table 3.5); this showed 2 exclusive classes of features corresponding to the exclusive style of labelling.

In both the human and mouse test sequences, [CWG]-learnt models frequently "wheel-labelled" Alu sequences (B1 in mouse), exons, and the upstream regions of genes. [W]-learnt models frequently labelled repeats of the Charlie, L1 and MER

types. This partitioning of features indicated an important feature about the learnt motifs: they had not learnt a signal related to coding DNA.

The features frequently labelled by [CWG]-wheel states included exons, which are protein-coding DNA and Alu sequences, which are derived from 7SL-RNA and which do not code for proteins (HGSC, 2001). The features frequently labelled by [W]-wheel states included transposase gene-coding repeats (The DNA-transposon derived Charlie and MER class of repeats) and endonuclease gene-coding repeats (L1 LINE repeats). Therefore, all the coding-sequences had not been grouped into the same class by the wheel-state labelling of either of the 2 motif-models.

The grouping of exons and Alu repeats (and B1 repeats) into the same class was intriguing as similar properties between the 2 features had not been reported previously. However, the similarity could be due to the presence of highly diverged SINE repeats, which have become too weak for current repeat-detection programs (for example *RepeatMasker*) to detect (Smit & Green, 1997; Smit, 1999). Representative sequence members of the 2 classes were compared to see if any general differences could be noted which could account for the observations (Figure 3.13). The consensus observation from Figure 3.13 was that the Alu sequence was not as poly(dA)•poly(dT) rich as the Charlie sequence. A strongly-periodic [CWG] motif was not visually apparent in the Alu sequence though. On the other hand, the Charlie sequence showed clumps of poly(dA)•poly(dT) which could be expected from the cyclicity of the model. The periodicity of the 2 motifs is discussed subsequently (Section 3.3.7).

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:

I	intermediate
-	unknown

			HUMAN		MOUSE	
TRAIN_SOURCE	MOTIF	STATES	%cycle - labelled	Features labelled by model and the ratio of their observed to expected frequencies	%cycle- labelled	Features labelled by model and the ratio of their observed to expected frequencies
chicken0	[CWG]	9	0.73		0.77	Charlie(1.21)
exon0	[CWG]	9	0.42	AluS(1.68) AluY(1.67) Alu(1.60) Exons(1.55) up2K(1.51) Down2K(1.23)	0.09	exons(3.03) B1(2.83) up2k(1.58) introns(1.58) down2K(1.44)
exon0	[CWG]	10	0.44	AluS(1.62) AluY(1.61) Alu(1.57) Exons(1.46) up2K(1.46) Down2K(1.22) AluJ(1.20)	0.11	B1(2.77) exons(2.50) down2K(1.47) up2k(1.45) introns(1.44)
exon1	[CWG]	10	0.44	AluS(1.64) AluY(1.60) Alu(1.58) Exons(1.46) up2K(1.45) AluJ(1.22) Down2K(1.22)	0.11	B1(2.81) exons(2.53) up2k(1.48) introns(1.47) down2K(1.47)
exon2	[CWG]	9	0.42	AluY(1.66) AluS(1.66) Alu(1.59) Exons(1.56) up2K(1.50) Down2K(1.23)	0.09	exons(2.99) B1(2.82) introns(1.58) up2k(1.53) down2K(1.45)
exon2	[CWG]	10	0.72	Charlie(1.37) MER(1.23)	0.93	
inter0	[CWG]	9	0.63	AluS(1.35) Alu(1.34) Exons(1.26) AluY(1.26) up2K(1.25)	0.36	B1(1.93) exons(1.69) introns(1.28) down2K(1.20)
inter2	[CWG]	9	0.64	AluS(1.36) Alu(1.35) AluY(1.28) Exons(1.26) up2K(1.25)	0.36	B1(1.94) exons(1.71) introns(1.29) down2K(1.20)
interMasked0	[CWG]	8	0.54	AluS(1.53) Alu(1.51) AluY(1.47) Exons(1.36) up2K(1.36) AluJ(1.27)	0.20	B1(2.42) exons(2.00) introns(1.35) up2k(1.33) down2K(1.33)
interMasked0	[CWG]	11	0.52	AluS(1.44) Alu(1.43) AluY(1.42) Exons(1.37) up2K(1.36) AluJ(1.21)	0.20	B1(2.35) exons(2.09) introns(1.37) up2k(1.35) down2K(1.31)
interMasked1	[CWG]	7	0.53	AluS(1.53) Alu(1.50) AluY(1.44) up2K(1.38) Exons(1.34) AluJ(1.25)	0.18	B1(2.38) exons(2.16) introns(1.38) up2k(1.37) down2K(1.32)

interMasked1	[CWG]	8	0.54	AluS(1.52) Alu(1.50) AluY(1.47) up2K(1.35) Exons(1.34) AluJ(1.27)	0.20	B1(2.43) exons(2.04) introns(1.36) up2k(1.32) down2K(1.31)
interMasked1	[CWG]	9	0.61	Charlie(1.59) MER(1.38) L1(1.24)	0.89	
interMasked1	[CWG]	12	0.54	AluS(1.50) Alu(1.48) AluY(1.46) Exons(1.35) up2K(1.34) AluJ(1.24)	0.22	B1(2.37) exons(1.99) introns(1.33) up2k(1.29) down2K(1.27)
interMasked2	[CWG]	6	0.54	AluS(1.55) Alu(1.52) AluY(1.50) Exons(1.37) up2K(1.35) AluJ(1.23)	0.20	B1(2.40) exons(2.11) introns(1.35) up2k(1.33) down2K(1.30)
interMasked2	[CWG]	7	0.52	AluS(1.51) Alu(1.49) AluY(1.41) up2K(1.38) Exons(1.35) AluJ(1.25)	0.19	B1(2.33) exons(2.17) introns(1.37) up2k(1.35) down2K(1.32)
interMasked2	[CWG]	12	0.54	AluS(1.50) Alu(1.48) AluY(1.45) Exons(1.36) up2K(1.34) AluJ(1.24)	0.22	B1(2.37) exons(1.99) introns(1.33) up2k(1.29) down2K(1.27)
intron1	[CWG]	9	0.63	AluS(1.36) Exons(1.35) Alu(1.34) AluY(1.27) up2K(1.26)	0.35	B1(1.99) exons(1.72) introns(1.29) down2K(1.20)
intronMasked0	[CWG]	11	0.62	AluS(1.34) Alu(1.32) up2K(1.28) AluY(1.27) Exons(1.27)	0.35	B1(1.92) exons(1.71) introns(1.31) down2K(1.20)
intronMasked0	[CWG]	12	0.64	AluS(1.36) Alu(1.35) AluY(1.28) up2K(1.26) Exons(1.24)	0.36	B1(1.97) exons(1.64) introns(1.27)
intronMasked1	[CWG]	6	0.63	AluS(1.40) Alu(1.39) AluY(1.32) up2K(1.27) Exons(1.26)	0.33	B1(2.02) exons(1.69) introns(1.28) down2K(1.20)
intronMasked1	[CWG]	8	0.63	AluS(1.37) Alu(1.36) AluY(1.28) up2K(1.26) Exons(1.25)	0.35	B1(1.99) exons(1.70) introns(1.27) down2K(1.20)
intronMasked1	[CWG]	10	0.63	AluS(1.37) Alu(1.37) AluY(1.32) up2K(1.27) Exons(1.27)	0.35	B1(1.99) exons(1.67) introns(1.27) down2K(1.20)
intronMasked1	[CWG]	11	0.62	AluS(1.34) Alu(1.33) Exons(1.28) AluY(1.27) up2K(1.27)	0.35	B1(1.93) exons(1.71) introns(1.31) up2k(1.20) down2K(1.20)
intronMasked1	[CWG]	12	0.63	AluS(1.36) Alu(1.35) AluY(1.29) Exons(1.27) up2K(1.26)	0.36	B1(1.94) exons(1.65) introns(1.27)
intronMasked2	[CWG]	8	0.63	AluS(1.38) Alu(1.37) AluY(1.29) up2K(1.26) Exons(1.24)	0.35	B1(1.99) exons(1.68) introns(1.28) down2K(1.21)
chicken2	[W]	10	0.87		0.80	
inter0	[W]	10	0.51	Charlie(1.85) MER(1.50) L1(1.46)	0.82	Charlie(1.25)
interMasked0	[W]	6	0.59	Charlie(1.64) MER(1.41) L1(1.28)	0.89	
interMasked0	[W]	7	0.59	Charlie(1.66) MER(1.38) L1(1.27)	0.88	
interMasked1	[W]	6	0.59	Charlie(1.65) MER(1.40) L1(1.28)	0.89	

interMasked2	[W]	8	0.61	Charlie(1.59) MER(1.39) L1(1.24)	0.89	
intron0	[W]	10	0.65	Charlie(1.52) MER(1.23)	0.84	
intron1	[W]	10	0.66	Charlie(1.50) MER(1.20)	0.84	
intron2	[W]	9	0.68	Charlie(1.47)	0.84	
intron2	[W]	10	0.66	Charlie(1.46) MER(1.20)	0.85	Charlie(1.20)
intronMasked0	[W]	6	0.46	Charlie(2.03) L1(1.57) MER(1.48)	0.78	Charlie(1.29)
intronMasked0	[W]	7	0.46	Charlie(2.01) L1(1.59) MER(1.51)	0.76	Charlie(1.33)
intronMasked0	[W]	9	0.47	Charlie(2.00) L1(1.57) MER(1.55)	0.77	Charlie(1.30)
intronMasked1	[W]	7	0.46	Charlie(2.05) L1(1.59) MER(1.52)	0.77	Charlie(1.33)
intronMasked1	[W]	9	0.48	Charlie(1.95) L1(1.54) MER(1.51)	0.79	Charlie(1.28)
intronMasked2	[W]	6	0.46	Charlie(2.02) L1(1.58) MER(1.51)	0.78	Charlie(1.28)
intronMasked2	[W]	7	0.46	Charlie(2.02) L1(1.59) MER(1.52)	0.77	Charlie(1.33)
intronMasked2	[W]	9	0.47	Charlie(2.03) L1(1.58) MER(1.55)	0.77	Charlie(1.30)
intronMasked2	[W]	11	0.47	Charlie(1.96) MER(1.56) L1(1.56)	0.77	Charlie(1.31)
levitsky0	[W]	9	0.39	Charlie(2.23) L1(1.79) MER(1.55)	0.68	Charlie(1.44)
chicken0	I	10	0.86		0.80	
chicken1	I	9	0.76		0.77	Charlie(1.21)
chicken1	I	10	0.68	Charlie(1.28)	0.76	Charlie(1.22)
chicken2	I	9	0.85		0.78	
interMasked0	I	9	0.61	Charlie(1.59) MER(1.38) L1(1.24)	0.89	
interMasked0	I	10	0.6	Charlie(1.61) MER(1.40) L1(1.25)	0.89	
interMasked1	I	11	0.6	Charlie(1.64) MER(1.37) L1(1.26)	0.88	
interMasked2	I	9	0.61	Charlie(1.58) MER(1.38) L1(1.24)	0.89	
interMasked2	I	10	0.6	Charlie(1.61) MER(1.40) L1(1.25)	0.89	
interMasked2	I	11	0.6	Charlie(1.59) MER(1.37) L1(1.25)	0.88	
intronMasked0	I	10	0.47	Charlie(1.94) L1(1.56) MER(1.49)	0.79	Charlie(1.29)
intronMasked2	I	10	0.48	Charlie(1.92) L1(1.56) MER(1.50)	0.79	Charlie(1.28)
Alu0	-	9	0.93		0.89	
Alu0	-	10	0.93		0.89	
Alu1	-	9	0.94		0.90	
Alu1	-	10	0.94		0.89	

Alu2	-	9	0.76	Charlie(1.30)	0.81	
Alu2	-	10	0.93		0.90	
archaea0	-	9	0.46	Charlie(1.71)	0.51	Charlie(1.40)
archaea0	-	10	0.5	Charlie(1.48)	0.60	Charlie(1.39)
archaea1	-	9	0.52	Charlie(1.62)	0.61	Charlie(1.33)
archaea1	-	10	0.45	Charlie(1.71)	0.50	Charlie(1.45)
archaea2	-	9	0.52	Charlie(1.68)	0.61	Charlie(1.33)
archaea2	-	10	0.44	Charlie(1.75)	0.50	Charlie(1.41)
inter1	-	9	0.64	AluS(1.35) Alu(1.34) AluY(1.28) up2K(1.27) Exons(1.26)	0.36	B1(1.96) exons(1.71) introns(1.30)
intron0	-	9	0.68	Charlie(1.47) MER(1.20)	0.84	
levitsky0	-	10	0.34	Charlie(2.47) L1(2.02) MER(1.48)	0.60	Charlie(1.53)
levitsky1	-	9	0.75	Exons(1.22) AluS(1.21) Alu(1.20)	0.60	B1(1.50) exons(1.42)
levitsky1	-	10	0.33	Charlie(2.52) L1(2.06) MER(1.45)	0.58	Charlie(1.57)
levitsky2	-	9	0.39	Charlie(2.23) L1(1.79) MER(1.55)	0.68	Charlie(1.43)
levitsky2	-	10	0.73	Exons(1.23) Alu(1.21) AluS(1.21) AluJ(1.20)	0.57	B1(1.55) exons(1.42)

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)

```
>aluY#SINE/alu
RGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGT
CAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCC
GGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCG
GGAGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCC
GTCTCAAAAAAAAAAAAAAAAAAAAAAAAAAAA
>Charlie1b#DNA/MER1_type
CAGCGGTTCTCAAAGTGTGGTCCGNGGACCCCTGGGGGTCCCCGAGACCCTTTCAGGGGGTCCGCGAGG
TCAAAACTATTTTCATAATAATACTAAGACGTTATTTGCCTTTTTCACTCTCATTCTCTCACGAGTGTA
CAGTGGAGTTTTCCAGAGGCTACATGACGTGTGATGTCGCAACAGATTGAATGCAGAAGCAGATATGAG
AATCCAGCTGTCTTCTATTAAGCCAGACATTAAAGAGATTTGCAAAAATGTAAAACAATGCCACTCTTC
TCACTAAATTTTTTTGTTTTGGAAAATATAGTTATTTTTCATAAAAATATGTTATTTATGTTAACATGT
AATGGGTTATTATTATTTTTAAATGAATTAATAAATATTTTAAAAATTTCTCAGTTTTAATTTCTAATA
CGGTAAATATCGATAGATATAACCCACATAAACAAAAGCTCTTTGGGGTCCTCAATAATTTTTAAGAGT
GTAAAGGGGTCCTGAGACCAAAAAGTTTGAGAACCGCTG
```

- **Lengths of wheel-labelled regions**

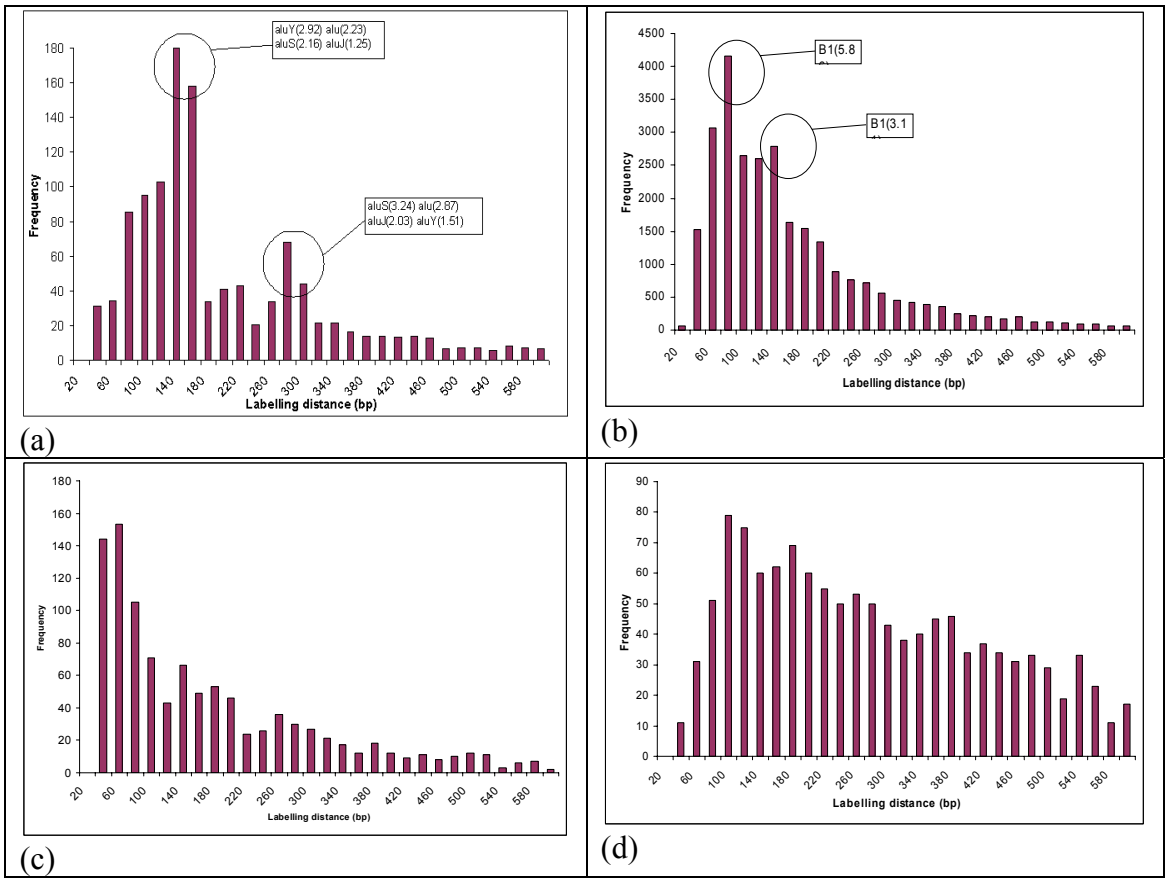
The lengths labelled by the 2 kinds of motif-learnt models were also compared in the range of 20–600 bp (Figure 3.14). This range was selected to scan for peaks which could resemble the length of a nucleosome (~146 bp). [CWG] model-labelling showed 2 distinct peaks in human: one was around 140-160 bp and another was around 300 bp (Figure 3.14(a)). In mouse, peaks were observed around 100 and 200 bp (Figure 3.14(b)) for [CWG]-wheel state labelled lengths. These peaks resembled "nucleosome-size" lengths. However, further analysis of the peaks showed that they were 3 times more frequently associated with Alu repeats than expected in human

(balloon text in Figure 3.14(a)). Similar results were observed for B1 repeats in the mouse peaks (Figure 3.14(b)).

Alu sequences are typically around 300 bp long; therefore, the two peaks most probably resembled half and full Alu lengths in human. This could be expected as Alu sequences have a polyA linker, of varying lengths, around position 150 bp in their sequence (Figure 3.13). From the opposing-style labelling observed, it could be expected that this polyA linker would not be labelled by the wheel part of the [CWG] models but by the wheel part of [W] models. This could account for the 2 observed peaks corresponding to full and half-Alu lengths. B1 repeats are half the size of Alu repeats; this could be why their [CWG]-wheel state labelling lengths appeared to be around 100 / 200 bp (Figure 3.14(b)).

[W] wheel-labelled lengths did not show any peaks within this range in human (Figure 3.14(c)). In mouse, however, peaks around 146 and 220 were apparent (Figure 3.14(d)); these peaks were not frequently associated with any repeats or known genomic features. However, the lack of similar peaks in human indicated that it was not a conserved feature.

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: *intronM1_c10* (Appendix B)). (c), (d) show data for human and mouse genomic sequences respectively, which were labelled with a [W]-learnt model (Model ID: *intronM2_c11* (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).



3.3.5 F3 model training results from Archaea and the 2 nucleosome datasets

The non-human training data included archaeal sequences, a set of chicken nucleosome sequences and Levitsky *et al*'s compilation of mapped nucleosome sequences from various organisms; a few of these models appeared to have similar properties to those learnt from the human training sets. Only 9 and 10 state F3 models were trained for these.

- **Models trained from Archaea**

9-state and 10-state models, trained from archaea, mainly learnt its background sequence composition which was poly-[W] rich (models shown in Appendix B). Archaea was an interesting organism to scan for nucleosome rotational positioning as SELEX-enrichment experiments had previously shown that DNA sequences, which bound histones in Archaea, were 10-periodic in [AA] motifs (Bailey *et al.*, 2000). This pattern was seen for the majority of the wheel states. This result probably arose from using a random DNA background model instead of the background archaeal sequence for all the emission states. However, models, which were trained using a background model of the *Archaeal* genome, showed similar results to using a random DNA background (results not shown). Therefore, enriched periodicities of ~9 or 10 bp could not be learnt for this organism using cyclical HMM-training. Aligning a human genomic sequence with these archaeal models wheel-labelled the sequence at roughly 50%; only Charlie repeats were labelled at a rate greater than expected (Table 3.5). The abundance of poly(dA)·poly(dT) regions in the example Charlie sequence (Figure 3.13) could account for this high rate of labelling using such a poly[W]-learnt model.

- **Models trained from the chicken nucleosome dataset**

For 9-10 state cyclical HMMs trained from the chicken nucleosome dataset, the [W] and [CWG] motifs were often seen; however, they were associated with a few other weak and inconsistent motifs (Appendix B). A difference between the models learnt in chicken and those learnt in human was that the chicken models learnt a strong [A] or strong [T] motif in the *Null* state whereas the *Null* state emission distributions in human-trained models were relatively flatter. The labelling properties of the chicken models were consequently different to sequences trained from human

(Table 3.5). Genomic sequences were usually labelled >76% with chicken models whereas this value was between 46-64% for human models. Therefore, although the wheel parts of the chicken models appeared similar to human, the *Null* state was different. The models were, therefore, not equivalent to those trained from human. The chicken models labelled human genomic sequences randomly with respect to known repeat types and coding regions (Table 3.5).

- **Models trained from the Levitsky dataset**

Models trained from Levitsky *et al*'s compiled nucleosome dataset learnt predominantly poly[W] motifs (Appendix B). Similar to the [W]-motif-learnt models trained from human data, many of the Levitsky models learnt [W] motifs in the wheel states and labelled the same genomic regions (Table 3.5). However, the [W] motif appeared in a number of wheel states rather than in a single wheel state as in human models. Similar to the human [W] models, *levitsky0_c9, levitsky2_c9, levitsky0_c10* and *levitsky1_c10* labelled MER and L1 repeats at a rate greater than random (Table 3.5, Figure 3.9); but wheel-state labelling was roughly 33% for these compared to 44% for the human [W] models. 2 models, *levitsky1_c9* and *levitsky2_c10* labelled complementary regions to the aforementioned models (wheel state labelling roughly 74%) (Table 3.5). Furthermore, they were enriched for the same features as the human [CWG] models (exons and Alu repeats). However, the Levitsky models did not learn a [CWG] motif in their wheel. The complementary labelling was more likely due to these last 2 models learning a [W] motif in their *Null* states. Therefore, although the labelling results suggested two complementary models like the human-trained models, the Levitsky models did not learn a counterpart [CWG] motif in their wheel components. The complementary behaviour was more likely due to modelling poly[W] motifs in the wheel as opposed to modelling [W] motifs in the null state.

3.3.6 Labelling analysis of chicken nucleosome sequences and chicken genomic sequences

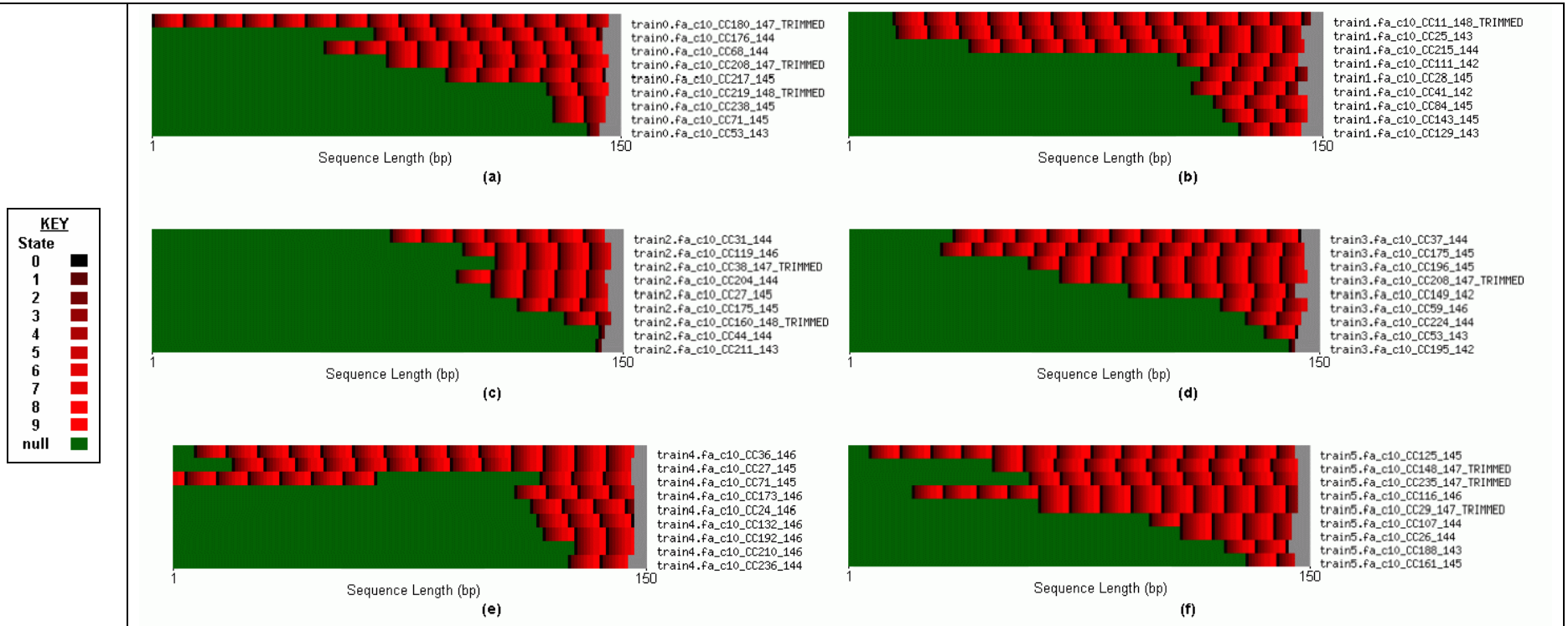
Labelling chicken nucleosome and genomic test sequences using chicken nucleosome-trained models highlighted some differences in the 2 types of test sequences. The models that were used to perform the alignments had all learnt [CWG] within the wheel component of the model.

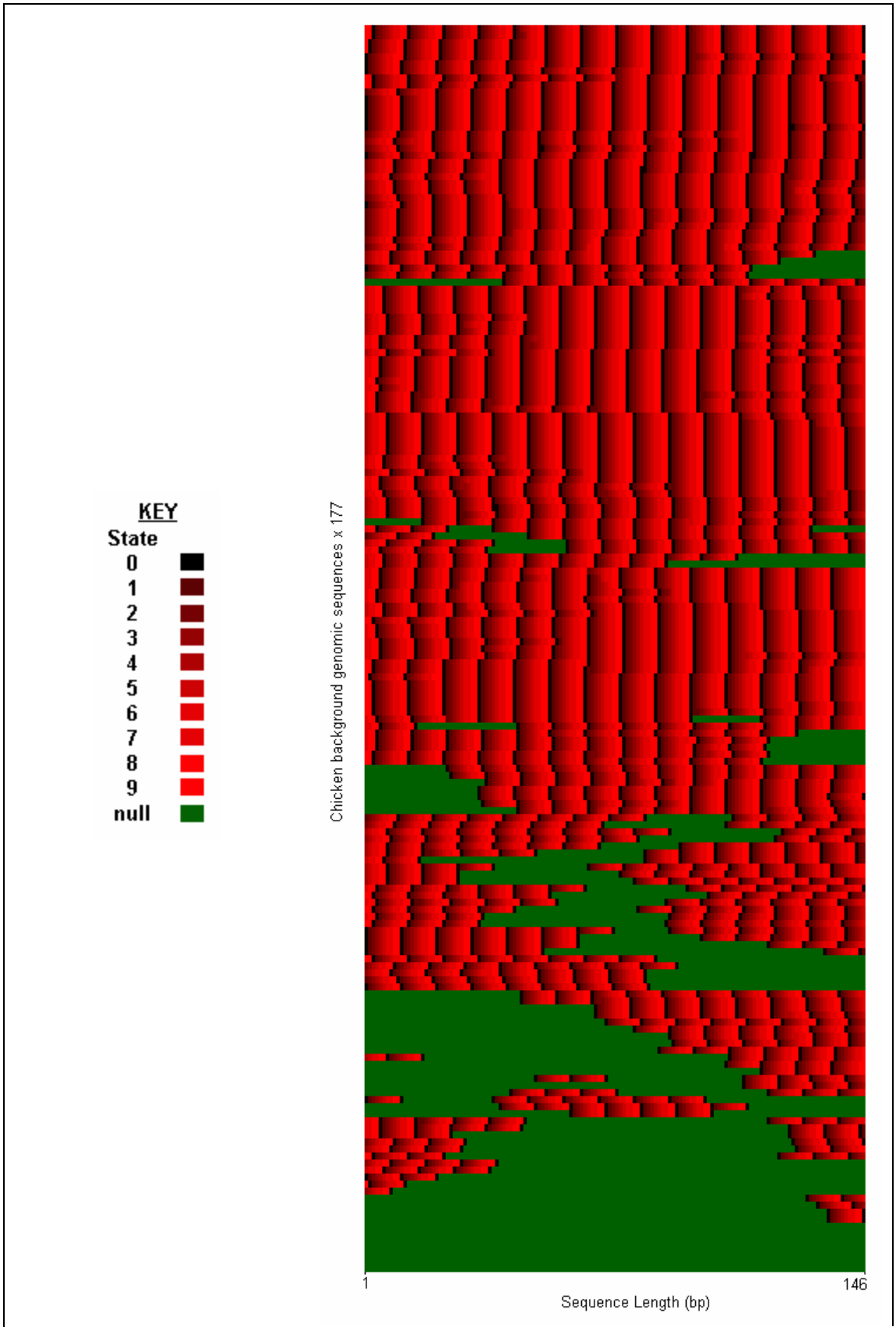
- **Alignment of chicken nucleosome sequences**

Firstly, the labelling of 10 chicken nucleosome test sequences, using a jack-knifing approach, showed that most times, only 1 or 2 sequences were aligned completely with wheel states (Figure 3.15(A)). The fact that only 1 or 2 sequences showed near 100% wheel-state labelling suggested that full turns of 10-phased [CWG] motifs around the complete core particle sequence was an unlikely requirement. Most of the other sequences showed mainly scattered labelling patterns but showed a slight bias to label the right ends of the sequences. Why there appeared to be this bias to label the ends of the sequences was not clear. Labelling of the genomic sequences did not show this kind of a bias though (Figure 3.15(B)).

The results of aligning the nucleosome sequences indicated no evidence of rotational positioning (10 bp-phasing) of the [CWG] motif. This was also the conclusion of the published analysis of the chicken nucleosome dataset (Satchwell *et al.*, 1986). Also, there did not appear to be any preference for the wheel states to align symmetrically about the centre of the sequences; this is understood about the [AA/TT] rotational positioning motif. However, the [CWG] motif was learnt from this same dataset so it could have some influence on nucleosome positioning; this data is too limited to suggest a possible mechanism though.

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.





(B) Background chicken genomic sequences

- **Alignment of chicken genomic sequences**

Aligning chicken genomic sequences with chicken nucleosome-trained models showed that ~60% of the sequences were labelled with almost 100% wheel-state labelling (Figure 3.15(B)). Only ~5% of sequences were not labelled at all with wheel states. Originally, it was expected that aligning the nucleosome test sequences would have shown 100%-wheel labelling if the [CWG] motif was involved in rotational positioning in the dataset. Instead observing it in the genomic sequences suggested that some aspect of [CWG] density and not necessarily any kind of preferential rotational positioning might have consequences for nucleosome positioning. This led to the analysis of [CWG] density (Section 3.3.8) and further analysis of the background trinucleotide distribution in different genomes and the 2 nucleosome datasets (Section 5.3.3).

3.3.7 Analysis of periodicity of the two opposing motifs

The 2 motifs, [CWG] and [W], were learnt using model architectures of a range of wheel sizes (6–12 states). Therefore, it was possible that the motifs themselves may occur quite regularly, with their periodicity corresponding to these different wheel sizes. However, to be an important motif for the rotational positioning of nucleosomes, it needed to be more strongly periodic at 10 bp compared to the other repeat periods. This made it interesting to investigate the periodicity of these motifs.

- **Model skipping and looping behaviour**

Firstly, there were no skips or loops observed for models in the wheel size range of 6–10 states (Table 3.4, page 3-76). However, for 11 and 12 state wheel models, which had learnt the [CWG] motif, a low frequency for looping was observed. This suggested that the models were probably trying to fit a higher-order wheel size to the wheel size-range examined. Analysis of an F2 model and an F3-

reproduced B&B model, however, suggested that 10 state wheel models had a slight tendency to skip to fit a 9 wheel (Section 3.3.3).

- **Forward scores of models of different wheel sizes**

The periodicity was investigated secondly by labelling both repeat-masked intergenic and coding DNA sequences with models of different wheel sizes and comparing their forward scores (Figure 3.16). For models, which learnt the [CWG] motif, the 9 and 10-state wheel models labelled intergenic sequences with a slightly better average forward score than the other wheel sizes (Figure 3.16(a)). In coding sequence, however, these same peaks were not seen (Figure 3.16(b)). There did appear to be a peak for the 6 state-models though, which suggests that the observation may be influenced by coding bias.

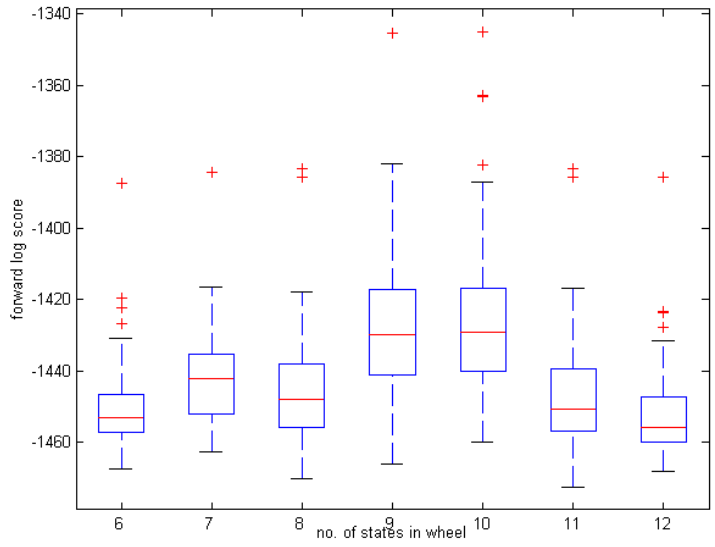
Models, which learnt the [W] motif, however, did not have any models of a specific wheel size which appeared to score better than the others (Figure 3.16(c)). So the [CWG] motif may have an enrichment at 9 and 10 bp in intergenic DNA but the [W] motif appeared random over the range of 6–12 bp; this suggested that the wheel states of the [W] models could be labelling mainly long runs of [W].

Figure 3.16: Boxplots of forward scores of test sequences labelled with F3 models of different wheel sizes.

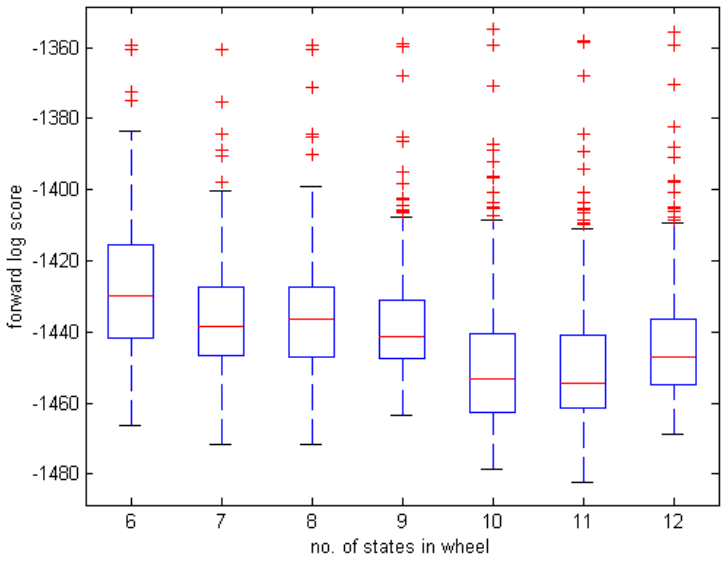
(a) Masked intergenic DNA labelled with [CWG]-learnt models,

(b) coding DNA labelled with [CWG]-learnt models and

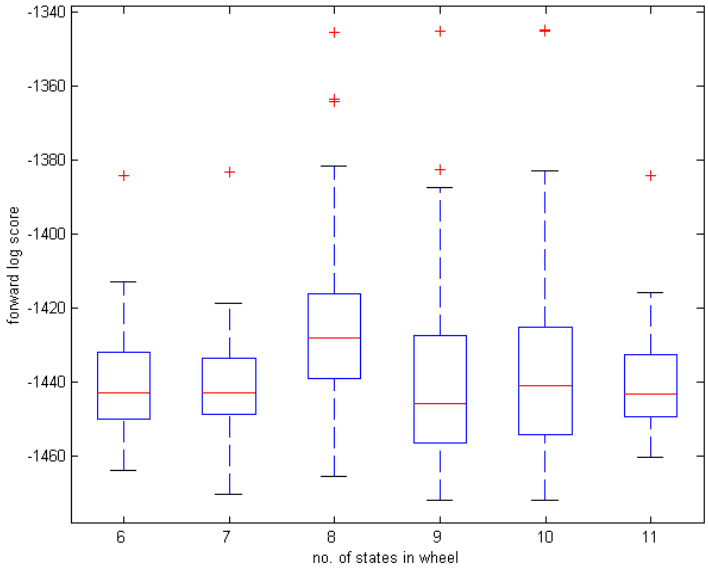
(c) masked intergenic DNA labelled with [W]-learnt models



(a)



(b)



(c)

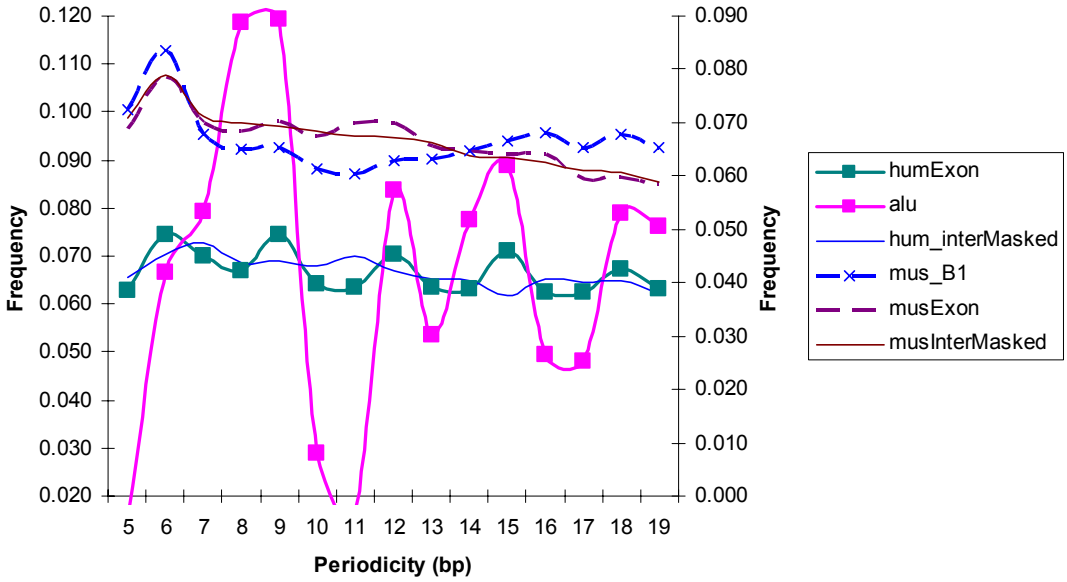
- **Motif-spacing frequency**

The final investigation of motif periodicity was to just calculate the frequencies of their repeat periods in different sequence types (Figure 3.17). For the [CWG] motif, the Alu sequences showed quite distinct periods at 8, 9, 12, 15 and 18 bp (Figure 3.17(a)). However, these peaks for Alu repeats seemed to weakly correlate with the same peaks in exons (correlation co-efficient: 0.62). The peaks in exons were, however, 3 modulo repeats which suggested effect of coding bias. This could explain why the [CWG]-motif models seemed to consistently wheel-label both Alu repeats and exons despite the fact that Alus do not code for proteins (Table 3.5). The peaks for mouse B1 repeats and mouse exons also appeared to visually correlate with each other but the correlation co-efficient was much weaker (0.46).

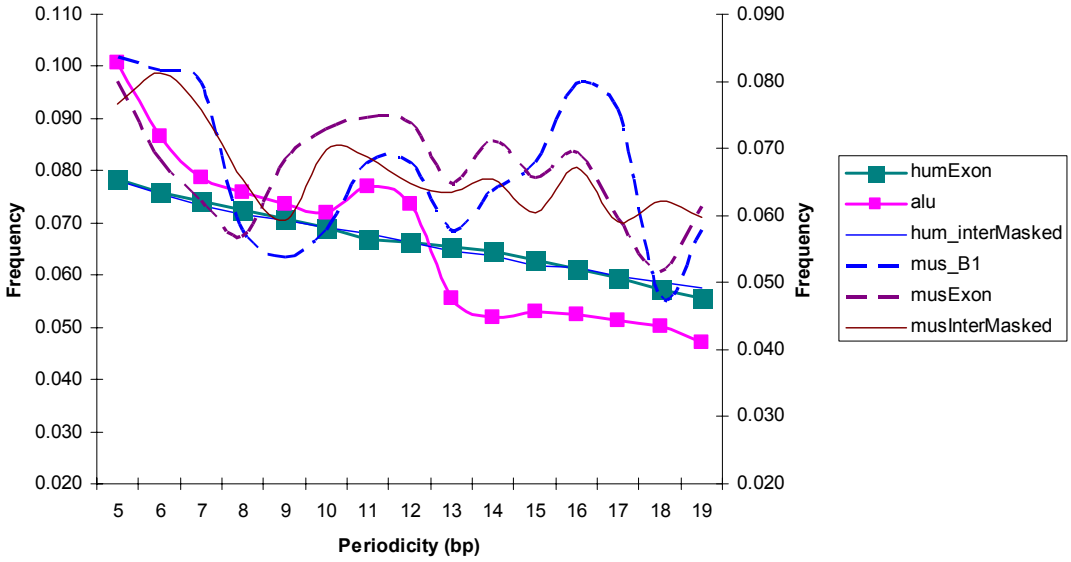
The repeat frequencies of [WWW][14] motifs, on the other hand, did not show any peaks which could suggest coding bias (Figure 3.17(b)).

[14] The periodicity of [WWW] motifs was calculated, rather than [W], because just counting [W]-occurrences would not have been informative.

Figure 3.17: Analysis of motif periodicity using a simple counting procedure: (a) [CWG] motif and (b) [WWW] motif



(a)



(b)

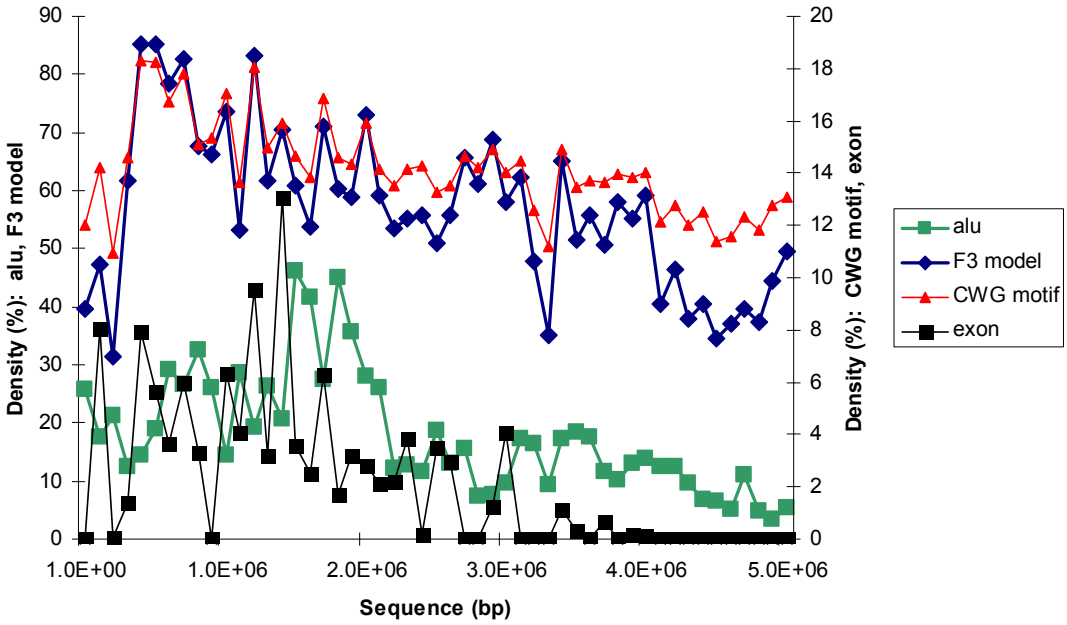
The overall impression was that the [CWG] motif did appear to be influenced by coding bias as a 3-modulo repeat of the pattern was observed. It was seen to be enriched at certain periodicities (8, 9, 12, 15 and 18 bp in human; 6, 9, 12, 16, 18 bp in mouse) and this appeared to be common for both exons and SINE repeats.

3.3.8 Labelling density of [CWG]-learnt models

The fact that different wheel-size F3 models, which learnt the same motif, all frequently "wheel"-aligned the same parts of the test sequences (Section 3.3.4) suggested that they were labelling regions having high density of the [CWG] motif. The model wheels did not skip or loop that frequently to fit other wheel sizes either (Table 3.4). To verify this, the density of a [CWG]-learnt model's wheel state labelling and windowed [CWG] density was compared (Figure 3.18). This showed that the two were correlated (correlation co-efficient: 0.98). Only these 2 variables, in Figure 3.18, appeared to be correlated. Alu and exon densities[15] did not correlate with these densities (Figure 3.18). In Figure 3.18(a), [CWG] density was seen to vary between 10 and 18%. Similar frequencies were obtained for [CWG] density in the chicken nucleosome dataset (data not shown). However, only the weak 9,10 bp-periodicity of the [CWG] motif, discussed earlier (Section 3.3.7), could suggest that the motif could be involved in rotational positioning. Models, trained and tested from the chicken nucleosome dataset, however, did not support this (Section 3.3.6).

[15] Genomic features earlier shown to be wheel-state labelled with [CWG]-learnt models (Table 3.5)

Figure 3.18: (a) Plot of a [CWG] motif-learnt 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.



(a)

	alu	F3 model	CWG motif	exon
F3 model	0.20	1.00	0.98	0.53
CWG motif	0.17	0.98	1.00	0.57

(b)

- **Windowed analysis of [CWG] motif density**

As discussed above, the [CWG]-learnt F3 models were also labelling [CWG] dense regions. Multiple expansion repeats of [CTG][16] had been seen to position nucleosomes experimentally (Section 1.5.2) although its exact mechanism in this was still unclear. Therefore, a scan was done to examine which parts of human genomic sequences frequently contained dense "blocks" of [CWG] (Figure 3.19). The highest densities that were found were around 35% within windows of 200 bp[17] (corresponding to 23 repeats of [CWG]). These dense windows appeared often, occurring once every 240 kbp in human genomic sequences and once every 300 kbp

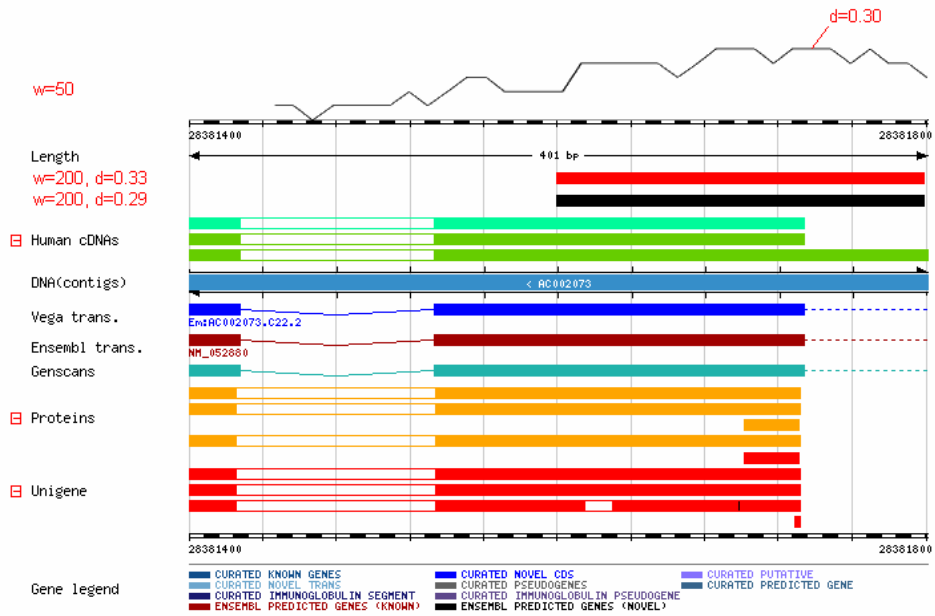
[16] A sequence member of the [CWG] motif
[17] A window size of 200 bp was chosen since it was close to ~146 bp, the nucleosome core particle size

in mouse sequence (data not presented). The features which were most frequently represented in these [CWG]-dense regions though included exons in both mouse and human (Table 3.6). This could perhaps explain Baldi and Brunak's observation of [VWG] motifs most often in coding sequence (Section 1.9.3) and the frequent labelling of exons shown earlier (Table 3.5).

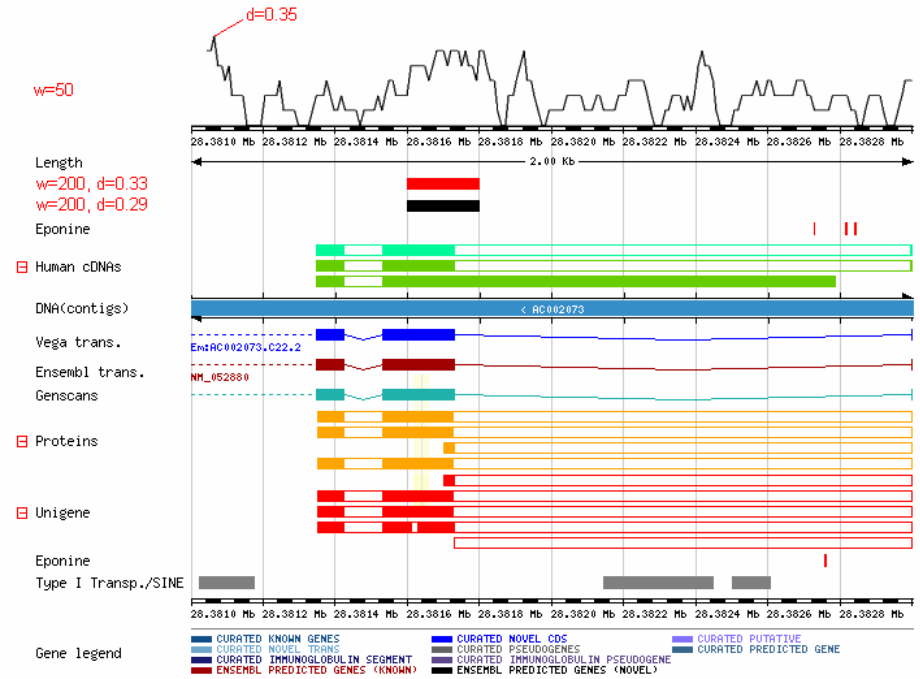
Table 3.6: Features observed to frequently have high densities of [CWG] repeats. A window size of 200 bp and cutoff threshold of 35% [CWG] density was used.

Genomic Sequence	Frequency ratio (Observed:Expected)
Human	Exons(1.37)
Mouse	Exons(2.50), Introns(1.31)

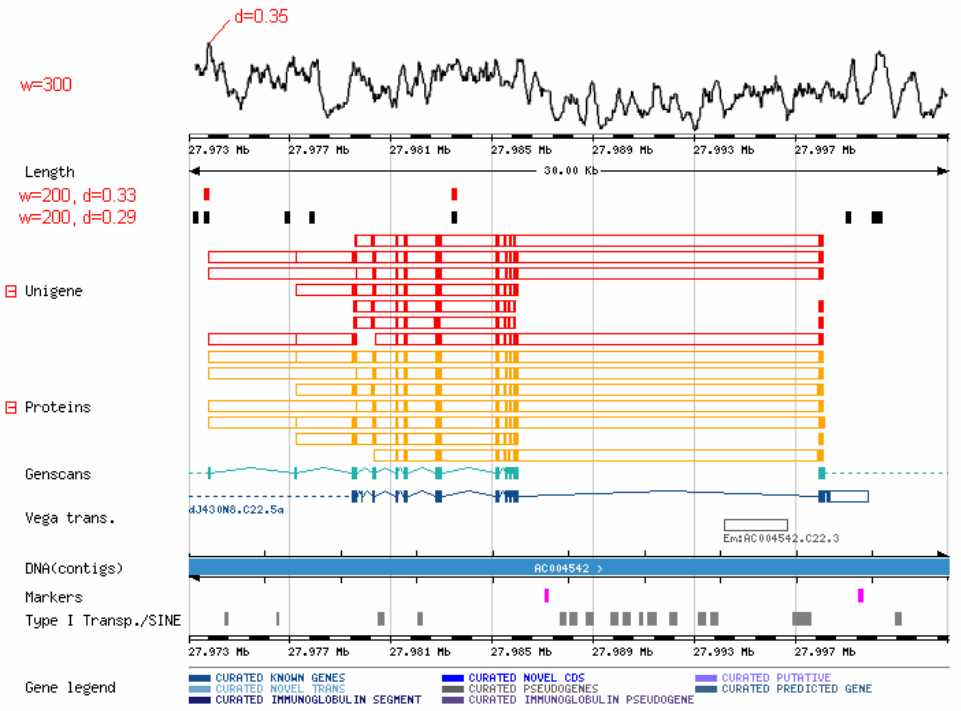
Figure 3.19: Density plots of [CWG] repeats in a human genomic sequence shown at different resolutions. 'w' is the window parameter and 'd' the threshold density of [CWG] within the window. The top density plot is a 'moving average' representation. The red and black boxes below represent non-overlapping 200 bp windows having >0.33 and >0.29 [CWG] densities respectively.



(a)


d=0.35
w=50
28.3810 Mb 28.3812 Mb 28.3814 Mb 28.3816 Mb 28.3818 Mb 28.3820 Mb 28.3822 Mb 28.3824 Mb 28.3826 Mb 28.3828 Mb
Length
2.00 Kb
w=200, d=0.33
w=200, d=0.29
Eponine
Human cDNAs
DNA(contigs)
< AC002073
Vega trans.
EmiAC002073.C22.2
Ensembl trans.
NM_052880
Genscans
Proteins
Unigene
Eponine
Type I Transp./SINE
28.3810 Mb 28.3812 Mb 28.3814 Mb 28.3816 Mb 28.3818 Mb 28.3820 Mb 28.3822 Mb 28.3824 Mb 28.3826 Mb 28.3828 Mb
Gene legend
CURATED KNOWN GENES
CURATED NOVEL TRANS
CURATED IMMUNOGLOBULIN SEGMENT
ENSEMBL PREDICTED GENES (KNOWN)
CURATED NOVEL CDS
CURATED PSEUDOGENES
CURATED IMMUNOGLOBULIN PSEUDOGENE
ENSEMBL PREDICTED GENES (NOVEL)
CURATED PUTATIVE
CURATED PREDICTED GENE

(b)


d=0.35
w=300
27.973 Mb 27.977 Mb 27.981 Mb 27.985 Mb 27.989 Mb 27.993 Mb 27.997 Mb
Length
30.00 Kb
w=200, d=0.33
w=200, d=0.29
Unigene
Proteins
Genscans
Vega trans.
dJ430N8.C22.5a
EmiAC004542.C22.3
DNA(contigs)
AC004542 >
Markers
Type I Transp./SINE
27.973 Mb 27.977 Mb 27.981 Mb 27.985 Mb 27.989 Mb 27.993 Mb 27.997 Mb
Gene legend
CURATED KNOWN GENES
CURATED NOVEL TRANS
CURATED IMMUNOGLOBULIN SEGMENT
CURATED NOVEL CDS
CURATED PSEUDOGENES
CURATED IMMUNOGLOBULIN PSEUDOGENE
CURATED PUTATIVE
CURATED PREDICTED GENE

(c)

3.4 Conclusion

Some interesting properties of the [CWG] motif have been observed. The motif represents some of the most frequent trinucleotides in the background trinucleotide density of human but not in mouse. However, the motif could also be learnt from mouse training sequences.

The evidence for this motif for effecting nucleosome rotational positioning remains unclear. Cyclical HMM results, trained using a flexibility emission alphabet, could not learn any motifs which were spaced around 9 or 10 bp (Section 3.3.2). This could mean that the background flexibility is in general not significantly different to the flexibility of [CWG], the motif which is learnt most often using models of the DNA alphabet. Also, the labelling of [CWG]-learnt models on chicken nucleosome sequences did not suggest any rotational preferences for this motif. A weak 9, 10 bp-periodicity of [CWG] was however seen in repeat-masked intergenic sequences (Section 3.3.7), which could indicate the presence of weak rotational positioning motifs.

High [CWG] density could be a factor in positioning nucleosomes though; multiple expansion repeats of [CTG] was seen to exhibit a high nucleosome density in previous research (Section 1.5.2). High windowed densities of this motif were seen in exons, which potentially suggests that exons could be preferentially wrapped in nucleosomes.

A simplistic suggestion could have been that [CWG]-dense regions, with a weak 9/10 bp periodicity, represented a greater density of nucleosomes (not necessarily positioned) whereas [W] dense regions did not. However, the comparison of the labelling properties were not the same (60% and 30% [CWG]-wheel state labelling in human and mouse respectively).