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 Chemistry[1]

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

List of Figures

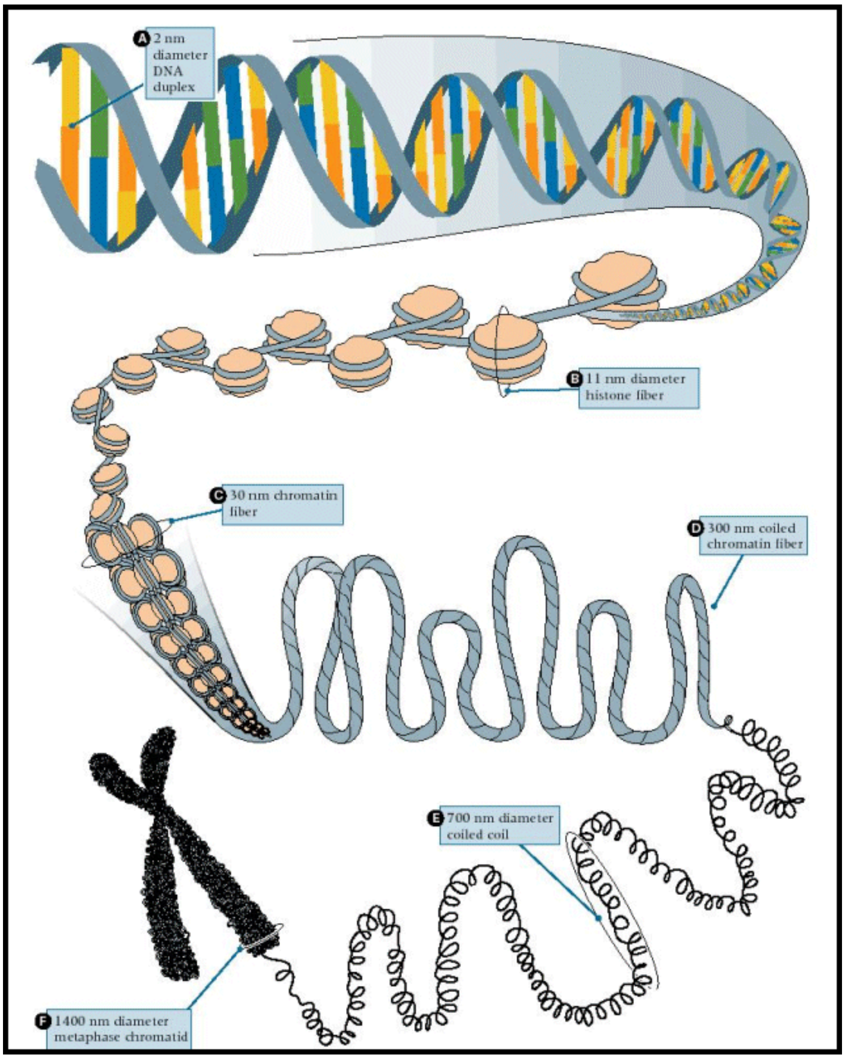
List of Tables

1 A General Introduction to Nucleosomes and Nucleosome Positioning

1.1 Nucleosomes: the Building Blocks of Chromatin

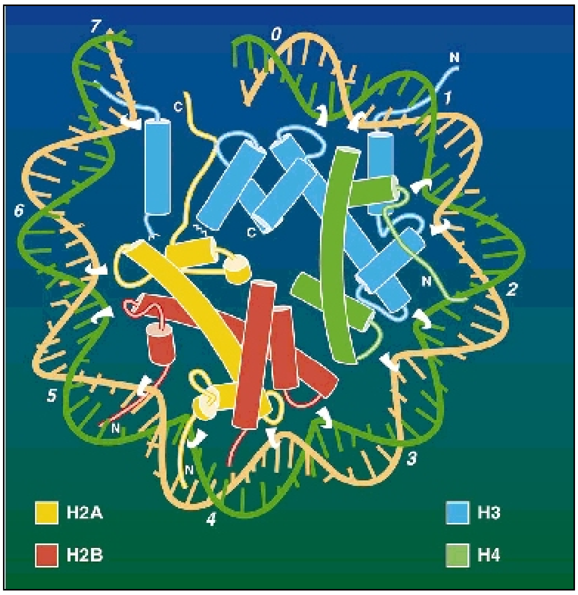
Chromatin is the complex of DNA and cellular proteins which form eukaryotic chromosomes. It is composed of an elementary repeating unit called the nucleosome, which is the major factor of DNA packaging in eukaryotic genomes (Figure 1.1).

Figure 1.1: A hierarchical view of chromatin structure. Reproduced figure (Hartl & Jones, 1998).


A 2 nm diameter DNA duplex
B 11 nm diameter histone fiber
C 30 nm chromatin fiber
D 300 nm coiled chromatin fiber
E 700 nm diameter coiled coil
F 1400 nm diameter metaphase chromatid

Nucleosomes are DNA-protein complexes, which are comprised of a *core particle* of 1.6 left-handed turns of DNA (roughly 146 bp) wound around a protein complex called the *histone octamer* (Figure 1.1(B)). The histone octamer is a set of 8 basic proteins, which are among the most well conserved proteins known in eukaryotes. It is comprised of a central tetramer, $(H3/H4)_2$, flanked by two H2A/H2B dimers (Figure 1.2). The structure of a single histone molecule includes three major α helices with positively-charged loops protruding at the N-terminals.

Figure 1.2: Top-level view of a nucleosome. Cylinders indicate alpha-helices; white hooks represent arginine/lysine tails. Reproduced figure (Rhodes, 1997)).



The DNA wrapped around the histone octamer is called the *core DNA* and the DNA joining adjacent nucleosomes is called *linker DNA*. Unlike core DNA, linker DNA exhibits great variability in length: anywhere between 8 to 200 bp. This variation in the length of linker DNA may be important for the diversity of gene regulation; however, chromatin structure formation is independent of the length of linker DNA (Kornberg & Lorch, 1999).

The constraint of the nucleosome on the DNA path forms the first level of higher-order packing, compacting DNA by a factor of ~6 (Lewin, 2000). An extra histone H1 (also called the linker histone) may also be present, clamping the DNA at the position at which it enters and leaves the histone core (Karrer & VanNuland, 1999; Satchwell & Travers, 1989; Widlund *et al.*, 2000).

The series of nucleosomes along a DNA sequence then coil into a helical array forming a fibre of ~30 nm (Figure 1.1(C)); this results in further compaction by a factor of ~40. In the recent crystal structure of the nucleosome (Luger *et al.*, 1997), it had been reported that the basic tail of H4 protrudes extensively and makes contacts with acidic patches of H2A and H2B on neighbouring octamers; this implies a role for H4 in stabilizing higher level structures. Histone H1 is thought to appear mainly towards the middle of the 30 nm fibre where it may play a role in stabilizing chromatin interactions (Staynov, 2000). Specialised nucleosomes are also known, for example the centromere-specific nucleosomes, which contain a variant of histone H3 called CENP-A; these occur in a range of organisms from yeast to human (Smith, 2002). Many non-histone chromatin proteins also interact with histones to enable formation of higher-order structures. The fibre itself undergoes further levels of packaging resulting in compaction by a factor of ~1000 in interphase euchromatin and ~10,000 in heterochromatin (Figure 1.1(D-F)).

The structure of chromatin is dynamic. It exists in a number of distinct functional states which can often be characterised by the level of transcriptional activity. The dynamic transitions between these states occur through a range of post-translational modifications of the histone tails which includes acetylation and phosphorylation (Jenuwein & Allis, 2001). This forms the basis of the "histone code hypothesis" which states that the combinatorial nature of these modifications results in the generation of altered chromatin structures that mediate specific biological responses (Turner, 2000).

1.2 DNA-Protein Interactions in the Nucleosome Core Particle

The earliest concepts for the association of DNA and histones in the core particle came from image reconstruction analysis using electron micrographs (Klug *et al.*, 1980). At 20 Å resolution, a left handed helical ramp was apparent on the octamer surface and proposals were made for how the DNA-protein interactions might occur. Since then, X-ray crystallography has helped to advance understanding of the DNA-protein interactions involved in the nucleosome core particle. Milestones included the solving of the nucleosome structure at 7 Å resolution (Uberbacher & Bunick, 1985), which reconfirmed the initially inferred arrangement of histones and DNA. This led to the highest resolution structures of the nucleosome core particle to date at 2.8 Å (Luger *et al.*, 1997) and 1.9 Å (Davey *et al.*, 2002).

The high-resolution structure of the core-particle firstly revealed that the core particle had a pseudo-dyad[1] axis of symmetry: 1 bp sat on the dyad axis of the octamer. It further revealed in fine detail that the histone-DNA interactions were confined towards the phosphodiester backbone of the DNA strand (Luger *et al.*, 1997). Arginine/lysine-rich tails, protruding from the core histones, made "hook-like" contacts every 10 bp where the minor groove of the double-helix faced inwards. The histone-DNA contacts were non-base-specific and included predominantly salt-bridges and H-bonds as well as non-polar contacts with DNA sugars.

The 10 periodic contact feature of the DNA backbone was suggested much earlier. It was suggested, for example, when 10 bp-phased digestion patterns were observed upon using the enzyme DNase I[2] to cut nucleosome-bound DNA (Wang,

[1] The central axis of the histone octamer is herein referred to as the dyad axis.

[2] DNase I is an endonuclease, which breaks phosphodiester bonds within DNA.

1982). The observed cutting periodicity of 10 bp, which is “in phase” with the helical periodicity of DNA, forms the basis of many computational approaches aimed at finding nucleosome rotational positioning signals (Section 1.9).

The helical periodicity of DNA is not constant as it traverses around the histone octamer. For example, experiments using hydroxyl-radical cleavage of nucleosome-bound DNA showed that the helical periodicity was 10.0 bp/turn in the vicinity of the dyad axis and 10.7 bp/turn towards the ends of the nucleosome (Puhl & Behe, 1993). Most experimental evidence for B-DNA in solution suggests that it has a helical periodicity of 10.5–10.6 bp in solution (Wolffe, 1998). This variation in DNA periodicity along the core particle is thought to be a consequence of local histone-DNA interactions.

1.3 The Concept of Nucleosome Positioning

Nucleosome positioning has been proposed to be a potential mechanism for regulating gene expression, providing the view that nucleosomes could play important roles in addition to organizing higher order chromatin structures in eukaryotic cells. The term 'positioning' refers to a pre-determined organization of nucleosomes on a DNA sequence. In contrast, in a random arrangement of nucleosomes, all DNA sequences will have an equal probability of binding histones (Sinden, 1994). This gives rise to the idea that the local DNA structure, which is affected by the underlying DNA sequence, may play a role in positioning nucleosomes.

Two kinds of DNA structural patterns may thus be envisioned to direct nucleosome positioning: those that strongly favour nucleosome formation and those that strongly obstruct it. Nucleosome positioning can help to either selectively expose functionally important DNA sequences by constraining their locations to the linker region or impede accessibility to functionally important sequences by constraining their location to within the core particle. This can impose another level of regulation in gene expression, for instance, by controlling the accessibility of binding sites available to RNA polymerases or specific transcription factors. Two kinds of DNA structure-based nucleosome positioning have been described previously and these will be discussed next (Sections 1.4, 1.5).

1.4 An Introduction to Nucleosome Rotational Positioning

Rotational positioning determines which side of a DNA double helix surface will face and contact the histone octamer; this kind of positioning has been attributed to intrinsically curved DNA. The theory that a nucleosome will fit an intrinsically curved DNA is that the DNA is already in a preferred physical conformation to allow it to easily wrap around the octamer surface.

This section will firstly introduce the physical basis of DNA which results in intrinsic curvature and then describe how this relates to rotational positioning preferences for nucleosomes.

1.4.1 Intrinsic DNA curvature: Bending based on 10-phased [A] tracts

Intrinsically curved DNA is thought to be a consequence of permanent bends in a DNA sequence. This was first proposed when it was noticed that a 414 bp piece of kinetoplast DNA from *Crithidia fasciculata* displayed limited or retarded migration compared to other sequence fragments of equal length in acrylamide gel but migrated normally in agarose gel (Marini *et al.*, 1983). This anomalous migration was attributed to the size of the pores in the respective gels: in acrylamide gels, pore sizes vary between 1-8 nm whereas pore sizes in agarose gels vary between 40-400 nm. It was proposed that a permanent bend or curvature in the kinetoplast sequence was probably what caused the fragment to get stuck in the smaller size pores of the acrylamide gels.

The sequence motif that caused the permanent bends was mapped using the circular permutation assay (Wu & Crothers, 1984). In this procedure, various 241 bp-long restriction fragments, of the 414 bp-long kinetoplast DNA, were prepared and cloned as dimers. The length of 241 bp was chosen as this is greater than the persistence length of DNA[3]. The dimerized fragments were then run on an acrylamide gel and scanned for the fragment causing the shortest end-to-end migration distance (this region contained the permanent bend). This experiment concluded the retarded migration property to be an effect of 10 bp-phased runs of $CA_{4\text{-}5}T$ in the kinetoplast DNA. This work led Wu *et al* to propose the *junction model* for DNA bending; this predicts that the poly(dA)·poly(dT) tracts, within the 10 bp-phased $CA_{4\text{-}5}T$ motifs, adopt a non-B-DNA helix called heteronomous DNA (Arnott *et al.*, 1983). It proposes that permanent bends are located at the junction between this kind of DNA and regular B-DNA.

An alternative model to explain how phased-A tracts caused permanent bending was proposed later called the *wedge model* (Ulanovsky *et al.*, 1986). In this assessment, "bend angles" were calculated by measuring the efficiency of ligation of small DNA fragments into closed circles. This model predicts that the bends are not located at the junction between 2 kinds of DNA structure but within the [AA] dinucleotides themselves.

Parameters estimated from X-ray analysis of DNA structure have also been used to explain how phased-A tracts could cause intrinsic DNA curvature. From X-ray crystal structures, 2 variables are considered important for the relative motion of DNA base pairs: *roll* and *slide* (Calladine & Drew, 1992). *Roll* describes the opening of base pairs towards the major or minor groove of the double helix. A positive roll value indicates a tendency to open up towards the minor groove whereas a negative roll value indicates a tendency to open up towards the major groove in the opposite direction; typical values for DNA bases range between $+20^0$ to -10^0. *Slide* refers to

[3] The persistence length of DNA is 150 bp, the minimum length at which random DNA is essentially linear: it cannot circularize.

the translation along the axis of the base pairs. Slide values, which are restricted by the sugar-phosphate chain, range from +2 Å to -1 Å. Estimates of roll angle from X-ray structure analysis predict [AA/TT], [AT] and [GA/TC] dinucleotides to be stable at low roll (0^0) and low slide (0 Å) (El Hassan MA & Calladine, 1997) making their overall conformation very restricted. On the other hand, dinucleotides such as [GC/GC], [CG/CG] and [GG] dinucleotides are predicted to exhibit a wide range of roll angles (-10^0 to 20^0) making their conformation "bistable" or "context-dependent". For the phased A-tract bending, this suggests that the [AA] dinucleotides prefer to align their side of the minor groove towards the centre of curvature because of their restricted low roll configuration and the [GC] dinucleotides prefer to align the major groove away from the centre of curvature because of their bistable configuration (more on this below; Section 1.4.2).

The latest evidence that tries to explain how phased-A tracts result in bending comes from NMR studies (MacDonald *et al.*, 2001). This estimates a total of 19^0 bending in phased A-tracts. Of this, 4^0 occurs at the 5'end of the A-tract, 5^0 occurs within the A-tract itself and 10^0 occurs at the 3' end of the A-tract.

1.4.2 Intrinsic DNA curvature and the initial assessment of nucleosome rotational positioning

A rotational preference for a circular piece of DNA sequence has been described as a bias towards aligning a specific face of the DNA surface towards the direction of curvature and aligning a specific face away from the direction of the curvature (Drew & Travers, 1985). To study the rotational preferences of 10 bp-phased [A] tract sequences, a 169 bp sequence, containing phased [A]-tracts, was circularly ligated and digested with DNase I[4] (Drew & Travers, 1985). The [GC]-tracts were seen to be easily digested by DNase I and therefore more likely to face away from the circle. On the other hand, the phased [A]-tracts were more likely to be oriented towards the circle and thus protected from DNase I digestion. This observation was consistent with the X-ray crystal structure explanation of [A]-tract DNA bending discussed in the previous section (Section 1.4.1). As part of the same experiment, the same sequence was reconstituted onto a histone octamer *in vitro.* Digesting this reconstituted nucleosome with DNase I showed the same rotational preferences as for the circularized DNA: the phased [A]-tracts of the sequence were seen to face in towards the histone octamer. A later study addressed the optimal number of [A] nucleotides required for [A]-tract bending (Koo *et al.*, 1986). The approach used gel anomaly analysis of several lengths of [A] nucleotides in 10 bp-phased [A]-tract sequences. This study showed that 3–5 [A] nucleotides, phased at 10 bp, resulted in optimal curvature.

Further analysis of rotational positioning of DNA sequences on histone octamers was carried out by cloning a library of 177 nucleosome core particle sequences from chicken genomic DNA and subsequently analysing its dinucleotide periodicity (this dataset is discussed again subsequently in Section 1.8.1) (Satchwell *et al.*, 1986). The sequence lengths in the final dataset, however, were not constant, most probably due to biases in micrococcal nuclease (MNase[5]) cutting specificity (Section 1.8.1). The lengths ranged from 142 to 149 bp with an average length of 145 (±1.5) bp. To deal with this uncertainty, the analysis was carried out using 3 bp-averaged representations of the data. Also, the authors had to shift all sequences,

[4] DNase I interacts with the surface of the minor groove and bends the DNA molecule away from the enzyme.

[5] Micrococcal nuclease is both an endonuclease and an exonuclease, which can break the phosphodiester bonds in linker DNA and remove nucleotides from the ends of the DNA molecule respectively.

which were not of length 145 bp, a few base pairs until a central reference point of 73.25 was obtained. Fourier analysis of the dinucleotides in the dataset showed 10 periodic patterns of [AA/TT] and [GC]. These 2 motifs were furthermore seen to occur phased at 5 bp from each other, reminiscent of the A-tract bent sequences discussed in the previous section.

In the same study (Satchwell *et al.*, 1986), the 3 bp-averaged positions of dinucleotide motifs were compared with the co-ordinates of the DNA sequence which faced the octamer in the nucleosome X-ray crystal structure available at that time (Richmond *et al.*, 1984). This showed a pattern for phased A-tracts to face the octamer a few turns symmetrically away from the dyad axis of the nucleosome core particle but not at the dyad itself. In the X-ray crystal structure of the nucleosome, the minor groove also faced away from the dyad axis (Section 1.2). This result also agrees with the previous discussion that there are 2 kinds of DNA helical periodicities at the dyad and end positions respectively (Section 1.2).

1.4.3 Further evidence to support nucleosome rotational positioning

Since the initial assessment of nucleosome rotational positioning, a big trend was to chemically synthesise DNA sequences with optimised rotational preferences for forming reconstituted nucleosomes *in vitro*. For example, sequences having repeats of the motif [TATAAACGCC] were shown to ligate more efficiently into a circle compared to random DNA (Widlund *et al.*, 1999). This sequence was shown to bind nucleosome core particles *in vitro* ~350 fold higher than random DNA. A few naturally phased A-tract sequences are also known to favour nucleosome reconstitution *in vitro*, for example the 5S RNA gene of *Xenopus laevis* (Tomaszewski & Jerzmanowski, 1997).

Analysis of whole genomic sequences has also shown that they may contain enriched phased A-tract bending motifs for positioning nucleosomes. For example, Fourier analysis of *Caenorhabditis elegans* and *Saccharomyces cerevisiae* showed enrichment of [AA] motifs at 10.2 bp (Widom, 1996); the same pattern was not seen in a prokaryotic genome. A different approach to analyzing whole genomic sequences is the SELEX protocol (Widlund *et al.*, 1997). This procedure works by starting off with a random pool of genomic sequences and performing a number of rounds of PCR, each time amplifying sequences based on their affinity to bind histones. This approach found [A]-tract bending sequences in *Methanothermus fervidus*, which belongs to a branch of the archaeal kingdom that contains histone like proteins (*Euryarchaeota*) (Bailey *et al.*, 2000). The same patterns were not found in *Crenarchaeota*, a branch of the archaeal kingdom which does not contain histones. This led to the suggestion that the evolution of eukaryotic genome sequences most likely originated in the archaea, before the split of the eukaryotic lineage.

1.4.4 Nucleosome rotational positioning and DNA regulatory regions

Generally, chromatin structure provides a repressive environment for transcription. The evidence for this comes from observations of increased transcription levels of prokaryotic RNA polymerases in histone-depleted eukaryotic cells compared to their levels in normal eukaryotic cells (Gonzalez & Palacian, 1989). Prokaryotic RNA polymerases have traditionally been used in such analyses since they do not require specific transcription factors as do eukaryotic RNA polymerases (Wolffe, 1998). One of the ways eukaryotic cells are understood to overcome nucleosome barriers to permit transcription is through the activity of ATPase-based remodelling complexes (Wolffe & Guschin, 2000). An example is the SWI/SNF complex, which is thought to disrupt the rotational positioning of nucleosomes as suggested from the loss of 10 bp-phased DNase I cleavage patterns (Lorch *et al.*, 1998).

The indication for nucleosome rotational positioning provided an incentive to map naturally bent DNA near important genomic sequences and assess whether these bends could position nucleosomes (Bash *et al.*, 2001; Nair, 1998; Pruss *et al.*, 1994; Wada-Kiyama & Kiyama, 1996; Wada-Kiyama *et al.*, 1999).

For example, the circular permutation assay (Section 1.4.1) was used to map bend sites in the 3,000 bp promoter region of the human oestrogen receptor gene (Wada-Kiyama *et al.*, 1999). A total of 5 bend sites were found using the circular permutation assay; [A]-tract bending was observed for 3 of these sites. Nucleosome positioning at one of these bend sites was then analysed in detail. These were mapped by firstly digesting the clone with MNase to isolate core particles followed by digestion with 2 different restriction enzymes, whose restriction sites were known on the clone. This showed that the position of the bend appeared 10–30 bp away from the experimentally-predicted location of the nucleosome dyad axis. Therefore, it seemed likely that the specific bent site could help to direct nucleosome positioning. Nucleosome mapping to an intrinsically bent site was shown previously as well in the human β globin locus (Wada-Kiyama & Kiyama, 1996).

A few specific cases are known where positioned nucleosomes are important for protein signal recognition. An example of this is the hormone responsive element (HRE) of the mouse mammary tumour virus (MMTV) promoter (Pina *et al.*, 1990). Footprinting[6] analysis showed that the sequence of HRE was able to precisely position nucleosomes both *in vivo* and in reconstituted chromatin. It was then shown that nuclear factor 1 (*NFI*), one of the transcription factors for this promoter, was not

[6] This technique identifies the site of protein-binding on DNA by determining which phosphodiester bonds are protected from cleavage by DNase I

able to bind to the promoter when it was wrapped in a nucleosome. Hormone receptor binding to the MMTV nucleosome was seen to shift the rotational position of the nucleosome rather than causing it to dissociate completely; this was detected as greater accessibility of the promoter-proximal end to exonuclease III digestion. Thus, hormone receptor binding could act as a primary switch by shifting the rotational setting of the nucleosome to permit *NF1* binding. Another example is the binding site of the human immunodeficiency virus (HIV)-encoded integrase enzyme: DNA distortion studies have shown that this enzyme recognises specific bends within a nucleosome core particle (Pruss *et al.*, 1994).

1.5 An Introduction to Nucleosome Translational Positioning

Translational positioning determines where a histone octamer will be positioned along a long stretch of DNA; “long”, in this case, refers to a length longer than the core particle length (~146 bp). The theory behind this kind of positioning is that certain regions of a long DNA sequence may be much worse or much better than random DNA in their ability to wrap a histone octamer. Two kinds of DNA structural features may be important in determining the translational position of a nucleosome:

- Highly rigid DNA – DNA, whose structural conformation is very restricted, compared to random DNA, will be more difficult to bend around a histone octamer. Therefore, such kind of DNA can be expected to repel nucleosome formation.
- Highly flexible DNA - The conformation of highly flexible DNA is such that it offers least resistance to being bent and wrapped around a histone octamer. Thus, DNA, which is significantly more flexible than random DNA sequences, may position nucleosomes more readily. Flexible DNA is different to bent DNA previously described (Section 1.4.1) in that it offers low resistance to being wrapped around a histone octamer whereas bent DNA is a permanent feature of the DNA molecule.

1.5.1 DNA sequences that repel nucleosome formation

Sequences that resist nucleosome formation may do so because they tend to form some other kind of DNA secondary structure unfavourable for wrapping around a nucleosome. They might also contain signals to bind a different cellular protein, which would compete with the histone octamer for the same position. Initial nucleosome reconstitution experiments, using salt dialysis, had reported a lack of success in reconstituting nucleosomes using poly(dA)·poly(dT) / poly(dG).poly(dC) sequences (Rhodes, 1979; Simpson & Shindo, 1979). Although it was not clear why such sequences would disfavour nucleosome formation, Rhodes *et al* suggested that the high salt conditions used in the reconstitution procedure could have caused the poly(dA)·poly(dT) sequences to form heteronomous DNA, a triple-strand DNA structure (Arnott *et al.*, 1983). Poly(dG).poly(dC) sequences were also known to easily adopt A-DNA conformation (Arnott & Selsing, 1974) so this could have been a possibility for their inability to reconstitute into nucleosomes using the high-salt experimental conditions.

In another nucleosome reconstitution experiment, it was also observed that tracts of poly(dA)·poly(dT) and poly(dG).poly(dC) were not present towards the dyad axis (Drew & Travers, 1985). However, poly(dA)·poly(dT) tracts appeared towards the ends of the core DNA sequences suggesting that they may have an influence on the translational setting of the histone octamer (Satchwell *et al.*, 1986). The basis for translational positioning was not clear at this point; a recent study, however, examined the translational and rotational positioning properties of a simple 20 bp-repeating sequence (Negri *et al.*, 2001). The approach was to study the effects of subtle changes to the original sequence by mapping the changes to rotational and translational positions using hydroxyl-radical and exonuclease mapping respectively. The main conclusion was that the sequence distortions which affected the rotational preferences of the core particle were not the same ones which affected the translational position. The exact features which resulted in translational positioning, however, were not confirmed but it was suggested that the exact sequence contexts of [GA] and [CT] dinucleotides could be important.

Why long runs of poly(dA)·poly(dT) might repel nucleosome formation is still unclear. However, one explanation, using X-ray crystal analysis, predicts A·T base pairs to have high propeller twist[7] (Nelson *et al.*, 1987). This would result in maximal base-stacking (the interaction of adjacent base pairs) in poly(dA)·poly(dT) sequences resulting in an overall rigid stretch of DNA. [AA/TT] dinucleotides were also discussed earlier to show restricted conformation in X-ray crystallography studies (Section 1.4.1). This may make it difficult to bend poly(dA)•poly(dT) sequences to easily fit around a histone octamer.

Expansion of [CCG] repeats, which are known to cause fragile X syndrome, have also been studied in relation to nucleosome positioning (Wang *et al.*, 1996). Using competitive nucleosome reconstitution and electron microscopy, it was shown that >50 repeats of [CCG] blocks tended to exclude nucleosome formation. Such sites, visible in patients suffering from fragile X syndrome, were referred to as "fragile" loci as they stained poorly and were hotspots for DNA strand breakage. It was possible that [CCG] repeats formed some other kind of secondary structure: the lack of nucleosomes could account for the high frequency of DNA strand breaks. The exact mechanism for extensive CCG repeats in excluding nucleosome formation is still unclear.

Cao *et al* had performed a negative-SELEX experiment on mouse genomic DNA to yield an enriched quantity of sequences that repel nucleosome formation (Cao *et al.*, 1998). 35% of the sequences finally isolated had long repeats of [TGGA] and the affinity of these were half that of background DNA.

[7] Propeller twist is a property of a single base pair which describes the angle between the plane of the paired bases.

1.5.2 DNA sequences that favour nucleosome formation

Expanded blocks of [CTG] have been shown to be strong positioning signals for binding nucleosomes (Wang & Griffith, 1995). This motif had been previously shown to form expanded blocks downstream of the myotonic dystrophy gene in affected patients (Mahadevan *et al.*, 1992). Such regions were seen to bind a large number of nucleosomes using electron microscopy. An *in vitro* nucleosome reconstitution experiment showed that 2 DNA sequences, having 75 and 130 [CTG] repeats respectively, formed nucleosomes 6 and 9 times more strongly compared to the 5S RNA gene (a naturally occurring nucleosome-positioning sequence containing 10 bp-phased [A]-tracts) (Wang & Griffith, 1995). A study involving DNase I digestion of trinucleotides has also shown [CTG] trinucleotides to have one of the highest cutting rates and therefore to be amongst the most flexible trinucleotides (Brukner *et al.*, 1995). So according to the DNase I digestion results, the high flexibility of [CTG]-expanded regions may lead to a relatively “easy” fit for binding nucleosomes. However, according to the analysis of the chicken nucleosome data, [CTG] motifs did not show any kind of rotational positioning preferences, i.e. to face inwards or outwards in the structure of the core particle (Satchwell *et al.*, 1986). This suggests that [CTG] may show preferential nucleosome binding only when it is present in dense clumps: its overall density along a DNA sequence and not its rotational preference may influence its strong nucleosome-binding feature.

SELEX enrichment of core DNA in the mouse genome found some other possible nucleosome-positioning motifs, all of which could not be explained by phased [A]-tract motifs (Widlund *et al.*, 1997). This study found some cases of phased runs of 3-4 adenines ([A]-tract bending), multiple [CA] repeats, phased [TATA] tetranucleotides and one sequence having [CAG] repeats. However, fluorescence *in situ* hybridization showed these sequences to strongly localise to centromeric DNA; some of the sequence motifs were also known centromeric satellite repeats. Such repeats may not represent the majority of nucleosome-binding sequences in the genome as centromeric nucleosomes contain specialised nucleosomes that have variant histones (Smith, 2002). Furthermore, a recent study showed that the exact histone variant in addition to the DNA sequence may be a factor in positioning nucleosomes (Bailey *et al.*, 2002).

1.5.3 Nucleosome translational positioning and DNA regulatory regions

As mentioned earlier, nucleosomes are considered a repressive environment for transcription (Section 1.4.4). To overcome this, eukaryotic cells also contain ATPase-based remodelling complexes which are understood to shift the translational positioning of nucleosomes, for example NURF complexes in Drosophila (Hamiche *et al.*, 1999; Kang *et al.*, 2002). These are thought to induce sliding of nucleosomes as they do not disrupt the 10 bp-phased DNase I digestion patterns.

Understanding of the role of translational nucleosome positioning in repressing transcription has come from the use of *in vitro* transcription systems (Wolffe, 1998). Such studies ask if transcription can still occur *in vitro* following nucleosome reconstitution. The general outcome is that if histone assembly takes place first, transcription activity is inhibited. Of course, this system is unlikely to represent what happens in eukaryotic cells *in vivo* as it is difficult to mimic the multitude of transcription factors, which are actively involved in the process. An experiment, using an *in vitro* transcription system, showed that Alu repeats positioned histones over and next to promoter elements, which are critical for its transcription activity (Englander *et al.*, 1993). The poly [A] linker region of Alu sequences was proposed to exclude translational positioning by a histone octamer.

1.6 Regions of Phased Nucleosomes

One of the consequences of nucleosome positioning may be genomic segments having 'phased nucleosomes': in this case, a constant length of linker DNA is maintained throughout a specific segment of genomic sequence. Possible models for demarcating such segments have been proposed (Kiyama & Trifonov, 2002):

- A perfect positioning model – The positions for all nucleosomes are defined in a genomic segment.
- A partial positioning model – Certain positions in a genomic segment are designated for nucleosome formation. The alignment of other nucleosomes is influenced by the initial allocation of these key positions.

A crude method of detecting nucleosome phasing in a genomic clone is by digesting it with micrococcal nuclease and observing the digested products using gel electrophoresis. If the bands produced by electrophoresis produce a unique band, it suggests that the linker lengths are roughly equal and that a specific phase is maintained. Conversely, "out of phase" nucleosomes yield a number of bands of varying lengths. Nucleosome-phasing was observed in a few randomly selected chicken genomic DNA clones using this method (Liu & Stein, 1997). This study concluded that phased regions (<2k bp) alternated with randomly-positioned regions in the sampled clones; the phased regions were reported to show 210 bp-phased nucleosomes. Possible underlying sequence factors were proposed in one of the phased regions, which contained a gene. These included a run of 10 [A] residues in the linker DNA between 2 specific nucleosomes (possible translational positioning motif) and apparently 10 bp-phased [VWG] motifs (Section 1.9.3; a motif that could affect rotational positioning).

1.7 Strength of Nucleosome Positioning Sequences In Vivo

Two very important problems have been looked at previously concerning the strength of nucleosome positioning sequences *in vivo*. The first was to estimate what proportion of genome sequences might be constrained for packaging nucleosomes. The second problem was to answer how efficient these sequences were at binding octamers compared to artificial sequences.

The first question was answered using competitive nucleosome reconstitution in which a library of random natural genomic mouse DNA sequences and a library of chemically synthetic DNA (Lowary & Widom, 1997) were made to compete for binding limiting amounts of histone octamer. The conclusion was that only 5% of the total genomic library was enriched to bind histones with a free energy of reconstitution higher than the synthetic library.

To address the second problem about the strength of naturally occurring motifs, a set of the strongest possible motifs in the whole mouse genome was enriched and analysed using SELEX enrichment (Widlund *et al.*, 1997). The free energies of these sequences were compared with artificial sequences, which were similarly enriched for nucleosome-binding using SELEX enrichment (Thastrom *et al.*, 1999). The first and second strongest sequences in the entire mouse genome were seen to have 6 fold and 34 fold lower affinities respectively for binding octamers than the random pool of synthetic DNA. It was concluded that even the strongest binding natural sequences were not evolved to be the most energetically favourable possible.

1.8 Experimentally Mapped Nucleosome Datasets

Two databases of experimentally-mapped nucleosome sequences were available during the course of work described in this thesis. Sequences in both databases, however, suffer from experimental limitations which hinder the precise mapping of the dyad axis.

1.8.1 Database of chicken core DNA sequences

The database of chicken core DNA, which was introduced earlier (Section 1.4.2) (Satchwell *et al.*, 1986) (177 sequences), was kindly made available by Andrew Travers. To isolate core DNA, MNase digestion was performed on DNA extracted from chicken red blood cells. This was followed by a further deproteination step to remove H5 (the chicken equivalent of the linker histone H1 in human). This resulted in 239 sequences, which were cloned using an M13 vector, and sequenced. However, many of the cloned sequences were finally discarded: these included those that were less than 142 bp and those that contained a double-length insert of roughly 290 bp. The sequence lengths in the final database ranged from 142 to 149 bp with an average length of 145 (±1.5) bp.

The length differences could be partly attributed to the cutting specificities of MNase. It prefers cutting pA and pT faster than pC or pG (Bellard *et al.*, 1989) resulting in an accuracy of ±3 bp in determining the translational positioning of the core particle (Hager & Fragoso, 1999). However, the authors reported that the A+T content in the core particles were the same as those in bulk chicken DNA (Satchwell *et al.*, 1986). Only a drop of 13% in TpA between core particle DNA and bulk chicken DNA was noticed that could be biased by MNase cutting specificity.

The authors also mention that this dataset did not necessarily represent the bulk of nucleosome positioning *in vivo* as one step of the isolation protocol, which involved removal of H1, "*allowed the exchange of histone octamers between DNA molecules*" (Satchwell *et al.*, 1986).

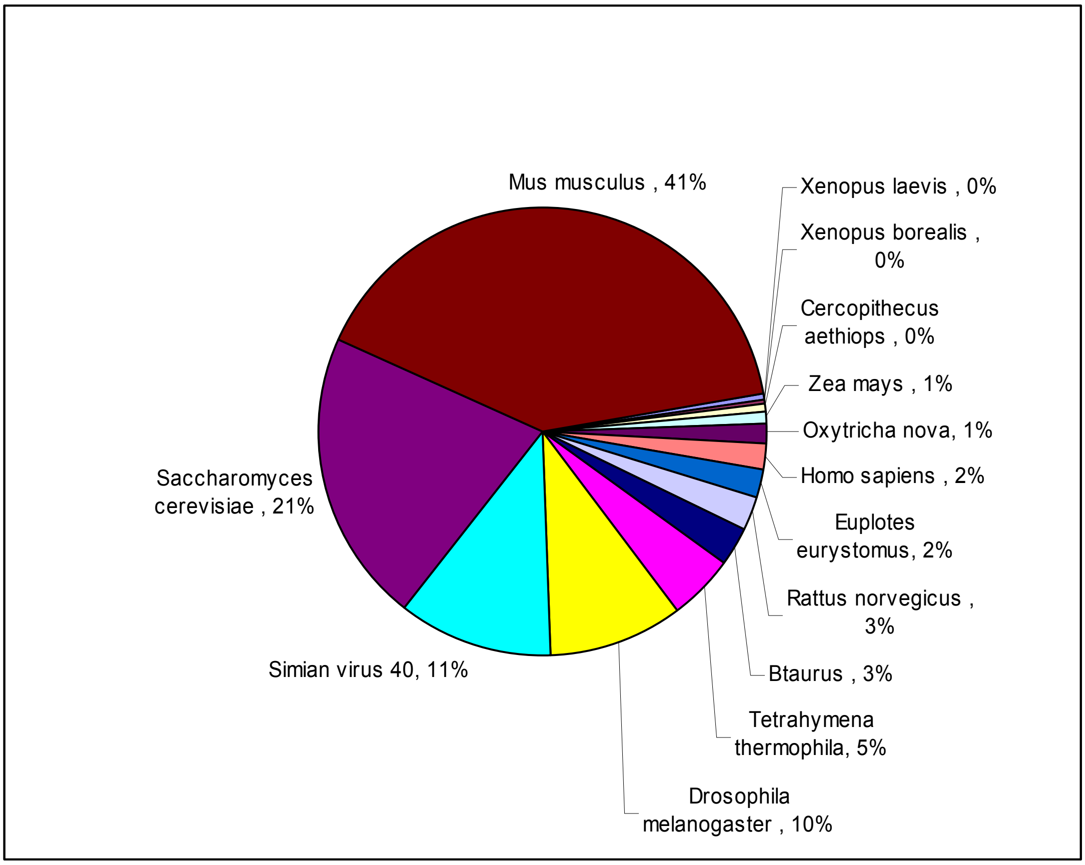
10 bp-phased [AA/TT] periodicity, along with 5 bp phase-shifted [GC], had been reported for this dataset (Section 1.4.2). Simple counting of [AA/TT] dinucleotide spacing (Figure 1.5, page 1-31) and multiple alignments of these sequences (Appendix A) were not sufficient to reproduce this result. The multiple sequence alignment in Appendix A, which is also sorted by pair wise identity, showed that the sequences were not highly similar to each other. A separate BLAST analysis (Altschul *et al.*, 1990)was also performed where each of the core DNA sequences was used to search for homologous members in the dataset (an "all against all" test; data not shown). This showed that these sequences were not highly similar to each other. This suggested that the reported periodicity was probably quite weak.

For some of the experiments performed in this thesis (Chapter 3 and Chapter 5), additional chicken genomic sequences were required which could be used as a background test set to these chicken core DNA sequences. Two chicken genomic clones were available for this purpose: AC092403 (144,369 bp) and AC120196 (202,027 bp).

1.8.2 Nucleosome database from mapping studies on various species

A second database of nucleosome sequences, which was publicly available (Levitsky *et al.*, 1999), essentially represented the same sequences from an earlier collection (Ioshikhes & Trifonov, 1993) and a more recent database of mouse nucleosomal sequences obtained using SELEX enrichment (Widlund *et al.*, 1997). A total of 193 sequences was present with the majority of sequences representing mouse and yeast data (Figure 1.3).

Figure 1.3: Organism sources of Levitsky et al's nucleosome sequence dataset (Levitsky *et al.*, 1999).



However, the length distribution of sequences was much more varied in this dataset compared to the mapped chicken sequences (Figure 1.4). The observed length variation necessarily resulted from the uncertainty of the technique used for nucleosome mapping. There were six main methods used, whose mapping accuracies are shown in Table 1.1 (Ioshikhes & Trifonov, 1993). The only technique unlisted in Table 1.1 is the SELEX protocol used to isolate many of the mouse nucleosome sequences: the lengths of these sequences ranged from 109 to 151 bp (average: 129 bp, standard deviation: 9 bp).

Figure 1.4: Length distribution of sequences in Levitsky *et al*'s nucleosome database.

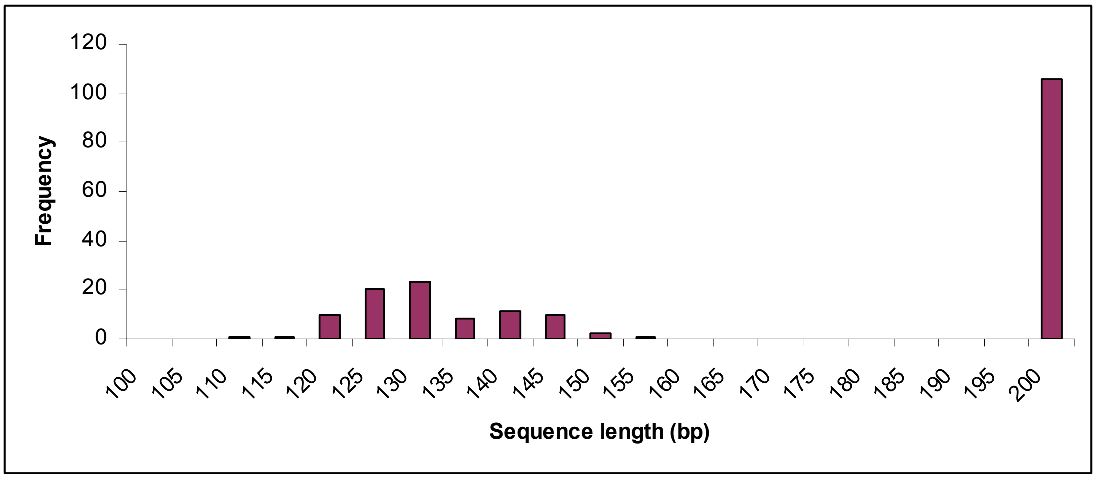


Table 1.1: Accuracy of different nucleosome mapping methods (Ioshikhes & Trifonov, 1993).

METHOD	MAPPING ACCURACY (bp)
MNase digestion of chromatin	>19
DNase I digestion of chromatin or reconstituted nucleosomes	10
Hydroxyl radical mapping	5
MNase digestion in combination with the cloning and sequencing of nucleosomal DNA sequences	5
DNase I digestion in combination with the highest possible accuracy	1
Exonuclease III with nuclease S1 digestion	1

The pair wise multiple sequence alignment of these sequences (Appendix A) showed that many of the mouse sequences were highly similar to each other (sequences 1-36 in the alignment). An "all against all" BLAST analysis also showed that these mouse sequences were highly similar to each other. However, they were more similar to the other sequences within the dataset compared to the chicken core DNA dataset (data not shown). The largely redundant mouse sequences were removed for any further analysis performed in this thesis. Unlike the chicken core DNA sequences, the sequence alignment of this dataset showed what appeared to represent phased [A]-tract motifs; these were in the first half of these sequences (Appendix A). [A]-tract bending was, therefore, more indicative in this dataset than in the chicken nucleosome dataset (this is discussed again subsequently; Section 1.9.2).

1.9 Computational Approaches to Understanding Nucleosome Positioning in Other Laboratories

This section will briefly introduce some of the computational approaches that have been developed till now to predict nucleosome formation.

1.9.1 Using DNA structural parameters to predict nucleosome positioning

The program BEND has often been used to predict DNA curvature and flexibility as a supplement to wet-lab mapping of positioned nucleosomes (Bash *et al.*, 2001; Blomquist *et al.*, 1999; Fiorini *et al.*, 2001; Wada-Kiyama *et al.*, 1999). The program accepts any DNA structural parameter set which can explain DNA bending along a DNA sequence, for example di-/tri- nucleotide parameter sets of twist, roll, tilt based on gel anomaly studies (Bolshoy *et al.*, 1991), cyclization kinetics (Ulanovsky *et al.*, 1986), X-ray crystallography (Calladine *et al.*, 1988) etc.. This software was useful to show that the binding of transcription factor *NF-1* depended on the position of curved DNA, which in turn affected nucleosome rotational positioning around the *NF-1* binding site (Blomquist *et al.*, 1999). The analysis was performed by introducing various sequence changes around the binding site and analyzing the potential effects of curvature. The software also helped to confirm bend sites, which were predicted using the circular permutation assay, in the promoter region of the GAL1-10 gene in yeast (Bash *et al.*, 2001).

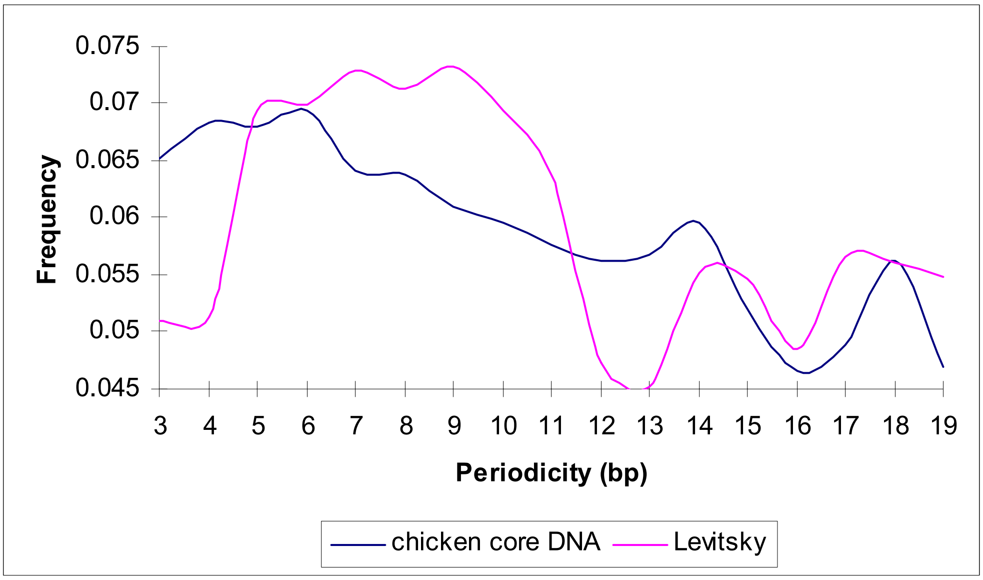
The wavelet tool (used in this thesis; Section 2.4.1, Chapter 4) is an example of a different approach which can use DNA structural parameters. It can be used to assess the occurrence and distribution of structural patterns that could affect nucleosome positioning (Arneodo *et al.*, 1995; Arneodo *et al.*, 1998; Audit *et al.*, 2001; Audit *et al.*, 2002). So far, it has been used to show that non-coding eukaryotic genomic DNA contain periodic flexibility patterns (>100 bp periodic) which do not appear in coding DNA or in prokaryotic DNA sequences. The size of such repeat periods, which reflects the size of a nucleosome, has been suggested to be potential nucleosome-positioning elements.

1.9.2 [AA/TT] rotational positioning pattern obtained using multiple sequence alignment

Ioshikhes *et al.* used five kinds of multiple alignment algorithms to create profiles of the nucleosomal database described earlier (Section 1.8.2) (Ioshikhes *et al.*, 1996; Ioshikhes & Trifonov, 1993). The algorithms considered only the positions of [AA/TT] dinucleotides because of their importance in rotational positioning described earlier (Section 1.4.1). These algorithms modelled an [AA/TT] dinucleotide positional frequency with a periodicity of 10.3(±0.2) bases towards the ends of a 146 bp sequence. [TT] dinucleotides also appeared to be distributed symmetrically relative to [AA] dinucleotides on the same DNA strand (phase difference: 6 bp). This result was reminiscent of the Fourier analysis results of the chicken core DNA dataset (Section 1.4.2) (Satchwell *et al.*, 1986) except the latter found [GC], rather than [TT], to be in phase with [AA]. A similarity, however, was that the periodic feature was seen to appear symmetrically away from the central 15 bp indicating that the DNA in the location of the dyad axis was not bent.

According to the multiple sequence alignment of these sequences using the software *Clustal W* (Appendix A), phased A-tracts were evident towards the first half of the sequences. However, the algorithms used to align the sequences by Ioshikhes *et al* were more strategic in that they did not model any ‘deletes’ and were specifically handling [AA/TT]-periodicity (*Clustal W* uses the 4-letter DNA alphabet and will align any given sequences). Therefore, the alignment results from using *Clustal W* cannot be expected to give exactly the same results. Simple counting of [AA]-spacing showed a smeared peak between 5-11 bp for this dataset (Figure 1.5) indicating that phased-A tracts were featured in this dataset.

Figure 1.5: Simple counting of [AA]-spacing in the 2 experimentally-mapped nucleosome datasets (Section 1.8).



Denisov *et al.* used this model to predict nucleosome-centering around splice sites in 2000 exon-intron boundary sequences (400 bp fragments) obtained from a variety of eukaryotic species (Denisov *et al.*, 1997). The sequences appeared to position the midpoint of the nucleosome towards the introns. However, the data presented in the analysis were averaged values and it is not clear what proportion of the sequences showed this trend.

1.9.3 10-periodic [VWG] pattern obtained using hidden markov models

A 10-periodic [VWG] motif was found serendipitously using hidden markov models (HMMs) (Baldi *et al.*, 1996). Initially, conventional left-right hidden markov models, which were being trained to recognize splice-site junctions, learnt this signal. A different kind of HMM architecture, the cyclical HMM was constructed which detected this motif with an apparent 10 bp periodicity in coding sequence. Many of the sequence members of the motif [VWG] were seen to be highly flexible in a DNase I – based flexibility table (Brukner *et al.*, 1995). This kind of proposed bending was different to the A-tract bending described earlier (Section 1.4.1); this suggests that 10-phased "flexible" motifs ([VWG]), rather than 10-phased "rigid" motifs ([AA]), could help to achieve nucleosome rotational positioning. The result was described as a flexible motif which appeared every 10 bp and which was superimposed over coding DNA[8]. This study suggested that exons could possess a nucleosome-binding signal superimposed over protein-coding signal.

Stein *et al.* used this observation as a model to predict nucleosome-positioning on the SV40 minichromosome simply by counting occurrences of 10-periodic [VWG] motifs (Stein & Bina, 1999). The results showed a weak correlation (correlation co-efficient: 0.52 with a P value <0.001) with experimentally-mapped nucleosomes in a 3,300 bp region (out of 5,200 bp) in the late SV40 region. It was described that in regions in the SV40 early region, where [VWG] could not be used to predict strong nucleosome positions, the 10-periodic [AA/TT] signal (Section 1.9.2) could. 5,000 bp is perhaps too short a sequence length for analysing nucleosome-positioning though: the maximum number of nucleosomes that could possibly fit on the whole SV40 minichromosome would be <30. Also, the reported correlation was observed in a specific part of the sequence rather than throughout the entire sequence.

[8] Coding DNA has harmonics of 3 bp.

1.9.4 RECON: A nucleosome prediction model based on dinucleotide relative abundance distance

A function to find 'nucleosome formation potential' was described recently (Levitsky *et al.*, 2001a). The prediction software, called RECON, was based on a function which calculated the optimal distance in dinucleotide space between mouse genome sequences that position nucleosomes (positive set) (Widlund *et al.*, 1997) and mouse genome sequences that repel nucleosomes (negative set) (Cao *et al.*, 1998). 86 sequences were available in the positive set and 40 sequences in the negative set. Using a jack-knifing procedure for model-testing, a model was trained which showed 80% accuracy at 94% coverage. Prediction analysis using this algorithm showed that introns and Alu repeats had a higher nucleosome formation potential than exons (Levitsky *et al.*, 2001b).

However, using fluorescence *in situ* hybridization, the positive set used in this study were found to belong to the mouse centromeric class of repeats (Widlund *et al.*, 1997). Centromeric nucleosomes are known to bind octamers, which have a variant of histone H3 in a large number of eukaryotes; this includes mouse (Smith, 2002). Therefore, it is unlikely that this positive set represents the majority of sequences that would bind nucleosomes in 'non-centromeric' genomic DNA.

The mouse positive sequences, used in RECON, were part of Levitsky *et al*'s nucleosome dataset introduced earlier (Section 1.8.2). However, the pair wise multiple sequence alignment of these sequences showed that a large number of the mouse sequences were highly similar to each other (Appendix A). These close variants were not reported to be discarded in the RECON software training. These could bias the results learnt in the RECON model.

1.10 Summary of Aims

The idea of nucleosome positioning, particularly its potential role in transcription regulation in eukaryotic cells, was an interesting prospect to research. With the large amount of eukaryotic genomic sequences now available from recent sequencing projects, particularly human and mouse data, an appealing option was to scan for evidence of nucleosome positioning, build models to predict nucleosome positioning and compare the predictions with known annotated features on these sequences.

1.10.1 The scope for studying nucleosome positioning

However, the scope for building good quality nucleosome models was limited. The restrictions arose partly from the limited experimentally-mapped data that supported nucleosome positioning. The 2 experimentally mapped nucleosome datasets (Section 1.8) each contained less than 200 sequences and also the initial sequence alignments of the 2 datasets did not show any obvious similarity between the 2 (Appendix A). About 36 sequences in the Levitsky dataset were also redundant.

Also, with regard to their role in events such as transcription regulation, the general view is that nucleosomes repress such activities (Section 1.4.4, 1.5.3); this could probably be a consequence of nucleosomes lying in the path of regulatory proteins such as RNA polymerase and transcription factors. This does not require nucleosomes to be positioned and it is not yet clear to what proportion positioned nucleosomes could repress transcription *in vivo*. Specific examples are available, for example *NF1*-binding to the MMTV promoter (Pina *et al.*, 1990) (Section 1.4.4). In this case, the position of a nucleosome is thought to be regulated by binding of a regulatory receptor protein, which in turn affects the accessibility of a transcription factor to its target site. From this, it could firstly be expected that it would not be energetically favourable to have a large density of specifically positioned nucleosomes throughout the genome. Secondly, the few nucleosome positioning signals that are available could be expected to appear near gene regulatory regions where they could carry out important functional roles. Overall, this does make it difficult to detect nucleosome positioning sequences with high sensitivity especially from using whole genome analysis techniques.

The role of chromatin remodelling complexes (Section 1.4.4, 1.5.3) in directing nucleosome positions near promoter regions provides additional speculation that many nucleosomes could be positioned. In other words, it could be hypothesized that the remodelling complexes target positioned nucleosomes *in vivo*. At the moment, this remains speculation as the roles of chromatin remodelling complexes have not yet been assessed *in vivo* (Tsukiyama, 2002).

It is also important to note that the current experimental procedures used to reconstitute and map nucleosomes may not represent positioned nucleosomes *in vivo*. Chromatin extracts often contain much higher levels of the HMG (high mobility group) of chromatin proteins than the cellular background (Wolffe, 1998). These proteins are known to interact with nucleosomes. *In vivo*, chromatin structure is dynamic and using reconstitution procedures it is difficult to mimic the activity of important factors such as chromatin assembly factors, post-translational modification of histones and the nucleosome assembly process itself (which occurs in stages). Also, in the reconstitution procedure, it is quite difficult to assess the non-specific association of DNA with histones.

1.10.2 Aims and benefits of predicting nucleosome positioning

Given the limitations above, predicting nucleosome positioning was always going to be a challenging task. Most of the evidence for nucleosome positioning itself was based on the results of *in vitro* experiments including the hypothesis of intrinsically curved DNA (Sections 1.4.1, 1.4.2). Possibly the major indication that nucleosomes could be positioned *in vivo* came from Lowary *et al*'s work, using competitive reconstitution (Section 1.7) (Lowary & Widom, 1997). From the results, it was estimated that only 5% of the mouse genome was probably enriched for binding nucleosomes

The aim in this thesis was to build computational models to predict nucleosome positioning. The first objective was to scan for evidence which could suggest that nucleosome positioning signals exist in the first place in eukaryotic genomic sequences. A second goal was to scan for evidence that suggests that nucleosome positioning could be involved in gene regulation. This would be carried out using 3 major modelling approaches (Section 1.11). If the positioning predictions, using any of the modelling techniques, indicated the following properties, it could suggest importance of nucleosome positioning in gene regulation *in vivo*:

- A high density of predictions in the vicinity of annotated genes
- Conservation of the prediction patterns in different eukaryotic species

If, however, the predictions were made randomly throughout the genome, it would suggest more that nucleosome positioning, if it does occur, is important only for maintaining and stabilizing higher order chromatin structures.

Being able to predict nucleosome positioning would definitely be beneficial in certain areas of genomic research. It may, for instance, aid in gene prediction if it can be shown that certain genes or regulatory DNA sequences have positioned nucleosomes over them or in their vicinity. This may, in turn, lead to clues about their expression patterns. Another area where it may be helpful is in the diagnostics of chromatin diseases, many of which are postulated to be due to aberrant nucleosome positioning (Hendrich & Bickmore, 2001).

1.11 Approaches proposed for modelling nucleosome positioning

The methods outlined below have been employed in this thesis to approach the problem of predicting nucleosome positioning. Chapter 2 will give a brief summary of the theories of these methods.

1.11.1 Potential for studying 10 bp-phased motifs

Chapter 3 of this thesis deals with the use of cyclical HMMs. The aim of this approach was to scan for 10 bp-phasing motifs in genomic sequences, which could potentially influence nucleosome rotational positioning. This modelling approach extended the cyclical HMM work of Baldi and Brunak (Baldi *et al.*, 1996), which was introduced earlier (Section 1.9.3). The results obtained by Baldi and Brunak suggested that 10-phased [VWG] could be a nucleosome positioning signal. Many of the sequence members of this motif were highly flexible according to a DNase I–based flexibility table (Brukner *et al.*, 1995). Baldi and Brunak's overall technique, however, involved only learning the motif from various kinds of human genomic sequences including exons, introns and intergenic sequences: the models were not used to perform any predictions. The architecture of their cyclical HMMs was extended in this thesis to additionally model the background distribution of learnt 10-cyclical motifs. This would allow a HMM to be trained which could be used as a prediction tool. The two experimentally-mapped nucleosome datasets were also used as training sets for this purpose.

1.11.2 Potential for studying nucleosome translational positioning

In Chapter 4, the wavelet transform tool (Section 2.4.1) was used to probe the locations of periodic flexibility patterns in genomic sequences. The aim for the investigation was to establish whether any evidence existed suggesting that translational nucleosome positioning was an important mechanism for positioning nucleosomes in eukaryotic species. This would be achieved by modelling DNA sequences as flexibility sequences (Section 2.3.1). Recent work had already reported that eukaryotic DNA exhibit significant flexibility patterns which correspond to the repeat length of the nucleosome and which do not appear in prokaryotic genomes (Audit *et al.*, 2001; Audit *et al.*, 2002). It has also been reported that such patterns appeared only in non-coding DNA (Arneodo *et al.*, 1995; Buldyrev *et al.*, 1998; Havlin *et al.*, 1999; Pattini L, 2001). However, the genomic contexts of such patterns had not been clarified yet.

In Chapter 4, the wavelet transform tool was used to establish both the distribution of strong periodic flexibility patterns in representative genomes as well as determine if such patterns appeared near gene dense regions in DNA sequences. In addition to establishing the locations of these periodic features, it could also be determined if previously known DNA sequence features were the major players in determining potential nucleosome translational positioning.

1.11.3 Using DNA weight matrices to model the existing nucleosome datasets

The two available nucleosome datasets (Section 1.8) have both been analysed for rotational positioning and have been described to contain such positioning signals a few turns away from and symmetrically about the nucleosome dyad axis (Ioshikhes *et al.*, 1996; Satchwell *et al.*, 1986) (Sections 1.4.2, 1.9.2). The methods applied themselves, however, were specifically aimed to find rotational positioning signals, namely patterns which recur at 10 bp periodicity in these datasets. For the chicken dataset, this was obtained using 3 bp window-averaged counts of dinucleotides along their position in the sequences (Satchwell *et al.*, 1986); this found the motif [AA/TT] to be enriched at 10 bp periodicity along with a relative 5 bp phase-shifted [GC/GC] motif. For Levitsky *et al*'s data, it was assumed that [AA/TT] was the major rotational positioning motif and the periodicity of this motif was analysed using several multiple sequence alignment algorithms (Ioshikhes *et al.*, 1996). This yielded a similar result to the chicken data except that [TT], and not [GC/GC], was reported to be phased at 5 bp to [AA] on the same strand.

However, to be a significant pattern, the suggested rotational positioning motifs should be present in the majority of these sequences; this has not yet been clarified for either dataset. Thus a motivation was formed to apply a rigorous classification system to each of the nucleosome datasets. This was the focus for the work in Chapter 5.

2 General Introduction to Computational Methods Used in this Thesis

2.1 The Application of Bayesian Methods in Sequence Analysis

Bayesian analysis (Grate *et al.*, 1996), a general class of stochastic modelling techniques based on Bayes' theorem of conditional probability (Equation 2.1), represent an important approach for studying biological sequences. The idea is to construct a model that describes a set of sequences. The model can then be used to find a set of related sequences or examined further to determine properties of the sequences. A model in this case can be described as a "black box" which does not necessarily represent a "real world" mechanism. The model's value depends solely on the accuracy of its predictions and not by the mechanism used to make those predictions.

Equation 2.1: Bayes' theorem of conditional probability. In the context of biological sequence analysis, *M* represents a Bayesian model and *s* a DNA or protein sequence.

$$P(M \mid s) = \frac{P(s \mid M)P(M)}{P(s)}$$

Bayes' theorem (Equation 2.1) is based on the idea that in many situations, an analysis can be commenced with an estimated prior probability for an event of interest. This probability can come, for example, from historical data or previous experience. The idea is to receive additional information such that the prior probabilities in Equation 2.1 can be updated. The updated probabilities are referred to as the posterior probabilities.

In Equation 2.1, above, one of two conditional probabilities to update is *P(M|s)*. This probability value answers the question "Given the sequence *s*, what is the probability that it came from the distribution described by *M*?". The other conditional probability to update is *P(s|M)*, which is the probability of the sequence *s* given *M*. Two prior probabilities are required to estimate these values: *P(M)*, the probability that *s* is drawn from model *M* and *P(s)*, the probability of the sequence *s*. It is not possible to know the real probabilities of *P(M)* and *P(s)* but a different approach can be used to overcome this. The approach is to calculate the odds that the sequence *s* came from model *M* rather than a null model *N* (Equation 2.2). As can be seen from Equation 2.2, *P(s)* is no longer required. The model probabilities *P(M)* and *P(N)* can be estimated using iterative training methods (the procedure for hidden markov models is described in Section 2.2.3).

Equation 2.2: Relative probability of model M and the null model N.

$$\frac{P(M \mid s)}{P(N \mid s)} = \frac{P(s \mid M)P(M)}{P(s)} \times \frac{P(s)}{P(s \mid N)P(N)} = \frac{P(s \mid M)}{P(s \mid N)} \times \frac{P(M)}{P(N)}$$

The null model defines what the null hypothesis is. Choosing a good null model is a tricky problem and depends on the problem at hand. A sequence *s* can then be said to fit model *M* if $P(M|s) > P(N|s)$. Usually, this result is scored in log values and the value $log\ P_M(s)$ - $log\ P_N(s)$ is referred to as the log-likelihood of the sequence. In practice, a threshold score is chosen: the higher the log likelihood score is than the threshold, the greater the confidence in the result. Bayesian methods have been used in this thesis in Chapters 3 and 5.

2.2 Hidden Markov Model Theory

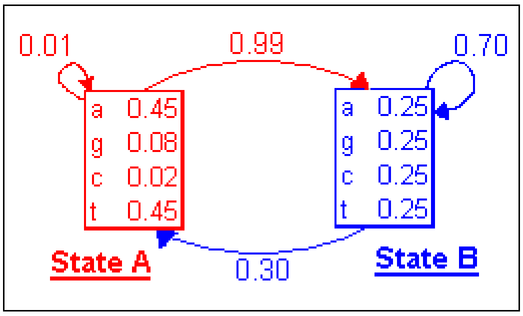
2.2.1 A general introduction to hidden markov models

Hidden Markov Model (HMM) analysis has widespread applications in Bioinformatics particularly in DNA and protein sequence analysis. These include creating multiple alignments of sequences to model protein families (Bateman *et al.*, 2002) and gene prediction (Meyer & Durbin, 2002). HMMs have also found importance as a pattern discovery tool; an example was seen recently where it was used to learn local composition patterns from chromosome 2 in the malarial genome *P. falciparum* and use that information to predict corresponding features in chromosome 3 (Pocock MR *et al.*, 2000). It has also been used as a discovery tool to find patterns that could be involved in nucleosome rotational positioning (Baldi *et al.*, 1996). This approach used a special kind of HMM referred to as the cyclical HMM. In this thesis, this approach has been extended to try to gain further insights into the patterns which were originally reported using cyclical HMMs: this is the focus of Chapter 3. This section will briefly introduce some basic HMM terminology and then introduce two algorithms which were used in this thesis for HMM prediction and training respectively (Sections 2.2.2, 2.2.3).

- **HMM terminology**

A hidden markov model (HMM) is in essence a vector of "states" connected with "transition paths"; each state contains 2 kinds of probability distributions associated with it: an emission spectrum and a transition spectrum respectively. Figure 2.1 shows a HMM which has an architecture of 2 states connected by a number of transitions.

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.



To model a specific kind of sequence with a HMM, it is first necessary to define the alphabet from which that sequence is composed; this alphabet is called the "emission alphabet". To model DNA sequences with a HMM, for example, it needs to be defined that DNA is composed of an emission alphabet of 4 symbols, "a,c,g,t".

The HMM shown in Figure 2.1 is a 2-state HMM, based on the DNA alphabet. *State A* has a strong probability of emitting "a" (0.45) or "t" (0.45) and a much weaker probability of emitting "g" (0.08) or "c" (0.02). *State A* has 2 transition paths out of it: one path to *State B* and one path back to itself. These paths form the transition spectrum of *State A*. In this case, it has a weak transition probability of going back to itself (0.01) and a strong transition probability of going to *State B* (0.99). *State B* has a random emission distribution (each symbol emitted at equal probability) and a set of 2 transitions (0.70 probability of going back to itself and 0.30 probability of going to *State A*). The entire set of emission and transition probabilities in the HMM define the HMM's parameters. This model can be used to score a sequence; this score is usually the product of all the emission and transition probabilities in the "path" of the model in that sequence (described below).

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.

```
(a) Possible High Scoring Sequence:

GAGCCGGCCGGGGGCCCGGGCCCGGGCTCGGGGACCCGCCCCCTCGCCCCAACCGCGG

(b) Possible Low Scoring Sequence:

AAAACCCTTAAAAATTTCGGGCCCTTTTTCCCTGTTTAAACGGTCCCTATTTACCCGG
```

To introduce HMM paths and HMM-based scoring, the 2 sequences in Figure 2.2 are considered. The first assumption is that the sequences in Figure 2.2 have been generated by the states of the HMM of Figure 2.1. But it is not known which part of the sequence was emitted by *State A* or *State B*; this is a "hidden" path from which the "hidden" term of HMMs is derived. However, it can be guessed that the sequence of Figure 2.2(a) was more likely to have been produced by a path through the HMM than the second sequence (Figure 2.2(b)). This is firstly because *State A*, whose emission spectrum represents [W] [9] motifs, has only a weak transition probability of going back to itself but a strong transition probability of going to *State B* (whose emission spectrum is random). Secondly, *State B* has a stronger probability of going back to itself compared to going back to *A*. This means that the HMM is more likely to spend more of its "energy" in *State B* than in *State A*. It effectively makes this HMM a model or predictor for sequences which display "short spurts" of [W] (*State A*) compared to a random background (*State B*). A path through the HMM which could have produced the sequence in Figure 2.2(a) could be as shown in Figure 2.3.

[9] Please refer to the ambiguity symbols for DNA at the beginning of the thesis

Figure 2.3: (a) A possible path through the HMM which could have emitted (b) the corresponding DNA sequence.

```
 (a) Possible path through the HMM:

BABBBBBBBBBBBBBBBBBBBBBBBBBBBBABBBBBABBBBBBBBBBABBBBBBBBBB

(b) DNA sequence:

GAGCCGGCCGGGGGCCCGGGCCCGGGCTCGGGGACCCGCCCCCTCGCCCCAACCCCCA
```

An algorithm for predicting the hidden path of states is described next.

2.2.2 Predicting the most likely path of a HMM through a sequence using the Viterbi algorithm

The Viterbi algorithm can be used to predict the most probable path, $\Pi_{(a)}$, through a HMM's states that could have emitted a given sequence. It uses a "dynamic programming" matrix where the columns are indexed by the states of the HMM, *S*, and the rows are indexed by the position x_i of the sequence *X*. The algorithm is outlined below using the following notations (Karchin, 1999; Shamir, 2001):

A general hidden markov model (HMM) is defined as $M=(A,S,Y)$ where:

- A = finite set of symbols (also called the emission alphabet).
- S = finite set of emission states.
- Y = finite set of probabilities comprised of:
  - State transition probabilities, denoted by t_{kl} for each $k,l \in S$.
  - Emission transition probabilities, denoted by $e_k(b)$ for each $k \in S$ and $b \in A$.

A sequence *X*, of length *L*, is defined whose positions are indexed as $(x_1,\ldots,x_i)$. $v_k(i)$ is denoted as the probability of the most probable path for the sequence that ends with state *k* ($k \in S$ and $1 \leq i \leq L$). $\Pi_{(a)}$ is found using the following steps:

- **Initialization:**

$v_{begin}(0) = 1$

For all $_{k \neq begin}$, $v_k(0) = 0$

- **Recursion:**

For each *i = 0, ... ,L-1* and for each *l* ∈ *S* the following is calculated recursively:

$$v_l(i+1) = e_l(x_{i+1}) \cdot \max_{k \in S} \{v_k(i) \cdot t_{kl}\}$$

During each recursive step, a backpointer is assigned from *l* back to the *k*.

- **Termination:**

$$P(X \mid \Pi_{(a)}) = \max_{k \in S} \{v_k(L) \cdot t_{k,end}\}$$

- **Path Reconstruction:**

$\Pi_{(a)}$ is found by re-tracing the backpointers.

2.2.3 Training a HMM using the Baum Welch algorithm

The HMM, shown in Figure 2.1, can be used to score any DNA sequence, for example by obtaining the Viterbi score, $P(X|\Pi_{(a)})$, as explained above. But the parameters of the HMM itself, *Y*, may not be realistic. To obtain realistic probabilities, it is necessary firstly to obtain a set of related sequences which contain a known motif or a set of known motifs. These sequences form the training set, $X_{(1)},...,X_{(n)}$, from which *Y* must be "learnt" or "trained". Training is an iterative process which keeps refining the parameters of the HMM to obtain an optimal score for $X_{(1)},...,X_{(n)}$ denoted as $Score(X_{(1)},...,X_{(n)}|Y)$. The Baum Welch algorithm is one such training algorithm, which was used in this thesis.

Before the Baum Welch algorithm can be introduced, it is important to point out that the individual statepaths of the HMM, $\Pi_{(1)},..,\Pi_{(n)}$, which produced $X_{(1)},...,X_{(n)}$

are unknown. The Baum Welch procedure has a step to overcome this. The step involves computing the probability of every statepath $\Pi_{(i,j)} = (\pi_{1(i,j)},..,\pi_{L(i,j)})$ for every $X_{(j)}$ in $X_{(1)},...,X_{(n)}$. These probabilities, $P(\pi_{(\iota,j)} = k|X_{(,j)})$, can be calculated using the forward and backward algorithms which are outlined first:

Forward algorithm (outlined for a single sequence *X*):

The parameter $f_k(i)$ denotes the probability of emitting X using the statepath $\pi_i = k$.

- **Initialization:**

$f_{\text{begin}}(0) = 1$

For all $_{k\neq\text{begin}}$, $f_k(0) = 0$

- **Recursion:**

$$f_l(i+1) = e_l(x_{i+1}) \cdot \sum_{k \in S} f_k(i) \cdot t_{kl}$$

- **Termination:**

$$P(X) = \sum_{k \in S} f_k(L) \cdot t_{k,end}$$

Backward algorithm:

The Backward algorithm works in exactly the same way as the forward algorithm except it is computed backwards from the end of *X*. The parameter $b_k(i)$ denotes the backward probability of emitting X using the statepath $\pi_i = k$.

Finally, it can be shown that $P(X, \pi_i = k) = f_k(i) \cdot b_k(i)$ (Shamir, 2001).

Baum Welch algorithm:

- **Initialization**

Y is initialized with reasonably-guessed parameters. For work done in this thesis, all $e_k(b)$ were initialized randomly and a reasonable guess was made for t_{kl}.

- **Expectation**

The probabilities $P(X_{(i,j)})$ for every statepath $\Pi_{(i,j)}$ for all $X_{(1)},\ldots,X_{(n)}$ is calculated as above.

The following 2 parameters can now be estimated:

  - T_{kl} – the number of transitions from state *k* to state *l.*
  - $E_k(b)$ – the number of times that an emission of the symbol *b* occurred in state *k*.

These are estimated as follows:

$$T_{kl} = \sum_{j=1}^{n} \frac{1}{P(X_{(j)})} \cdot \sum_{i=1}^{L_{(j)}} f_{k(j)}(i) \cdot t_{kl} \cdot e_l(x_{i+1(j)}) \cdot b_{l(j)}(i+1)$$

$$E_k(b) = \sum_{j=1}^{n} \frac{1}{P(X_{(j)})} \cdot \sum_{\{i|x_{i(j)}=b\}} f_{k(j)}(i) \cdot b_{k(j)}(i)$$

- **Maximization**

The new values of *Y* are estimated from T_{kl} and $E_k(b)$. These are estimated using maximum likelihood estimators for the transition and emission probabilities respectively. The maximum likelihood estimators are:

$$a_{kl} = \frac{T_{kl}}{A_{q\in S} A_{kq}}$$

$$e_k(b) = \frac{E_k(b)}{A_{a\in A} E_k(a)}$$

- **Terminaton**

Steps 2 and 3 are repeated until the improvement in $Score(X_{(1)},\ldots,X_{(n)}|Y)$ is less than a given parameter ε.

2.3 The Use of Flexibility Sequences

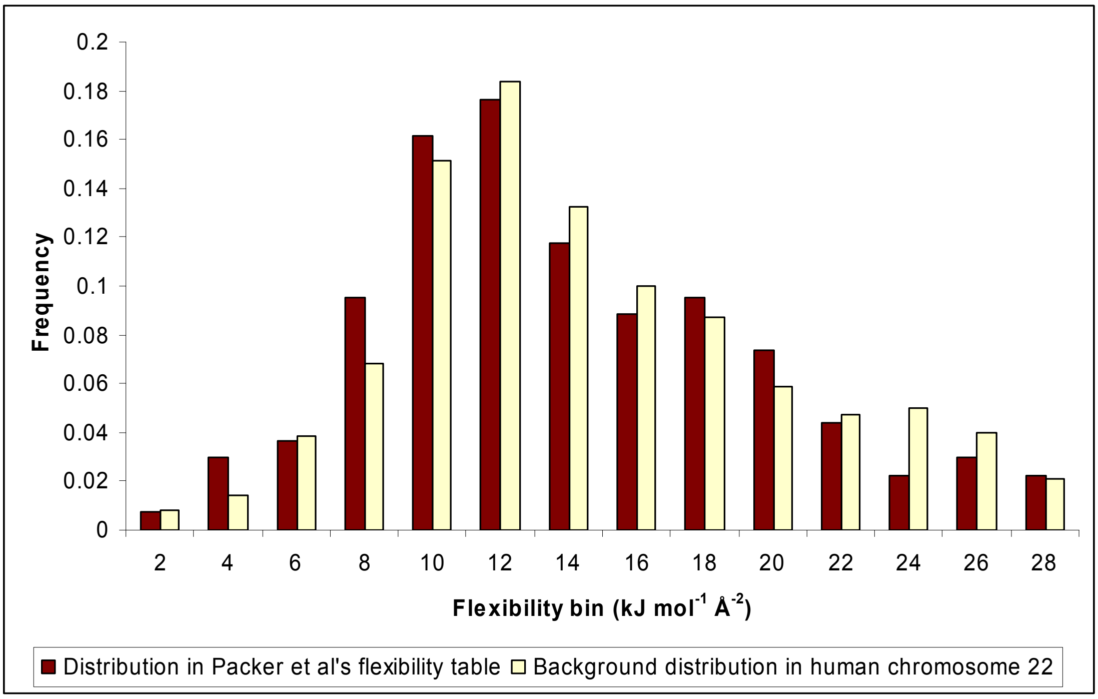
2.3.1 An Introduction to flexibility sequences

One of the fundamental concepts of nucleosome positioning is that it is an effect of the physical properties of the underlying DNA sequence. This made it necessary to model DNA sequences as sequences of physical DNA parameters. This section will introduce these kinds of sequences, herein referred to as “flexibility sequences”. The flexibility sequences described in this section was used for wavelet analysis (discussed in Section 2.4.1). Section 2.3.2 will introduce a simpler kind of flexibility sequence for using as emission symbols for HMMs.

For the work carried out in this thesis, a table which provides flexibility values for all 256 possible tetranucleotide steps (4^4 combinations) (Packer *et al.*, 2000b) was used to translate a given DNA sequence into its corresponding flexibility sequence. According to these studies, certain dinucleotide steps, represented within the larger tetranucleotide steps, were ‘sequence-independent’. Their conformation appears to be constant regardless of neighbouring sequences; an example of this is [AA/TT] whose physical basis was discussed earlier (Section 1.4.1). At the other extreme, sequences such as [CA/TG] are ‘sequence-dependent’ as their conformation is strongly influenced by the immediate DNA sequence context. This is why a tetranucleotide-based flexibility table was used rather than a lower di- or tri- nucleotide based flexibility table since it would be able to model the contexts of the sequence-dependent dinucleotides slightly better.

The parameters in this table were estimated using force field measurements, which are mathematical formulas for expressing energy as a function of physical conformation (Sprous, 1996). Such functions are usually sums of terms which correspond to bond angle, torsion, Van der Waals forces and electrostatic interaction energies. These parameters correlated reasonably well with the limited tetranucleotide parameters available from X-ray crystallography (Hunter & Lu, 1997; Packer *et al.*, 2000b). The values in the flexibility table range from 1.9 (most flexible) to 27.2 (most rigid) and there are a total of 102 unique flexibility values. As can be seen from Figure 2.4, the distribution of the flexibility values is negatively skewed in both the flexibility table and in background human genomic DNA. Those tetranucleotide sequences which exhibit the highest rigidity generally contain [AA/TT] dinucleotides.

Figure 2.4: Histogram of DNA flexibility values (Packer *et al.*, 2000b)



A DNA sequence was converted to this kind of flexibility sequence using the following steps:

- A 4 bp window was taken at position 1 of the DNA sequence.
- Its corresponding flexibility value was looked up and stored as the first symbol of the flexibility sequence.

- The window was shifted by 1 bp and the next value looked up; this was stored as the second symbol of the flexibility sequence.
- Steps 2-3 were repeated until reaching 3 bp from the end of the DNA sequence.

2.3.2 Flexibility emission alphabet for using with HMMs

A simple flexibility emission alphabet was derived from the tetranucleotide-based flexibility table described above for using with HMMs. In the original form of this table, 102 unique symbols would have been an exhaustive emission alphabet for HMM training (compare with 4 symbols for the DNA alphabet for example). Therefore, the number of symbols had to be sized down to form a reasonable emission alphabet. This was done by firstly splitting the 256 unique tetranucleotide sequences into 6 equally binned categories ranked by ascending values of flexibility. Each of the 6 bin categories represented a symbol of the new compressed alphabet: these new symbol values were assigned from 1 for most flexible to 6 for most rigid. So for example, the ‘most flexible’ category would contain the 42 (256/6) most flexible tetranucleotide sequences of the original table. In this way, a compressed 6-symbol flexibility lookup table for tetranucleotide DNA sequences was derived. This table was used to convert a DNA sequence into its corresponding 6-symbol flexibility sequence using the same steps outlined in Section 2.3.1.

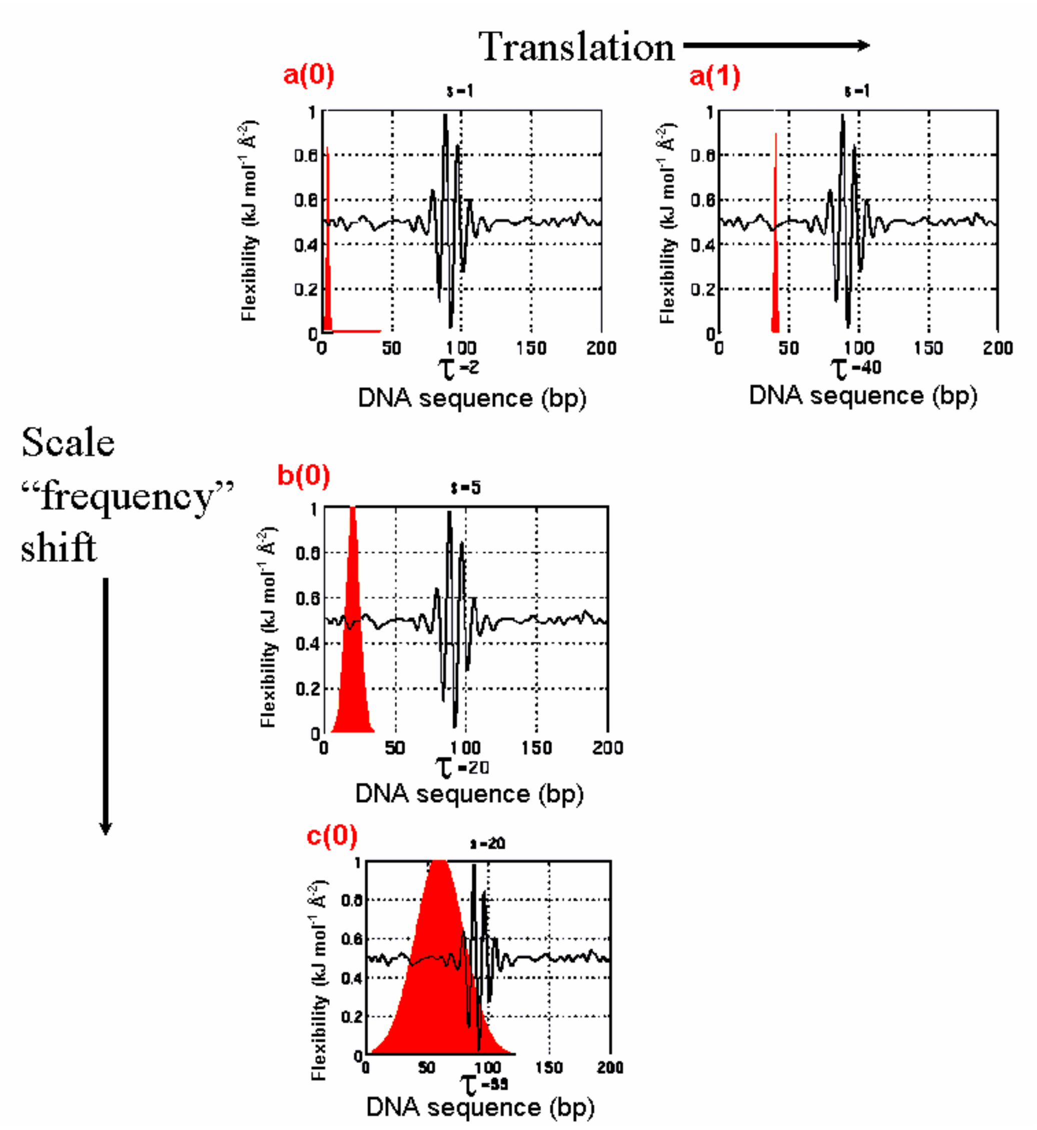
2.4 A Basic Introduction to Wavelets

2.4.1 An introduction to wavelets

Wavelets are a family of mathematical transformations which reveal information about the strength and localisation of periodic patterns in a signal; this information is not apparent in the raw format of the signal. A DNA sequence can be considered as a specific kind of signal. The flexibility sequence is another representation of the same signal but from which it is easier to derive information about the sequence of structural features in the DNA sequence. There are 2 parameters which define a wavelet (Figure 2.5):

- Translation (τ) which defines a specific position along a signal and
- Scale (s) which defines a specific frequency.

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).



In Figure 2.5a(0), the wavelet function is seen as a red sine curve; it is located at its initial position 2 (the value of τ) along the DNA sequence and with a scale parameter of 1 (the value of s). This is the wavelet function at its original position and is called the mother wavelet. The following shifts in size and location are then applied to the mother wavelet:

- Firstly, the function is moved or ‘translated’ along a sequence to scan for any localised frequencies which correspond to the present value of s = 1 (Figure 2.5a(1)). In Figure 2.5a(1), the function has been shifted to a τ value of 40. τ = 80 will receive a high score at this present s value as it is very similar in size and shape to the current value of s. In this way, a score is obtained for each point along the DNA sequence which represents how strongly correlated the part of the sequence is to the present shape and size of the wavelet function.
- The scale parameter, ‘s’, is now ‘dilated’ to 5 (Figure 2.5b(0)) increasing the width of the function. It is also translated across the sequence to obtain a score for each point along the DNA sequence. One important feature is that since the scale has increased, the resolution along the ‘x’ axis has also diminished. This is a property of multiresolution which is explained in the next section. Note that the initial τ value is now at 20 which is due to the increase in width of the wavelet function.
- In Figure 2.5c(0), ‘s’ is further dilated to 20. In this way, a number of co-efficient scores are obtained for different values of ‘s’ and τ. The results can be plotted as a 2D contour map as in Figure 4.2 (page 4-122), where the intensity of the colours represent the strength of different frequencies in different regions of the DNA sequence (dark blue is strongest).

Equation 2.3 is the formula for the continuous wavelet transform. For different values of τ and s, the wavelet function is obtained as the product of the original sequence, x(t), and the wavelet function. This product is referred to as the convolution of the signal and the wavelet function; it is analogous to a correlation co-efficient between the wavelet function and a specific region of the signal. The convolved product is further multiplied by a normalisation factor 1/√|s|, which ensures that the energy of the co-efficients is distributed evenly along different scales.

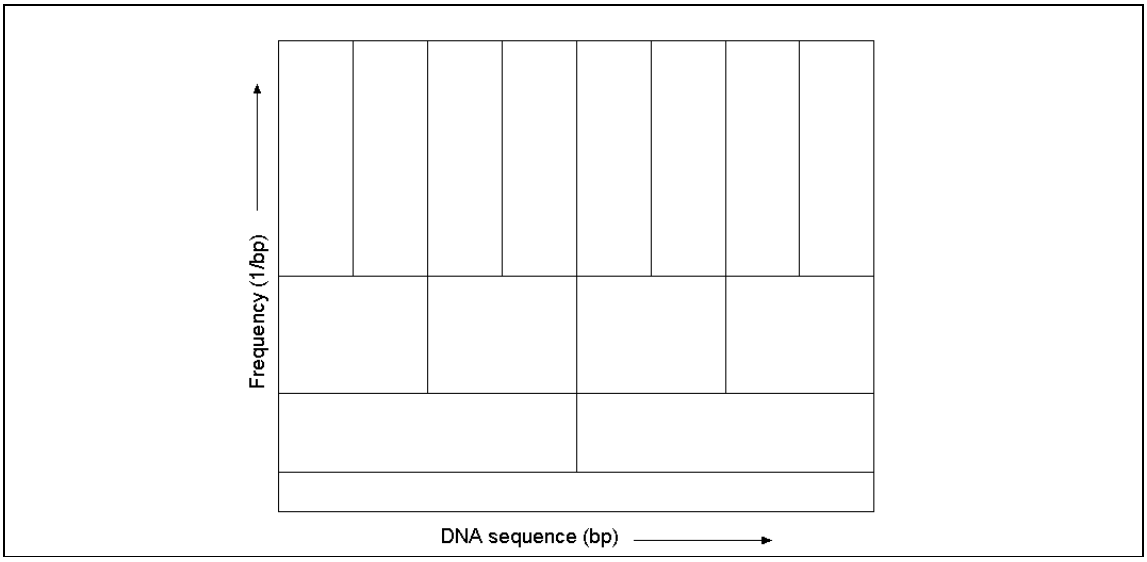
Equation 2.3: Continuous wavelet transform

$$CWT_x^{\psi}(\tau, s) = \Psi_x^{\psi}(\tau, s) = \frac{1}{\sqrt{|s|}} \int x(t)\psi^*\left(\frac{t-\tau}{s}\right) dt$$

2.4.2 The multiresolution property of wavelets

The output from a wavelet transform provides a 2 dimensional representation where the strengths of different frequencies against a DNA sequence can be viewed. However, an important feature with this kind of transformation is the multiresolution property. This states that high frequency components are resolved well in time and low frequency components are resolved well in frequency. As can be seen in Figure 2.6, as the frequencies get higher, the width of the boxes get narrower; thus this value can be resolved well along the DNA sequence. The reverse is true for low frequencies which will be resolved poorly along the DNA sequence but better along the frequency axis; this is seen as the wide box at the bottom of the frequency axis.

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.



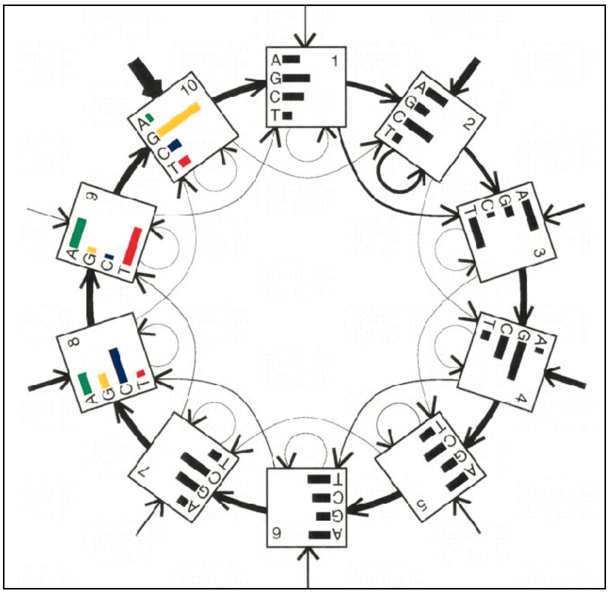
3 Cyclical Hidden Markov Model Analysis to find Signals Involved in Nucleosome Rotational Positioning

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.

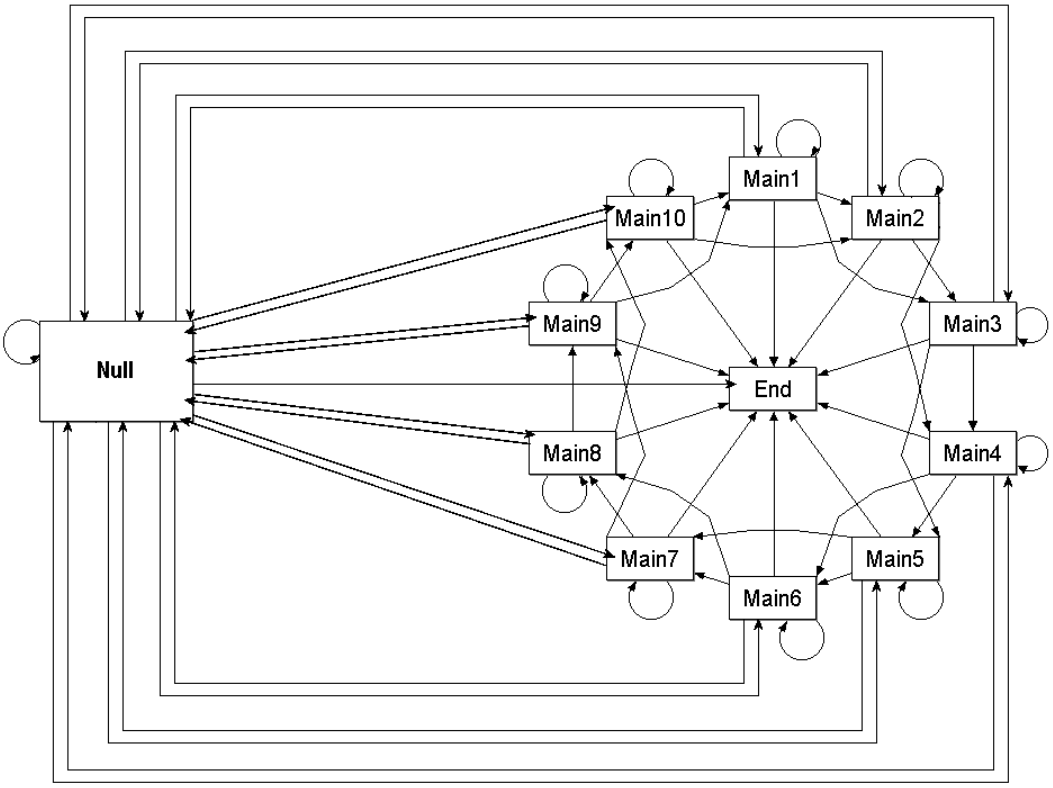
To model the physical aspect of rotational positioning more directly, a flexibility-emission alphabet was also developed to model DNA sequences as flexibility sequences (Section 2.3.2, page 2-53).

3.2 Methods

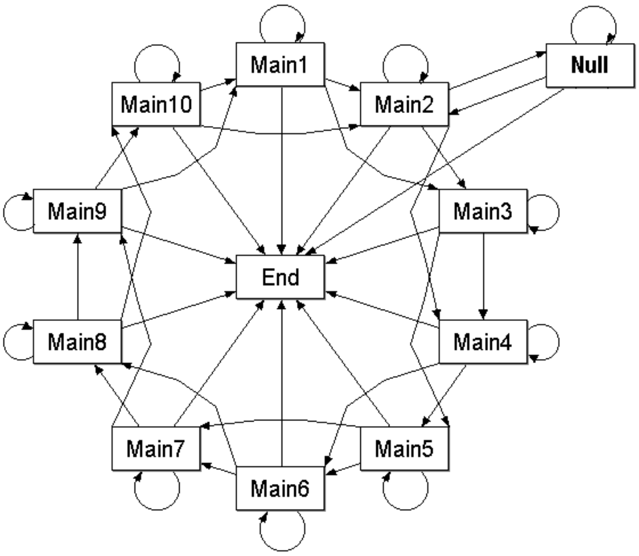
The main techniques used in this chapter involved HMM training and prediction. HMMs are introduced more generally in the introduction chapter of this thesis (Sections 2.2.1-2.2.3). This section will outline the construction, training and prediction procedure for a general architecture of HMMs, the cyclical 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

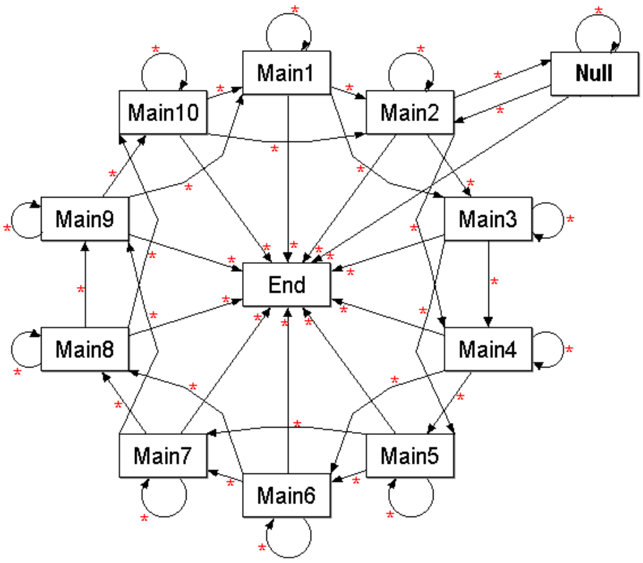
Figure 3.2: Different cyclical HMM architectures: (a) F1, (b) F2 and (c) F3.



(a)



(b)



*All transition parameters kept constant

(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

nnnnnnnnnn0123456789nnnnnnnnnnnnnnnnnnnnnnnnnnnnn0123456789n
```

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.

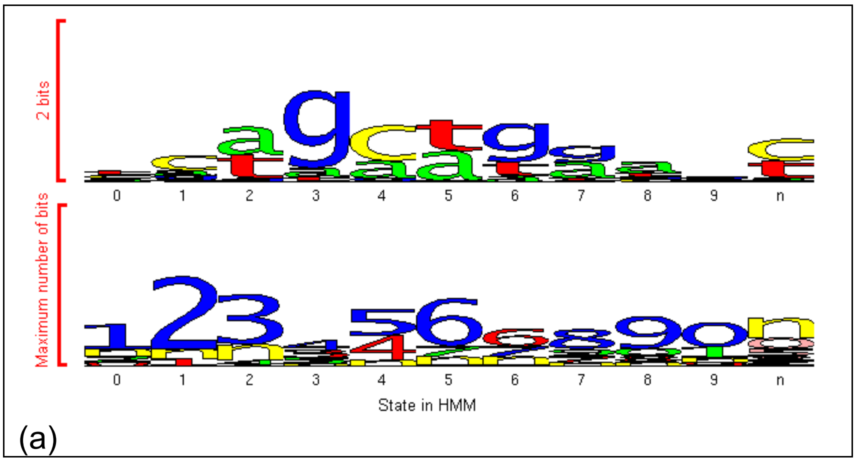
[11] The product of the wheel-labelling frequency and the frequency of the genomic feature in the chromosome

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.



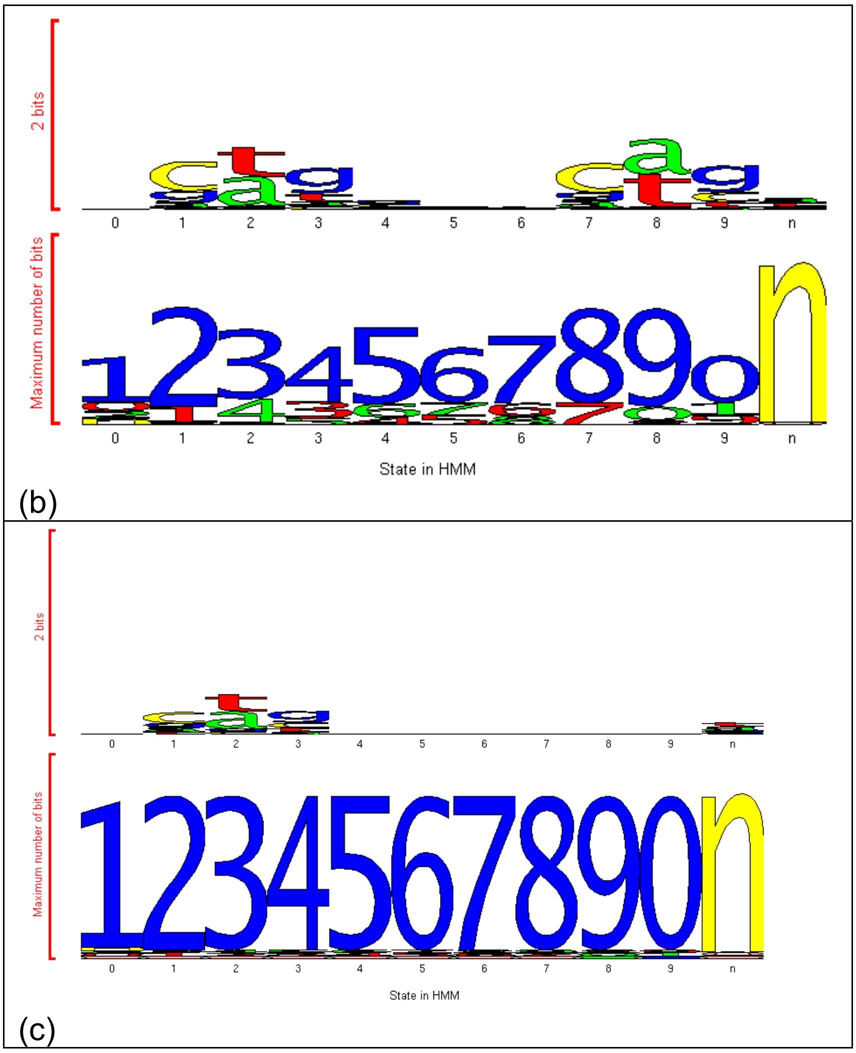
Legend for viewing transition probabilities

🟦 Next

🟥 Self

🟩 Skip

🟨 Null



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).

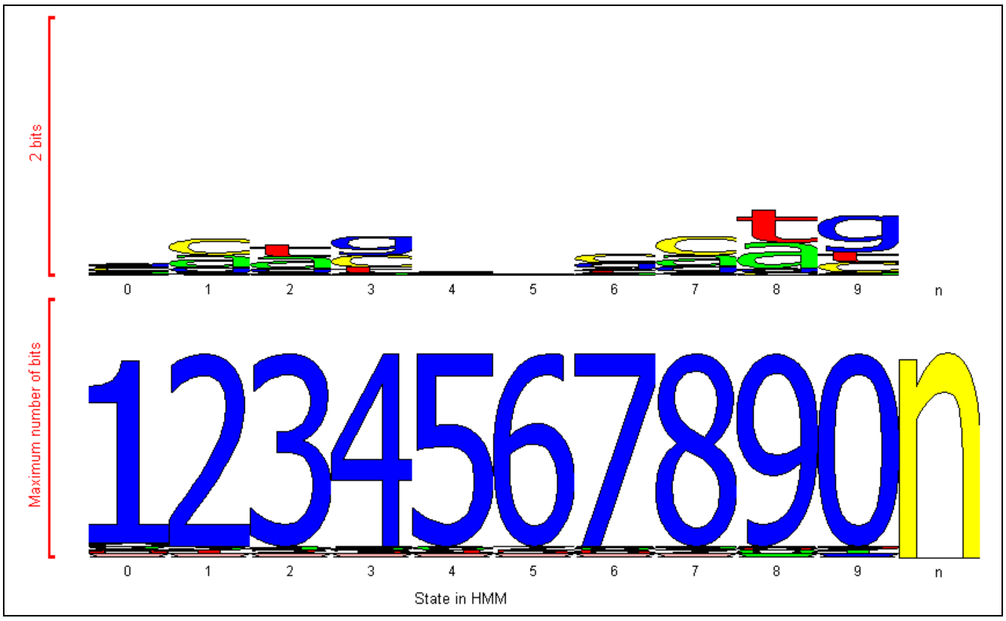
TRAINING SOURCE	STATES	SKIP	NEXT	LOOP	MOTIF
intronMasked0	6	0	2276	0	
intronMasked2		0	2283	0	
interMasked0		0	2491	0	
intronMasked1		0	2381	0	
interMasked1		0	2457	0	
interMasked2		0	2602	0	
interMasked1	7	0	2728	0	
intronMasked2		0	2199	0	
interMasked2		0	2796	0	
interMasked0		0	2458	0	
intronMasked1		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).

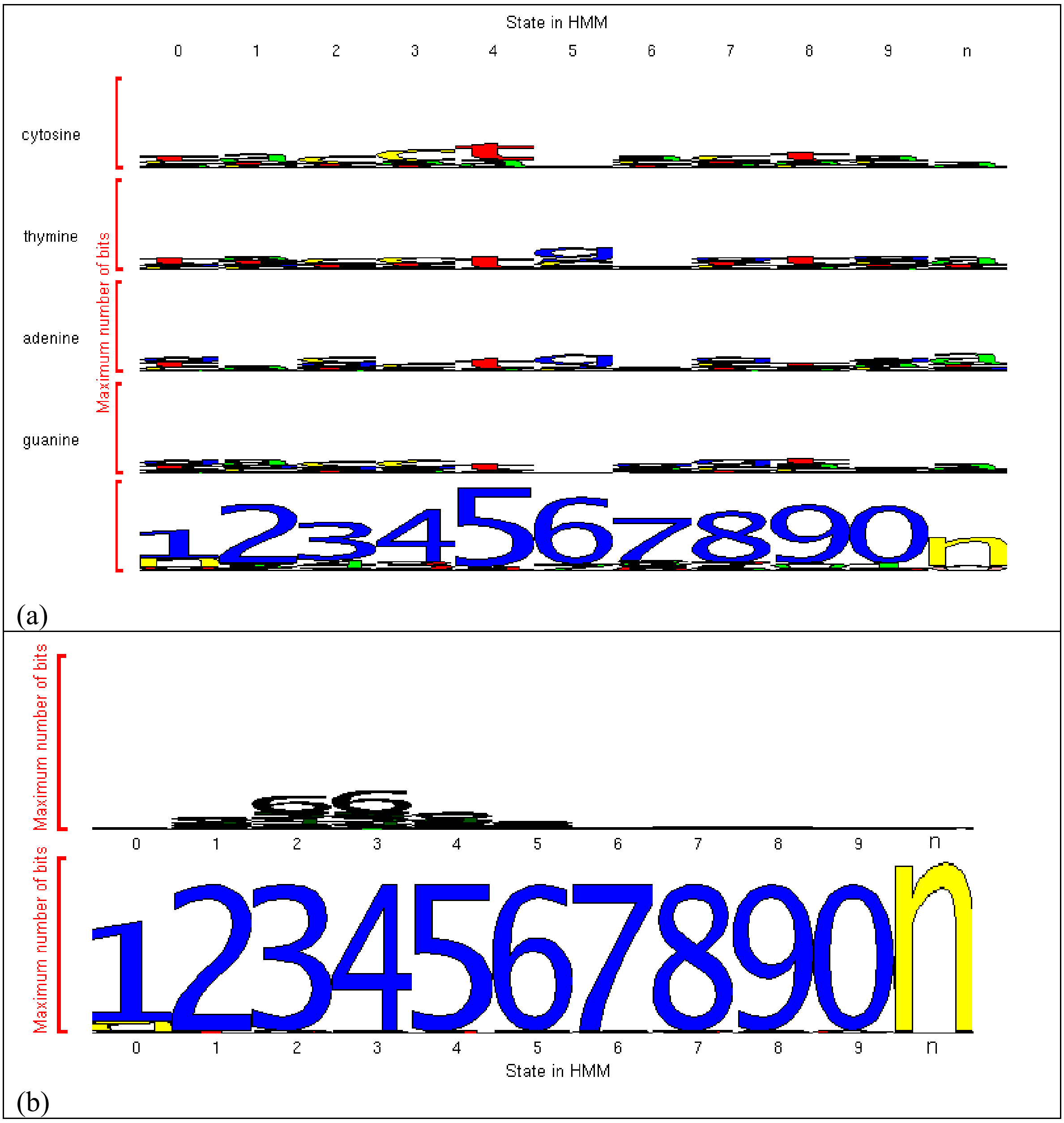
Figure 3.5: An F3 model, whose emission parameters have been crudely reproduced from the B&B model. The transition parameters were all fixed to the same value since the original parameters were not available.



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 2nd 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.

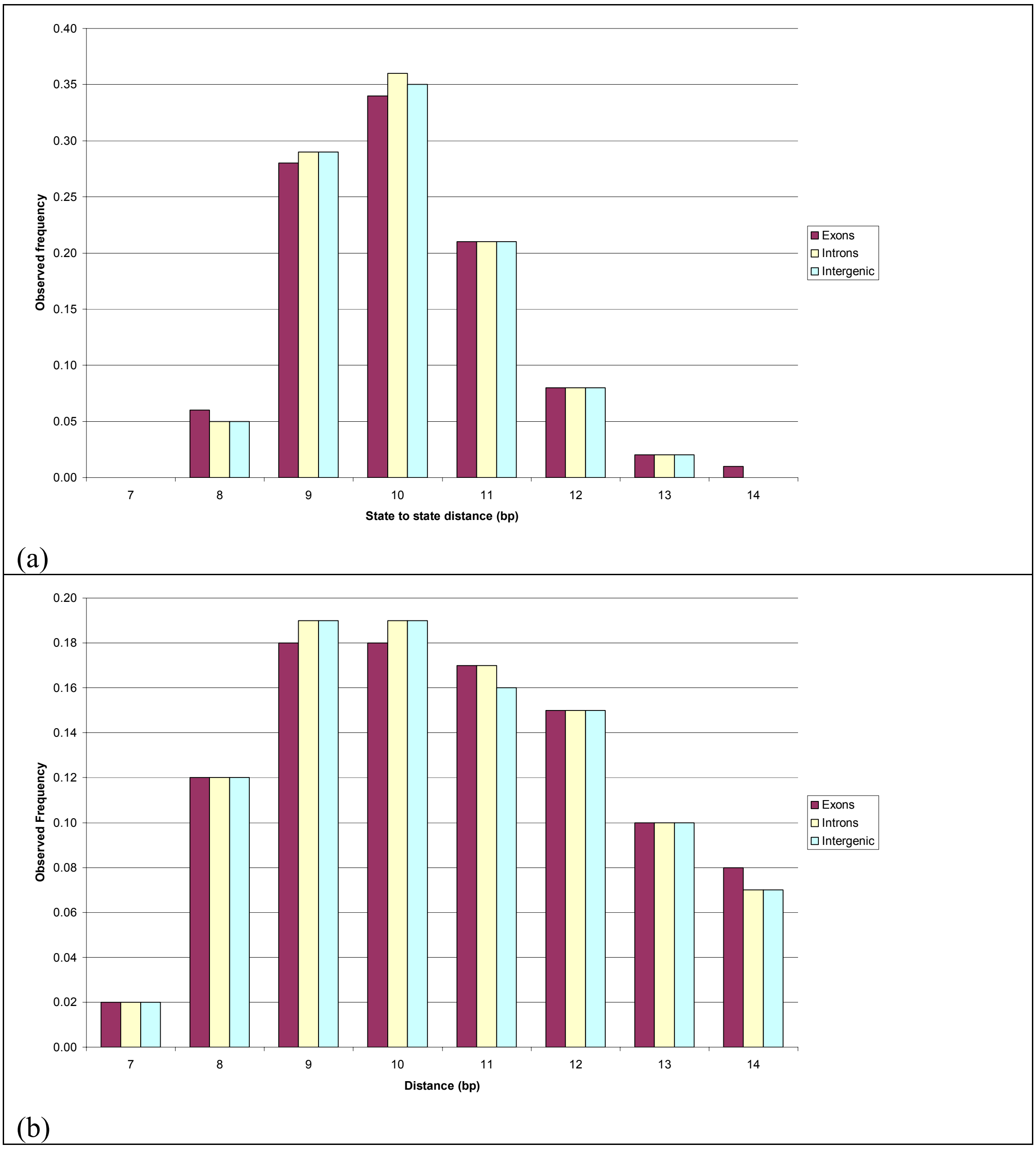
This result indicated that the structural basis for the [VWG] motif to effect nucleosome rotational positioning was perhaps not as convincing as was suggested earlier (Baldi *et al.*, 1996). However, the [VWG] motif itself was quite intriguing as it was being learnt both in coding and non-coding DNA 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).

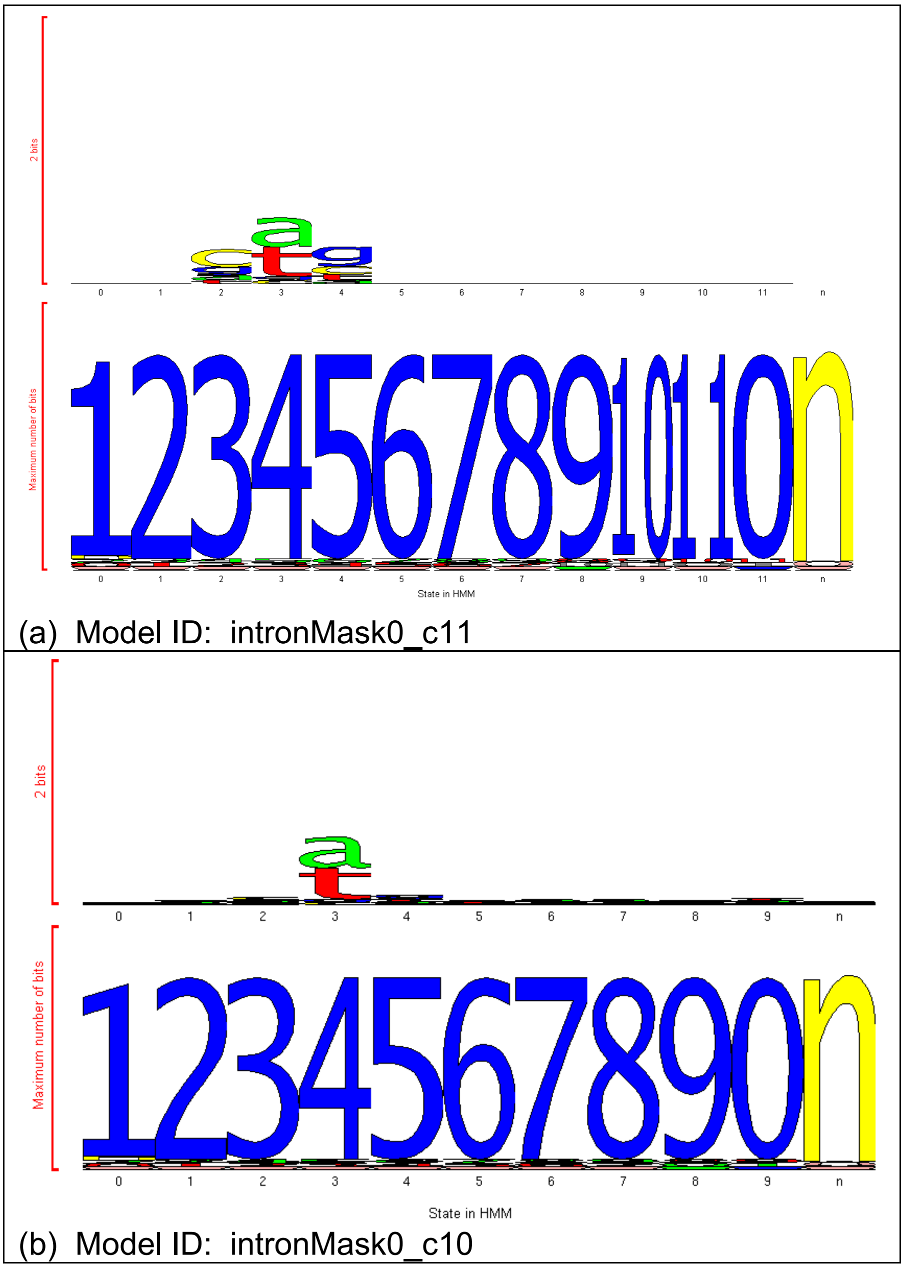
3.3.4 The [VWG] motif in retrospect and the distinction of two apparent motifs learnt in F3 human models

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).



(a) Model ID: intronMask0_c11

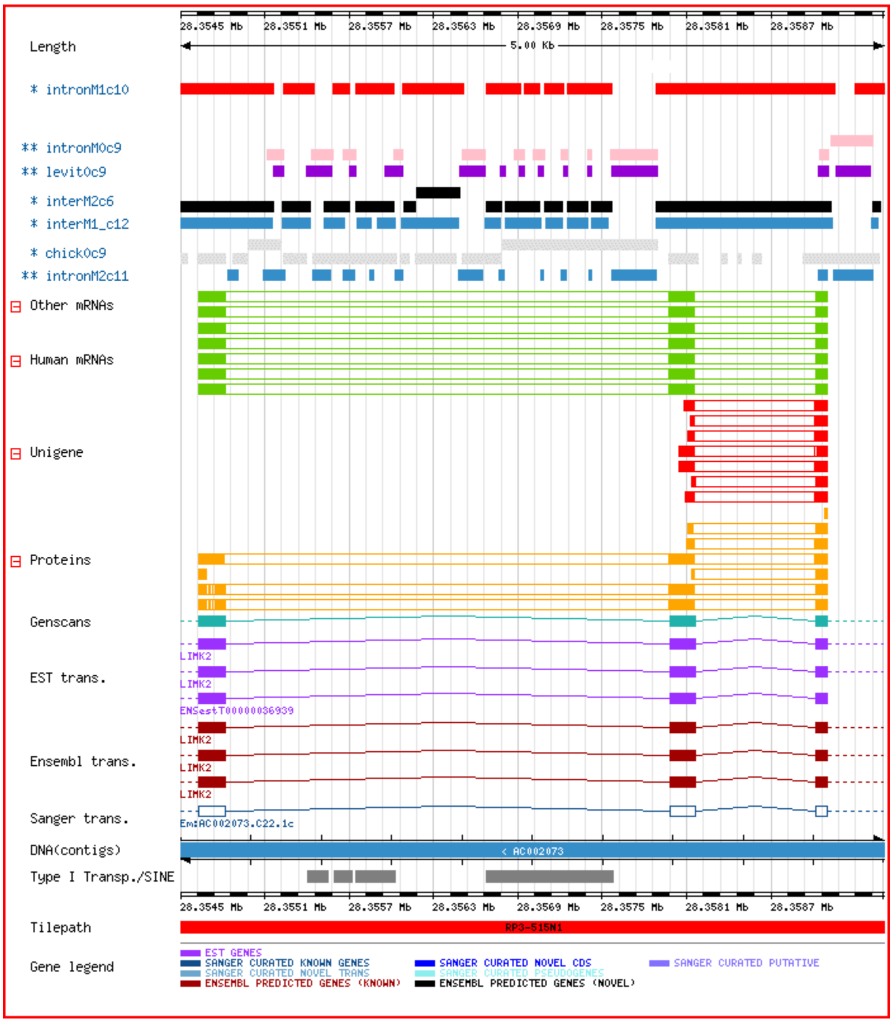
(b) Model ID: intronMask0_c10

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).

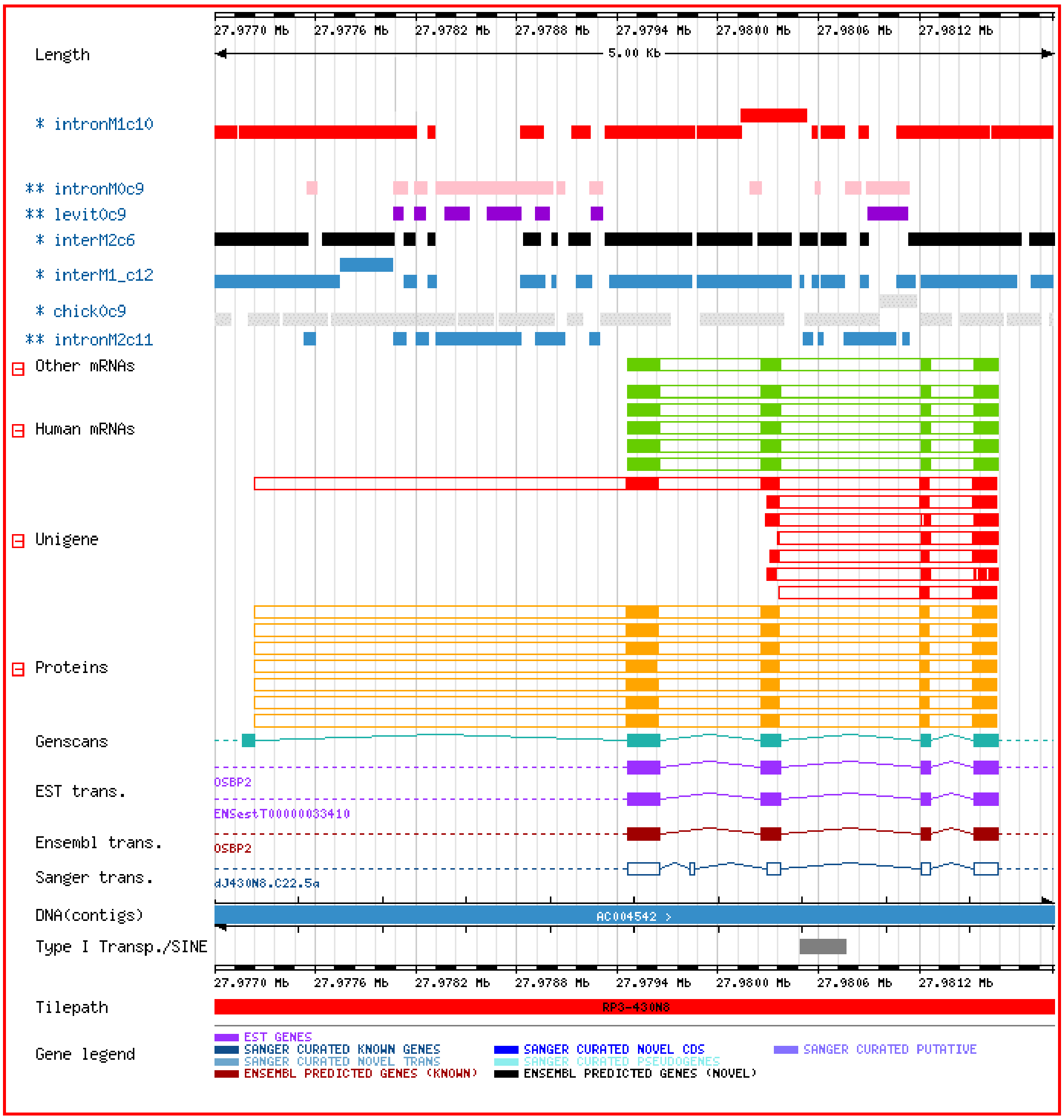
Labelling a human chromosome 22 contig with models trained from repeat-masked non-coding human sequences, showed that 2 kinds of models with complementary labelling patterns had been learnt (Figure 3.9). Figure 3.9(a)-(c) shows that there were 2 opposing labelling patterns. Of the 5 models trained from 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.

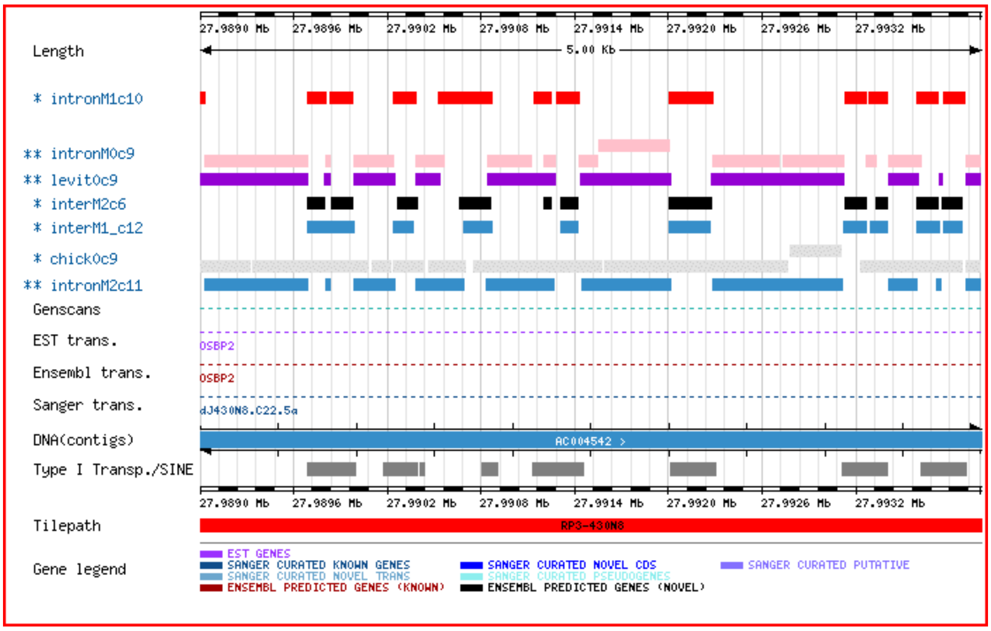
Legend: * = F3 model which learnt a [CWG] motif; ** = F3 model which learnt a [W] motif.



(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)

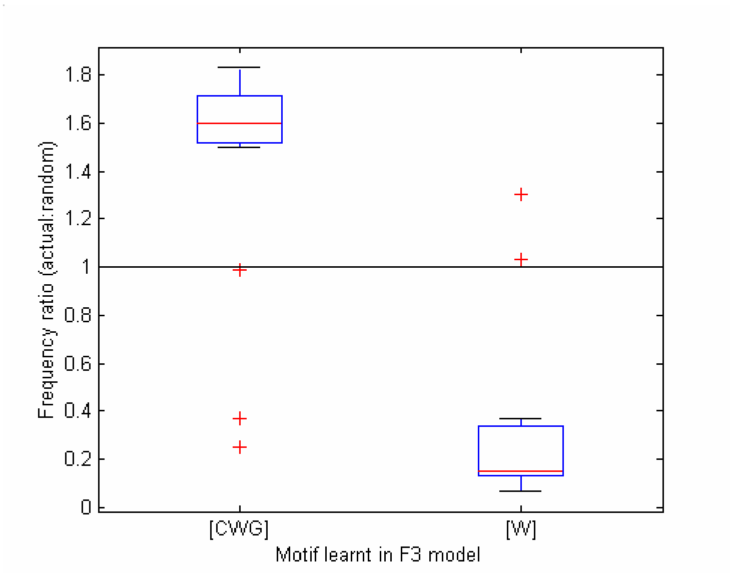

27.9890 Mb
27.9896 Mb
27.9902 Mb
27.9908 Mb
27.9914 Mb
27.9920 Mb
27.9926 Mb
27.9932 Mb
Length
5.00 Kb
* intronM1c10
** intronM0c9
** levit0c9
* interM2c6
* interM1_c12
* chick0c9
** intronM2c11
Genscans
EST trans.
OSBP2
Ensembl trans.
OSBP2
Sanger trans.
dJ430N8.C22.5a
DNA(contigs)
AC004542 >
Type I 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).

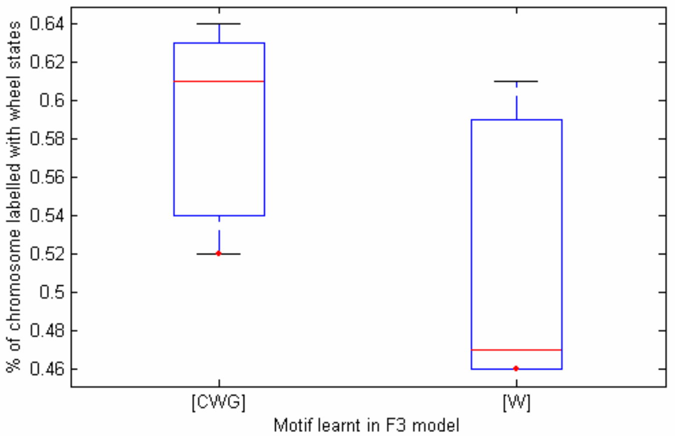
Figure 3.10: Comparison of model to model labelling. An F3 model, which had learnt a [CWG] motif (Model ID *interMask1_c12* in Appendix B), was used to label a 2.5MB sequence of human chromosome 22. The labelling of this was compared to the labelling of other models, of different wheel sizes, whose apparent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



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

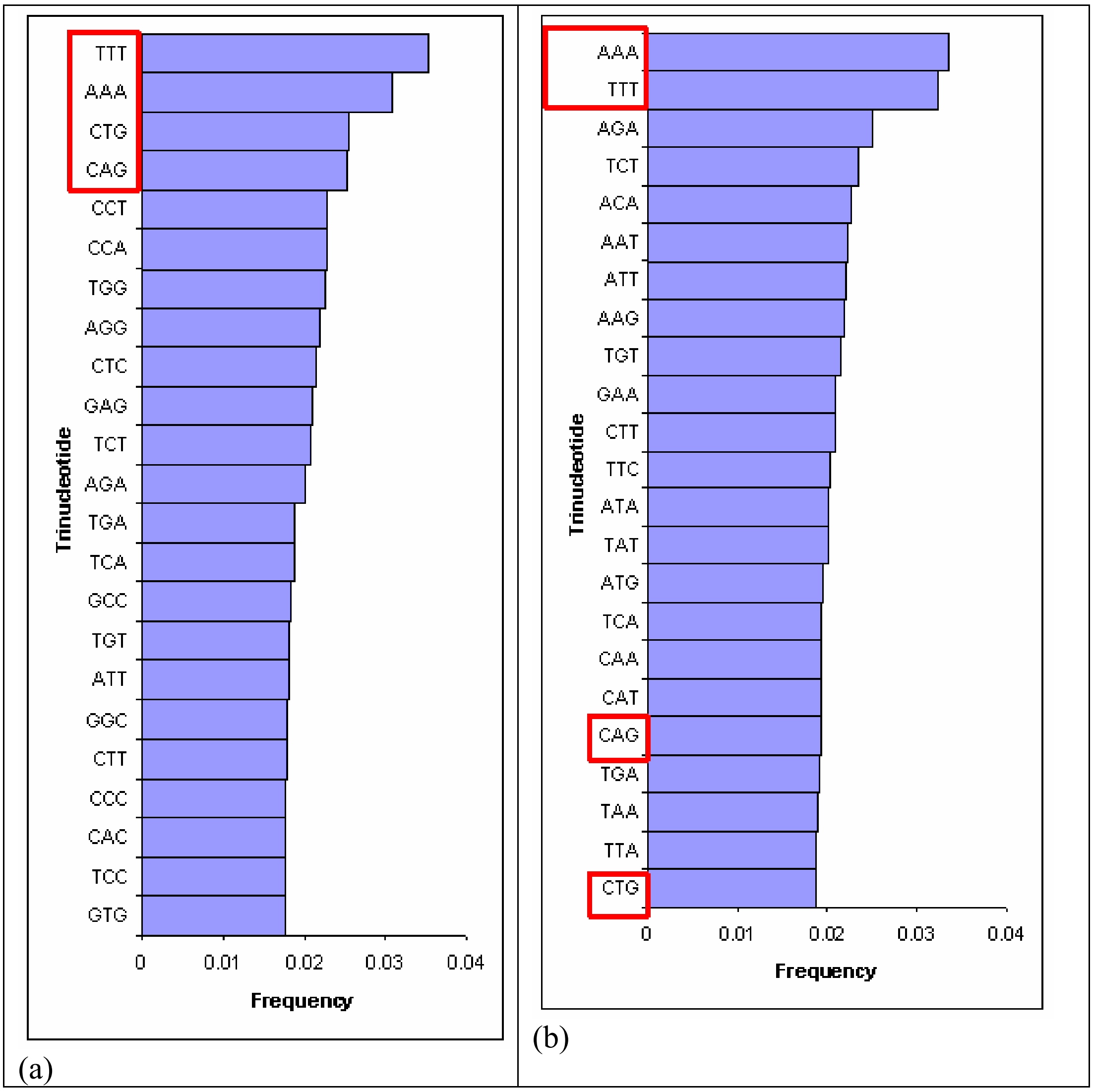
On average, in human, 60% of the test chromosome 22 contig was labelled as wheel states by [CWG] models and 52% by [W] models (Figure 3.11); therefore, there was likely to be some overlap (~8%) between the 2 mostly opposing labelling patterns.

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

GTCTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

>Charlie1b#DNA/MER1_type

CAGCGGTTCTCAAAGTGTGGTCCGNGGACCCCTGGGGGTCCCCGAGACCCTTTCAGGGGGTCCGCGAGG

TCAAAACTATTTTCATAATAATACTAAGACGTTATTTGCCTTTTTCACTCTCATTCTCTCACGAGTGTA

CAGTGGAGTTTTCCAGAGGCTACATGACGTGTGATGTCGCAACAGATTGAATGCAGAAGCAGATATGAG

AATCCAGCTGTCTTCTATTAAGCCAGACATTAAAGAGATTTGCAAAAATGTAAAACAATGCCACTCTTC

TCACTAAATTTTTTTGTTTTGGAAAATATAGTTATTTTTCATAAAAATATGTTATTTATGTTAACATGT

AATGGGTTATTATTATTTTTAAATGAATTAATAAATATTTTAAAAATTTCTCAGTTTTAATTTCTAATA

CGGTAAATATCGATAGATATAACCCACATAAACAAAAGCTCTTTGGGGTCCTCAATAATTTTTAAGAGT

GTAAAGGGGTCCTGAGACCAAAAAGTTTGAGAACCGCTG
```

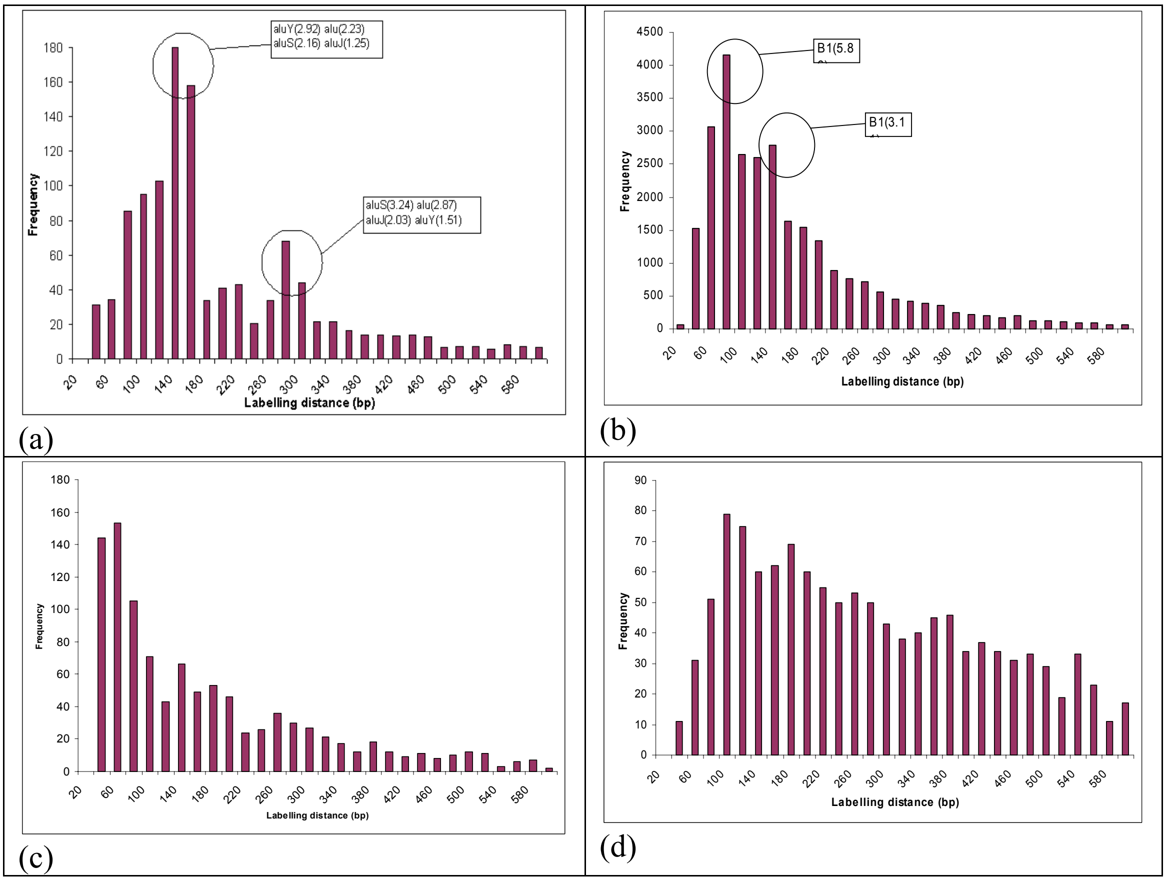
- **Lengths of wheel-labelled regions**

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

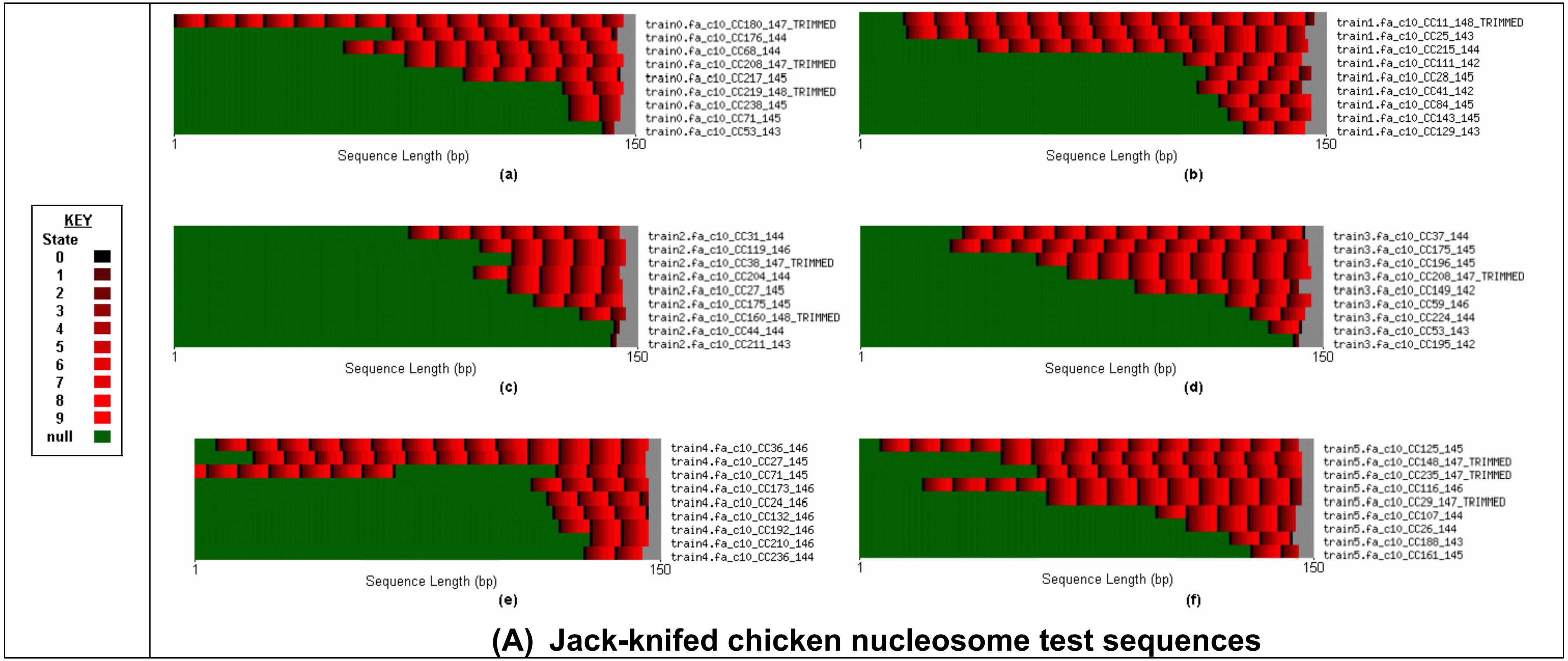
Labelling chicken nucleosome and genomic test sequences using chicken nucleosome-trained models highlighted some differences in the 2 types of test sequences. The models that were used to perform the alignments had all learnt [CWG] within the wheel component of the model.

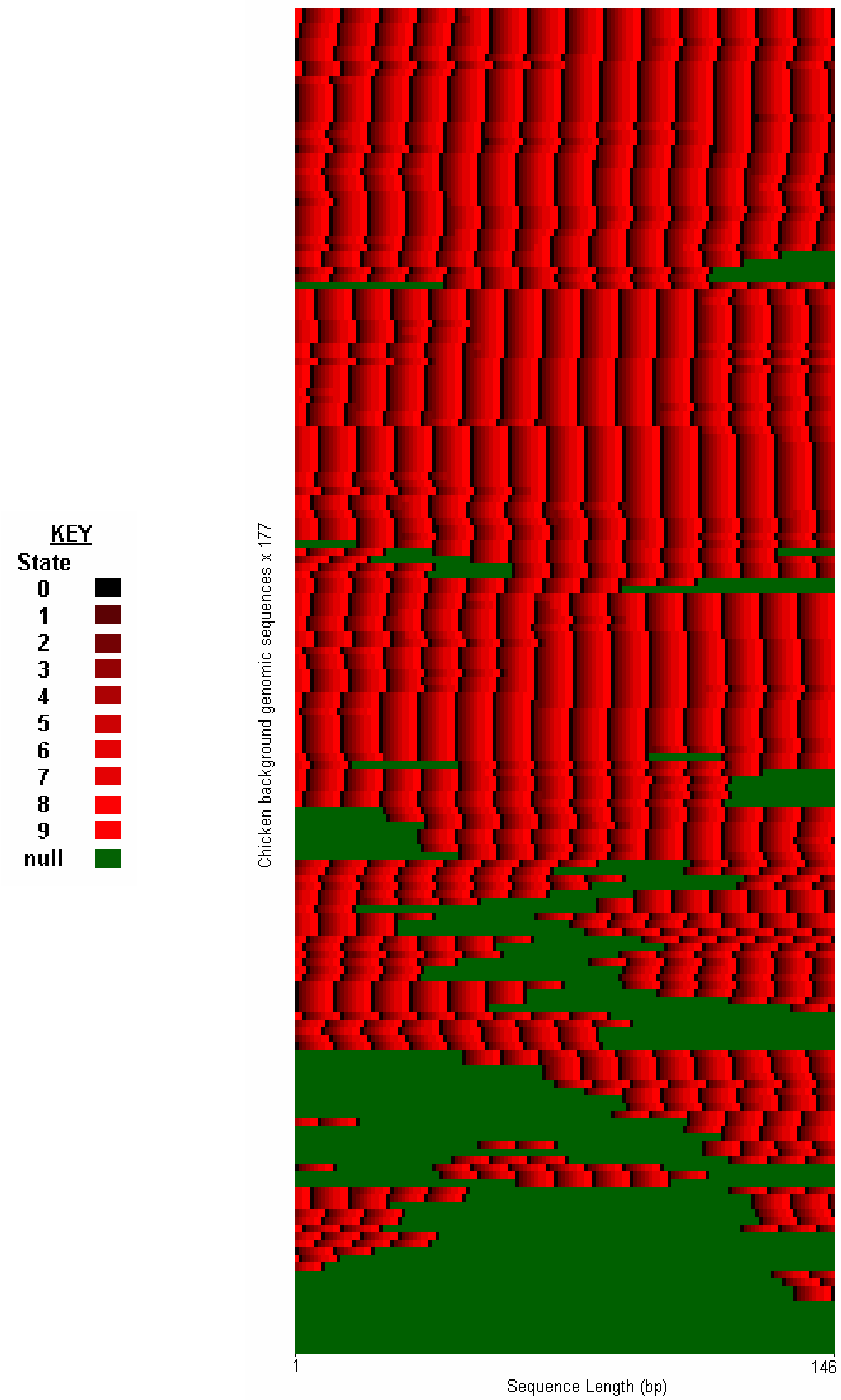
- **Alignment of chicken nucleosome sequences**

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.

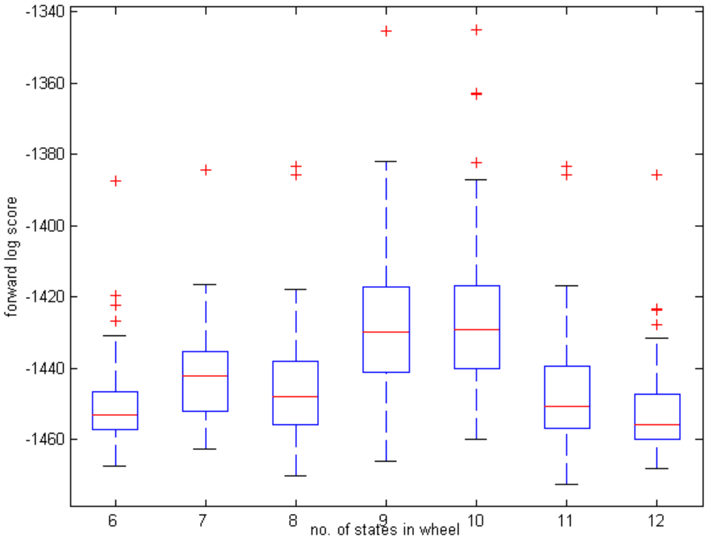
Models, which learnt the [W] motif, however, did not have any models of a specific wheel size which appeared to score better than the others (Figure 3.16(c)). So the [CWG] motif may have an enrichment at 9 and 10 bp in intergenic DNA but the [W] motif appeared random over the range of 6–12 bp; this suggested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

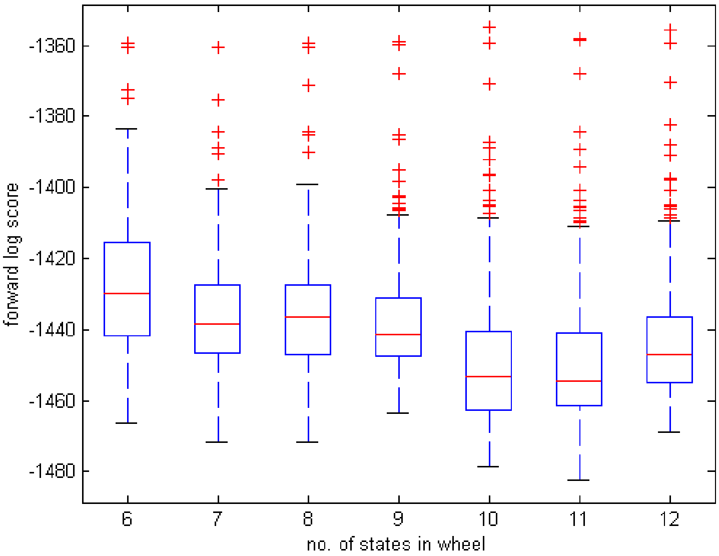
(a) Masked intergenic DNA labelled with [CWG]-learnt models,

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

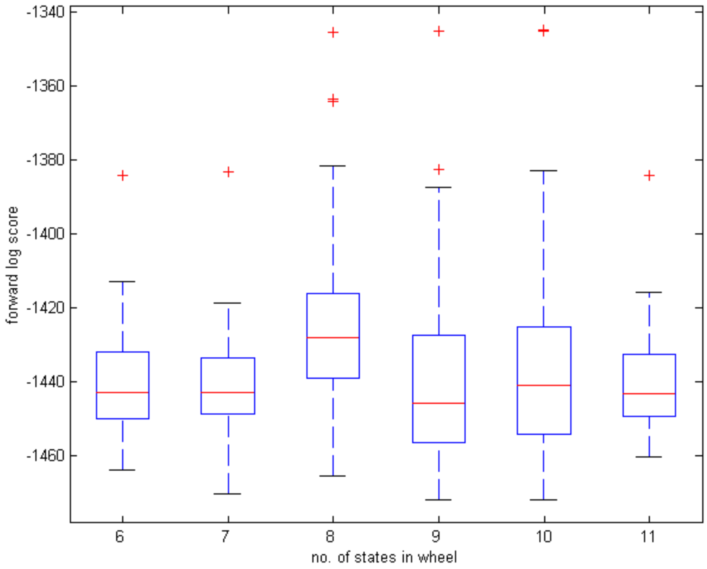
(c) masked intergenic DNA labelled with [W]-learnt models



(a)



(b)



(c)

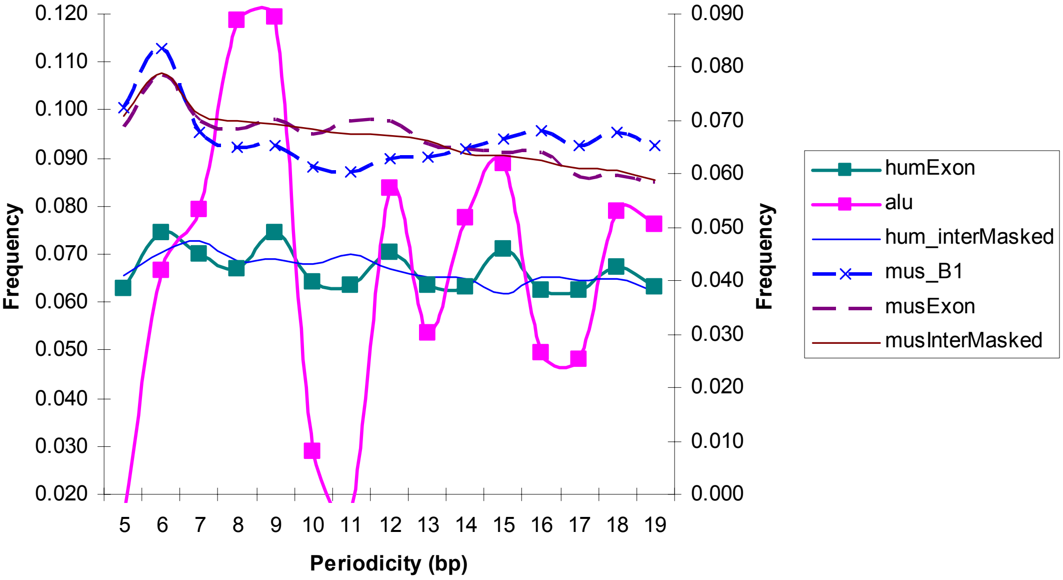
- **Motif-spacing frequency**

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).

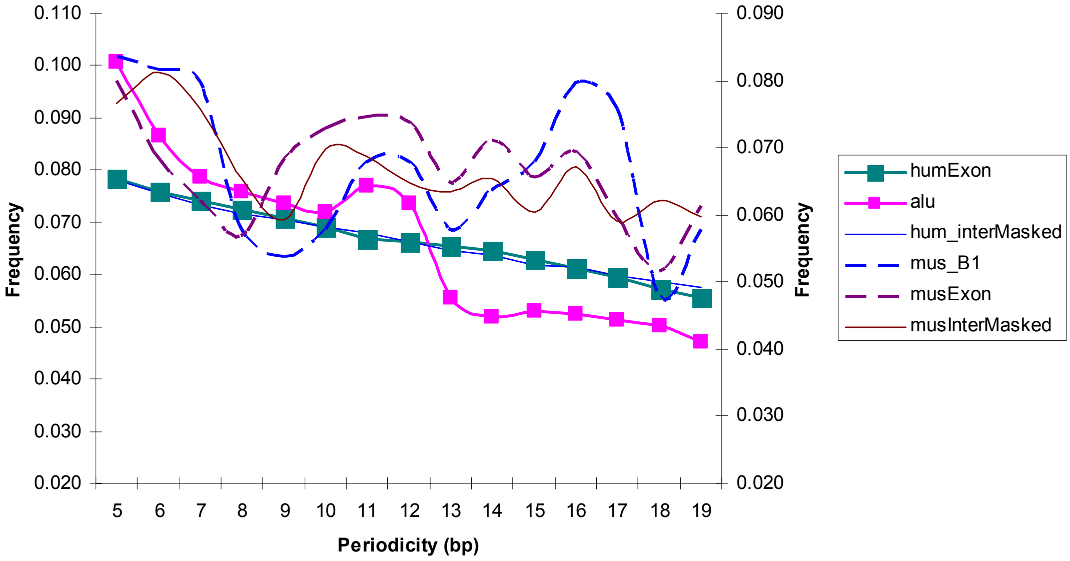
The repeat frequencies of [WWW][14] motifs, on the other hand, did not show any peaks which could suggest coding bias (Figure 3.17(b)).

[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)

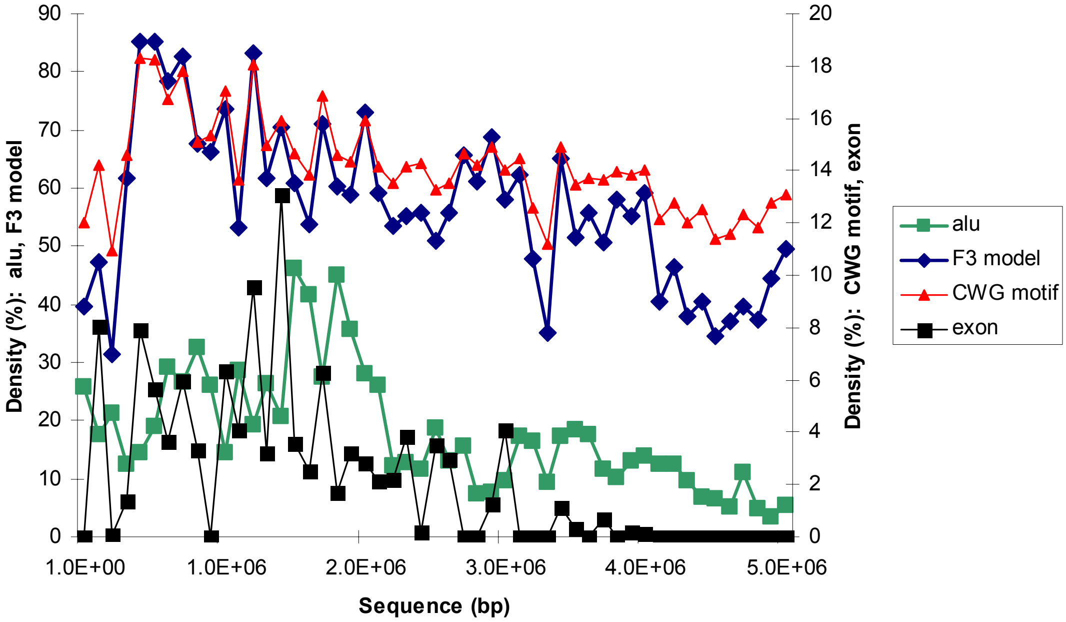
The overall impression was that the [CWG] motif did appear to be influenced by coding bias as a 3-modulo repeat of the pattern was observed. It was seen to be enriched at certain periodicities (8, 9, 12, 15 and 18 bp in human; 6, 9, 12, 16, 18 bp in mouse) and this appeared to be common for both exons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

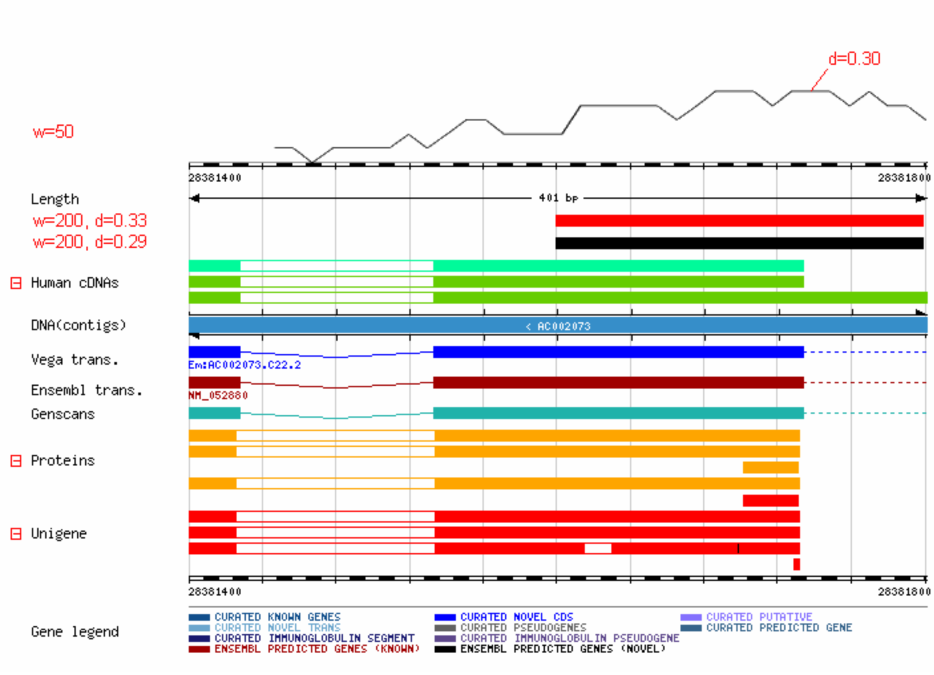
[17] A window size of 200 bp was chosen since it was close to ~146 bp, the nucleosome core particle size

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).

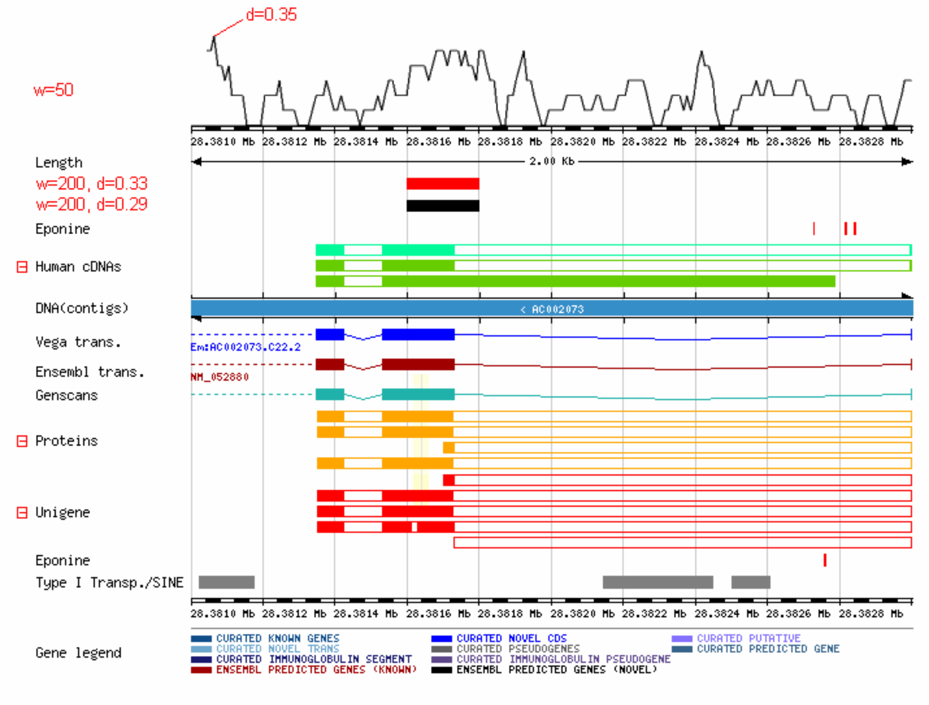
Table 3.6: Features observed to frequently have high densities of [CWG] repeats. A window size of 200 bp and cutoff threshold of 35% [CWG] density was used.

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

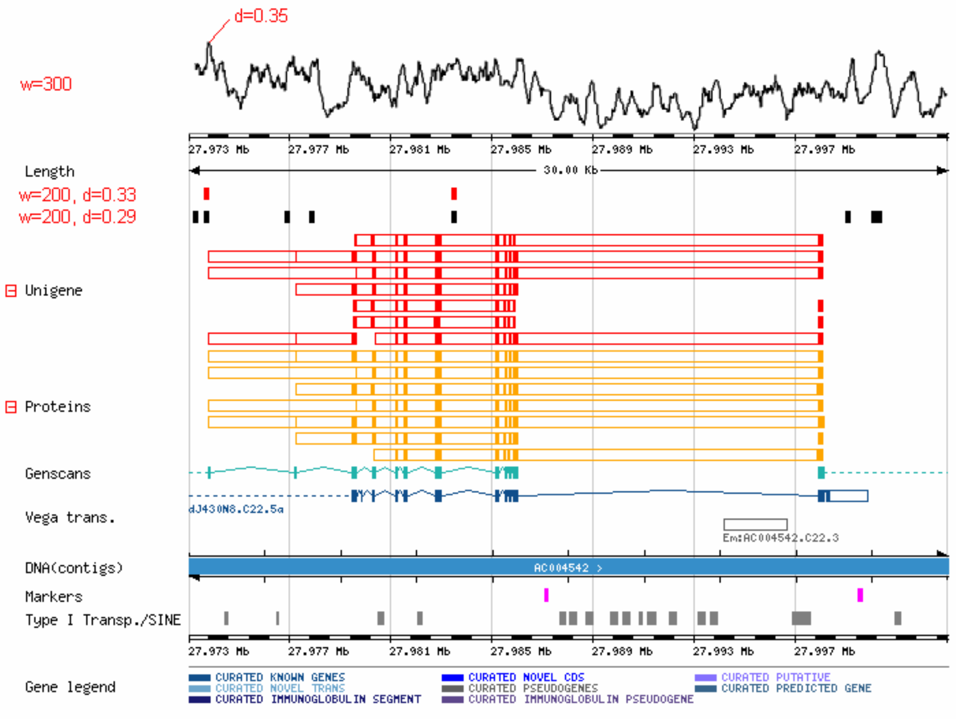
Figure 3.19: Density plots of [CWG] repeats in a human genomic sequence shown at different resolutions. 'w' is the window parameter and 'd' the threshold density of [CWG] within the window. The top density plot is a 'moving average' representation. The red and black boxes below represent non-overlapping 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.
Em:AC002073.C22.2
Ensembl trans.
NM_052880
Genscans
Proteins
Unigene
Eponine
Type I Transp./SINE
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
Em:AC004542.C22.3
DNA(contigs)
AC004542 >
Markers
Type I Transp./SINE
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).

4 Periodic Flexibility Patterns in DNA: a Scan for Signals Involved in Nucleosome Translational Positioning

4.1 Introduction

Some recent computational approaches have indicated periodic occurrences of flexibility patterns in the range of 100-200 bp in eukaryotes but not in prokaryotes (Section 1.11.2). This suggests that these flexibility patterns could be involved in positioning nucleosomes, owing to their size which is of the size order of a nucleosome core particle (146 bp). This made it interesting to examine where such flexibility patterns are located with respect to gene features in eukaryotic genomes. The availability of mouse genomic sequences, particularly syntenic regions shared with human, was a benefit to this investigation as it could also be investigated whether such potential translational positioning signals were a general mechanism conserved in evolution. The approach taken was to use the wavelet tool (Section 2.4.1) to analyse the occurrences and distribution of flexibility patterns in genomic sequences.

4.2 Methods

4.2.1 Construction of flexibility sequences

Flexibility sequences (Section 2.3.1) were used to represent DNA as sequences of conformational flexibility values.

4.2.2 Wavelet transform of whole chromosomal flexibility sequences

Wavelet transforms were performed on whole chromosomal flexibility sequences using the software *Autosignal* (Clecom, 1999). The *Morlet* family of wavelets was used. This wavelet family is considered 'crude' in the respect that once transformed, the original data cannot be reliably reconstructed. However, signal reconstruction was not required in this analysis. The *Morlet* was an appropriate family to use for transforming flexibility sequences as it is suited for decomposing continuous data series such as flexibility sequences. The particular implementation of the *Morlet* family that was used was also a fast one, which calculates the Fourier transform of both the *Morlet* waveform and the raw signal (flexibility sequences) to achieve fast convolution.

The main datasets that were transformed and analysed were[18]:

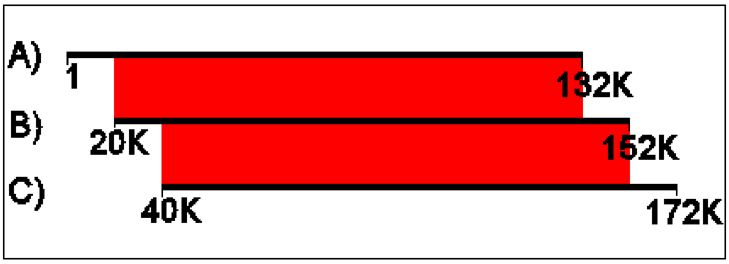
- Mycobacterium tuberculosis (Genbank ID: AE000516),
- Saccharomyces cerevisiae (Genbank ID: NC_001147),
- Human chromosome 20,
- Human chromosome 22,
- Mouse chromosome 19,

[18] Human and mouse data were extracted from the Ensembl database (Clamp *et al.*, 2003; Hubbard *et al.*, 2002)

- A 30MB syntenic region between human chromosome 20 (29.4MB to 62.9MB) and mouse chromosome 2 (172.1MB to 202.3MB).
- BRCA2 syntenic region between human and mouse (a 1.2 MB sequence alignment)

The period range which was analysed was 50-1000 bp; this range was selected such that periodic patterns of the length order of the nucleosome core particle (~146 bp) could be detected. Due to memory limitations as well as the software design constraints, the maximum sequence length that could be transformed at a time was 132,000 bp. Therefore, to handle chromosome-size data which covered several MB, a windowing scheme was used. Apart from the maximum data size, another limitation was the occurrence of edge effects associated with this wavelet family. These would result in a large amount of false classification towards the window edges. Therefore, an overlapping windowing scheme was adopted to minimize these effects (Figure 4.1). The start of each window was offset by a small amount (20,000 bp) relative to the size of the full analysis window (132,000 bp). So, for instance in Figure 4.1, strong patterns between co-ordinates 40,000 bp and 132,000 bp would only be considered if they appeared in all 3 analysis windows A, B and C.

Figure 4.1: Overlapping windowing scheme for removing edge effects in 'wavelet transform' analysis windows.



4.2.3 Thresholding by wavelet co-efficient strengths

The wavelet co-efficients, which represent the strength of a specific period along a flexibility sequence, are complex numbers. For purposes of visualisation and thresholding, these values were converted to decibels (dB) in *Autosignal*. This is measured as:

$$10.0 \text{ x } \log_{10}(r^2+i^2)$$

where *r* and *i* are the real and imaginary components of the wavelet co-efficients respectively. The strongest co-efficients, thus obtained in chromosomal flexibility sequences, were around 30.0 dB and the weakest were around -248.0 dB (0.0 dB is considered to be the lower limit for comparing 2 signals). 2D contour maps of the strengths of different wavelet co-efficients were plotted as in Figure 4.2 (page 4-122). For visualising the locations of strong patterns on sequences longer than the size of the wavelet analysis window, only regions stronger than 28.0 dB were plotted (for example, Figure 4.3, page 4-124).

4.2.4 Probability distribution of periodic flexibility patterns

The probability of observing a flexibility pattern, corresponding to a specific repeat period in the genome, was calculated as the total length occupied in a chromosome by such patterns divided by the total length of the chromosome. This was done separately for both introns and intergenic regions (for example, Figure 4.4, page 4-126).

4.2.5 Estimation of genomic features frequently associated with periodic flexibility patterns

The ratio of observed to expected frequencies was used to indicate which genomic features were frequently associated with flexibility patterns. The same procedure was used earlier (Section 3.2.9).

4.2.6 Alignment of flexibility sequences

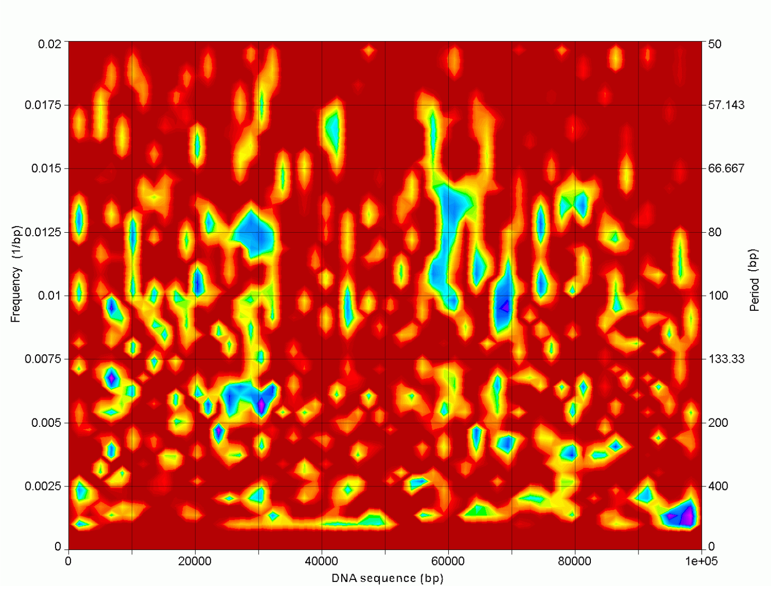
Sequences were aligned by their flexibility values in regions where strong wavelet co-efficient strengths (>28 dB) were obtained. A flexibility-sequence dataset was constructed by trimming 300 bp fragments around such regions. Following this, one sequence from this dataset was chosen randomly as a reference sequence. All other sequences were rotated until the offset of these sequences, having the strongest correlation co-efficient with the reference sequence, was found. The strongest offset flexibility sequences were then clustered and plotted as in Figure 4.6, page 4-129.

4.3 Results and Discussion

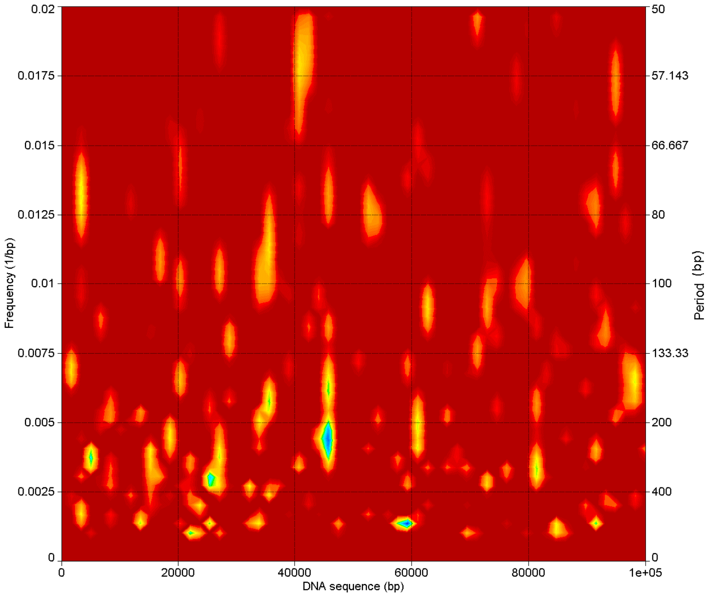
4.3.1 Differences in wavelet spectra between eukaryotic and prokaryotic flexibility sequences

The lack of nucleosome formation in prokaryotic genomes and their ubiquitous distribution in eukaryotic ones provides a reasonable basis for comparing their flexibility landscapes. To investigate this, 100 kbp-long flexibility sequences from human, *Saccharomyces cerevisiae*, and *Mycobacterium tuberculosis* were randomly selected and broken down using wavelet transformation (Figure 4.2). It was observed that in human, there was a dense distribution of periodic flexibility patterns, which was periodic between 50-1000 bp (Figure 4.2(a)). However, such patterns were not seen in *Saccharomyces cerevisiae* (Figure 4.2(b)) or in *Mycobacterium tuberculosis* (Figure 4.2(c)). Whereas the wavelet co-efficients in the human flexibility sequences were as high as 32 dB, the highest observed in *M. tuberculosis* or *Saccharomyces cerevisiae* was 24 dB. In the latter 2 genomes, there was still some weak periodicity, which was distributed sparsely. This distribution was not as dense as the stronger patterns seen in human. Upon completely randomizing the DNA sequence of the human clone and performing the wavelet transform on the corresponding flexibility sequence, the strong peaks were diminished yielding co-efficients which were now as high as 22 dB (data not shown). The lack of periodic flexibility patterns in *Saccharomyces cerevisiae* suggested that if such patterns did influence nucleosome positioning, then they would probably do so only in higher eukaryotic species.

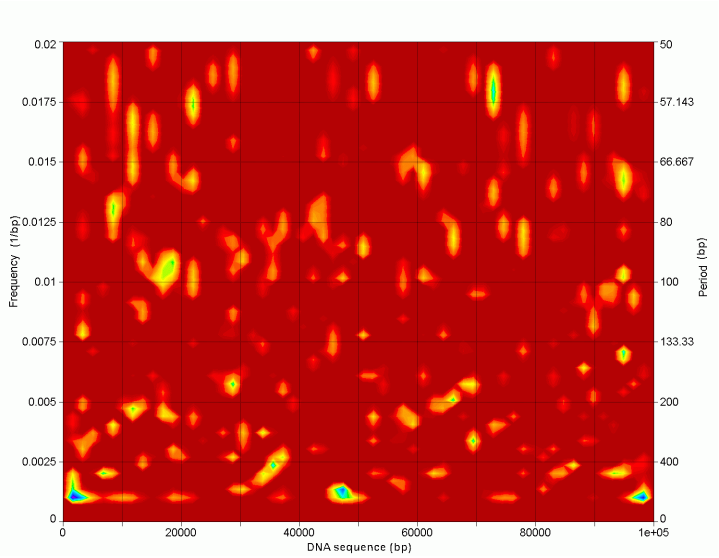
Figure 4.2: Continuous wavelet transform spectra compared between eukaryotic DNA flexibility sequences and a sample prokaryotic DNA flexibility sequence. The figures were obtained by performing the wavelet transform on randomly chosen 100,000 bp segments of the following sequences: (a) a clone from human 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.



(a)



(b)

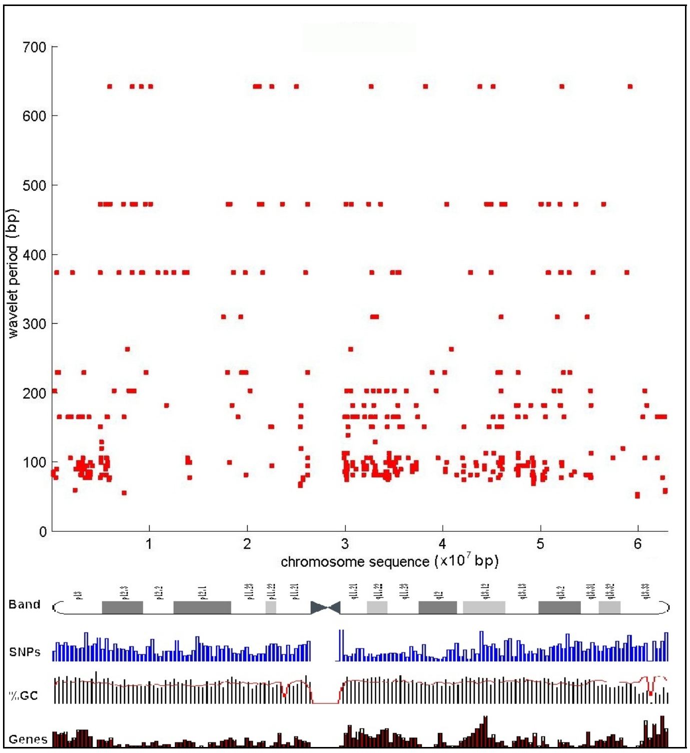


(c)

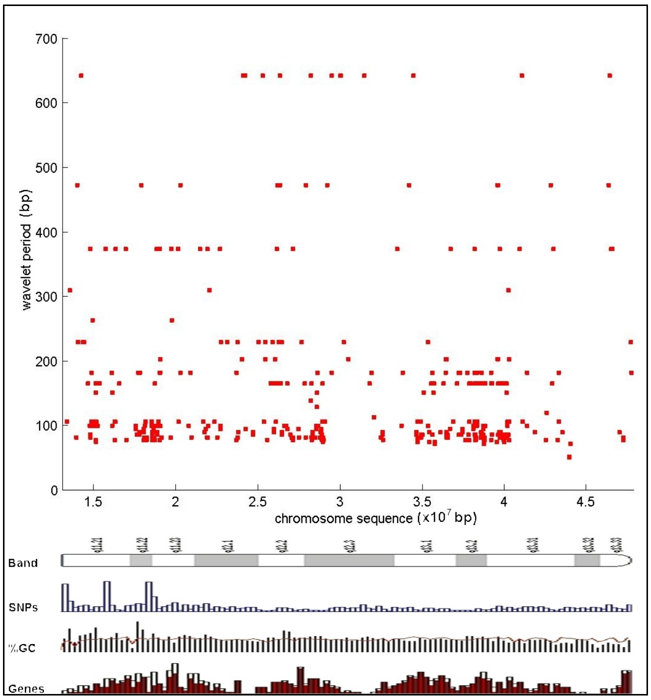
Such an examination of the flexibility landscapes of eukaryotic and prokaryotic DNA had been done before (Audit *et al.*, 2001; Audit *et al.*, 2002) utilising a different flexibility model (Goodsell & Dickerson, 1994). Using the wavelet technique to estimate a parameter called the Hurst exponent, Audit *et al* estimated that the occurrence of long range correlations of the order 10 – 200 bp was strong in several eukaryotic genomes including *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and human as well as in some of the viral genomes which infect them. The results obtained for *Saccharomyces cerevisiae* above, however, contradict this observation. They had also noted the lack of strongly periodic features in this range in bacterial genomes such as *Aquifex aeolicus* and *Bacillus subtilis*.

4.3.2 Whole chromosomal flexibility landscape in higher eukaryotic chromosomes

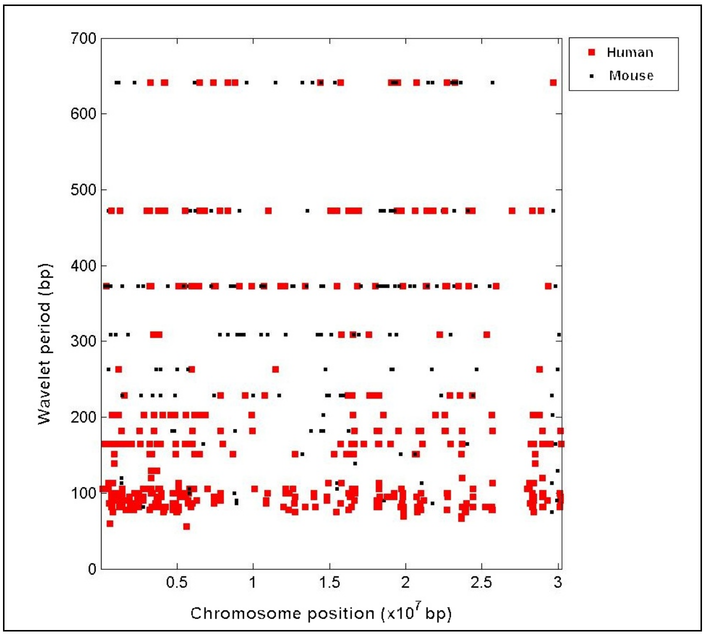
Figure 4.3: Continuous wavelet transforms of 3 large eukaryotic DNA contigs. These 2D plots were obtained by thresholding the wavelet co-efficients at 28 dB and plotting only those regions which were above this threshold. These results were obtained from transforming (a) 63 MB of human chromosome 20, (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).



(a)



(b)



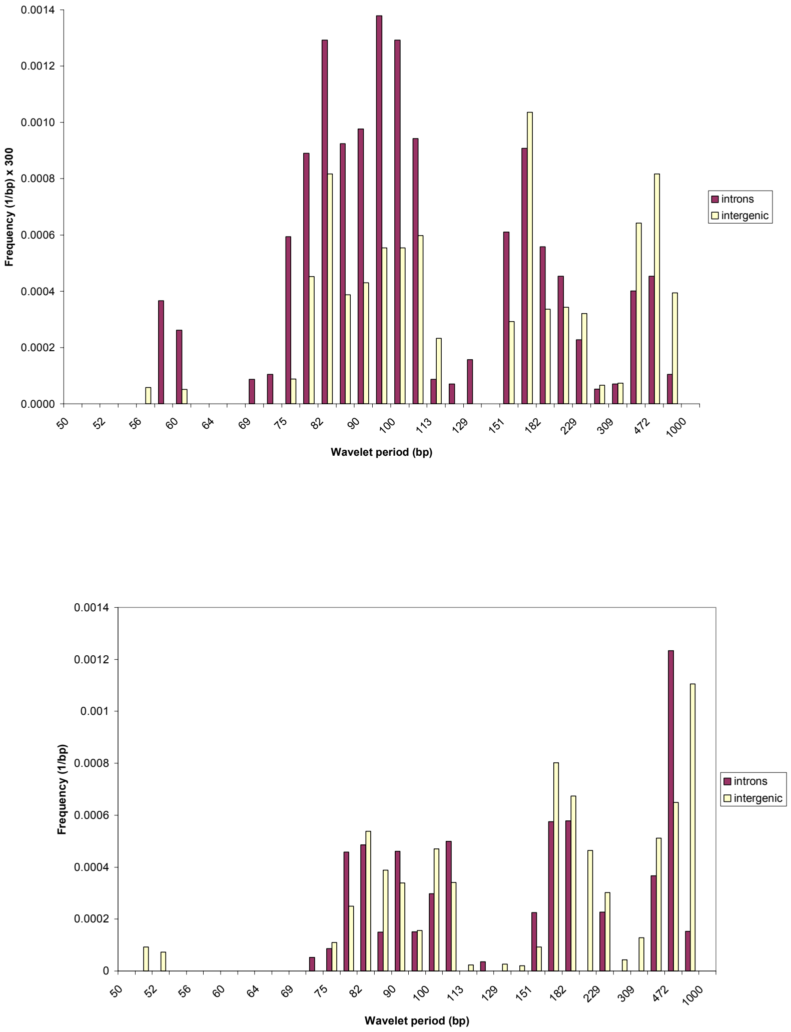
(c)

Figure 4.3(a),(b) shows the flexibility wavelet spectrum in relation to gene density in 2 human chromosomes. Distinct clumps of periodic flexibility patterns, in the range of 80–120 bp, were observed. In addition to these, there was a slightly less dense distribution of patterns observed in the range of 120–200 bp. The locations of these two "periodic classes" appeared to roughly coincide. Periodic patterns, above the 200 bp scale, occurred relatively sparsely.

The dense clumps of 80–120 bp patterns also appeared to roughly coincide with gene density (Figure 4.3(a),(b)). This closeness was apparent along the following co-ordinates:

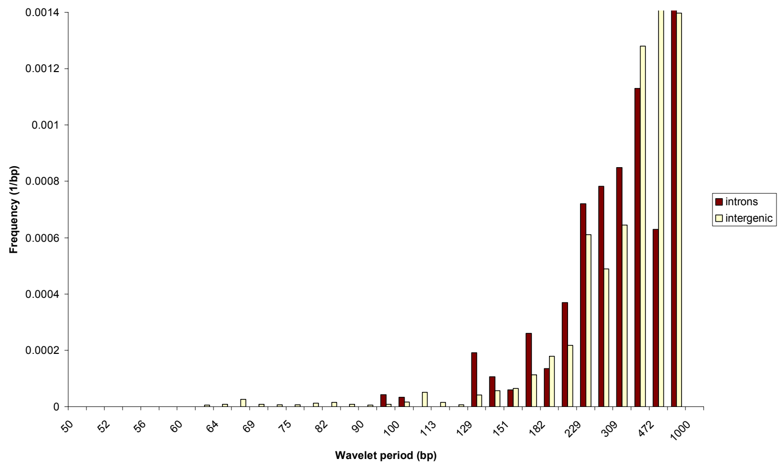
- Human chromosome 20 (Figure 4.3(a)): 1–7 MB; 30–38 MB; 40–50 MB
- Human chromosome 22 (Figure 4.3(b)): 17–20 MB; 25–30 MB; 35–40 MB

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.



(a)

(b)



(c)

To gain further insight into the distribution of these flexibility patterns, the probabilities of observing each of the sampled periods were compared for the 2 human chromosomes (Figure 4.4(a),(b)). In both graphs, there were 3 distinct peaks visible, which corresponded to the 3 aforementioned “classes” of periodic patterns.

4.3.3 Genomic features frequently associated with strongly periodic flexibility patterns

The occurrence of strongly periodic flexibility regions could have simply been the result of recoding previously known eukaryotic DNA sequence elements. Especially given the fact that the periodic features aligned closely with gene-dense regions, this observation required a closer inspection.

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.

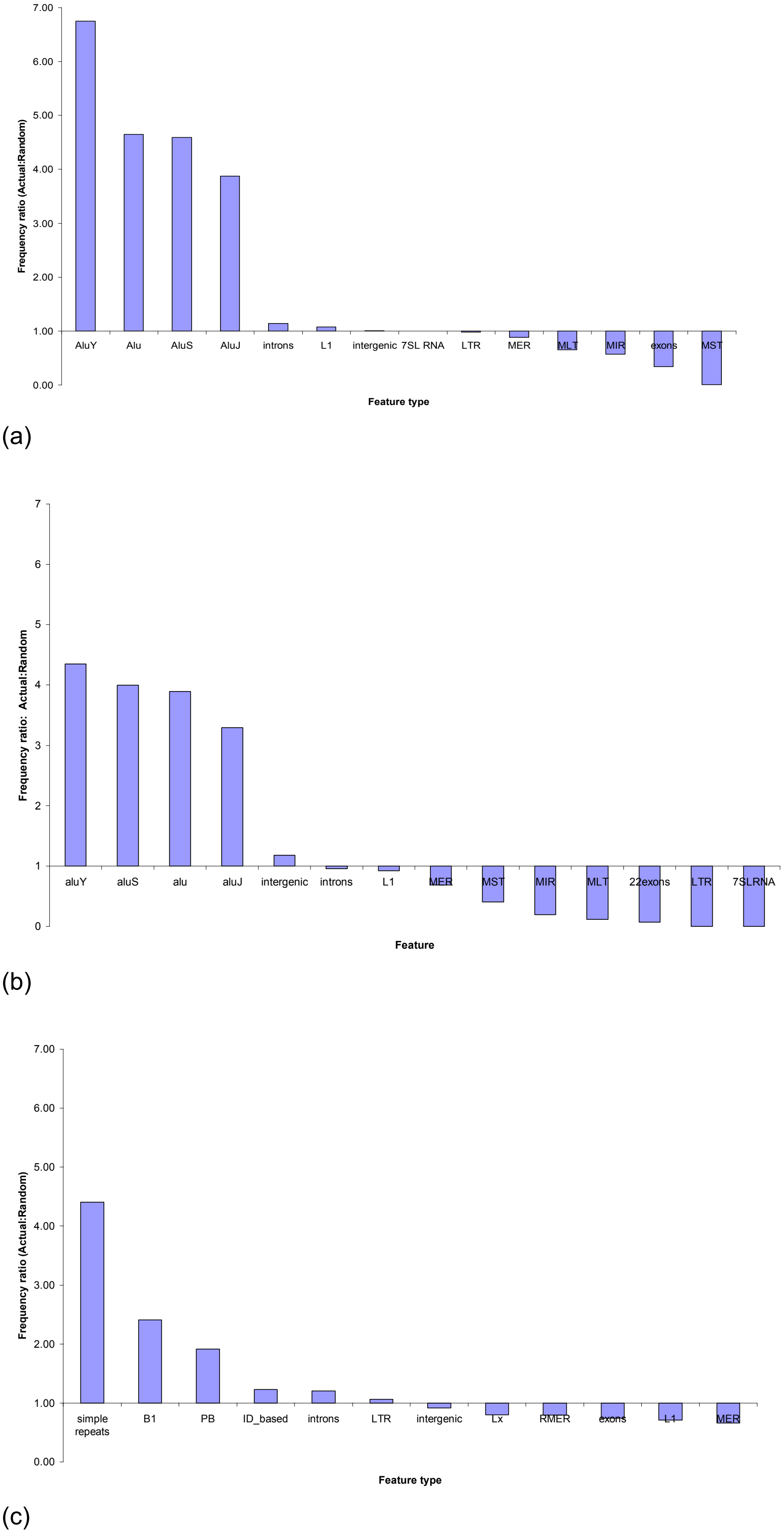
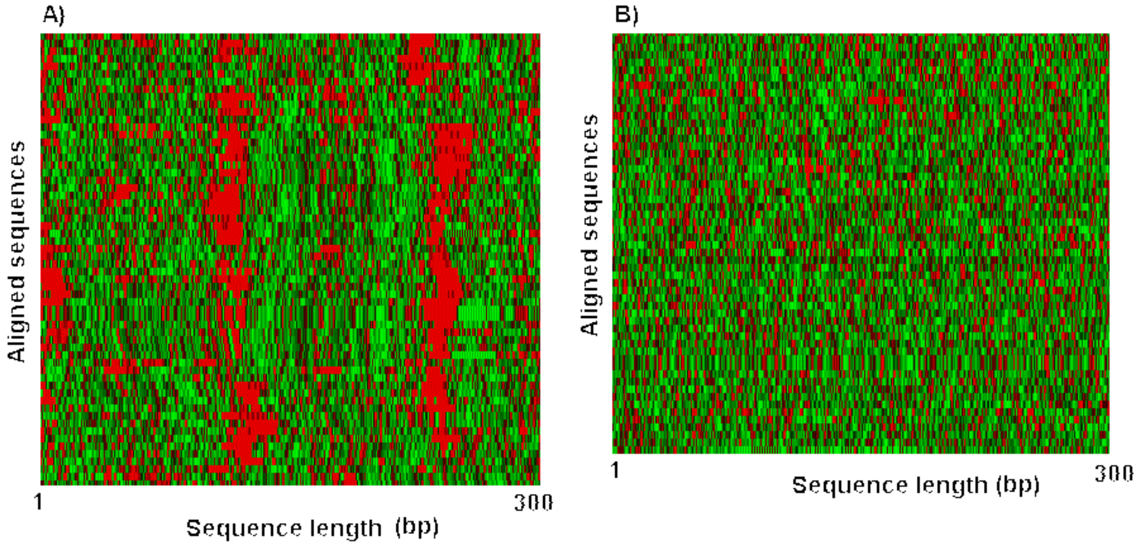


Figure 4.5(a,b) show features, which were frequently associated with the flexibility patterns of the 80–120 bp class in human. Clearly only the Alu repeat category was enriched: this repeat category was 4 times more frequently associated with these periodic flexibility patterns than expected. Aligning these sequences based on their flexibility (Figure 4.6) showed the linear arrangement of the periodic flexibility patterns that were detected. However, *RepeatMasker* analysis (Smit & Green, 1997) showed that the sequences themselves were mostly Alu repeats. So the observed patterns were in fact recoded Alu repeats (discussed in the next section; Section 4.3.4). Other notable observations from this analysis were that exons were not associated with these flexibility regions. This observation was consistent with other work, which suggests that long range correlations in eukaryotic DNA sequences exist only in non-coding DNA and not in coding sequences (Arneodo *et al.*, 1995; Arneodo *et al.*, 1998; Buldyrev *et al.*, 1998; Havlin *et al.*, 1999; Pattini L, 2001).

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.



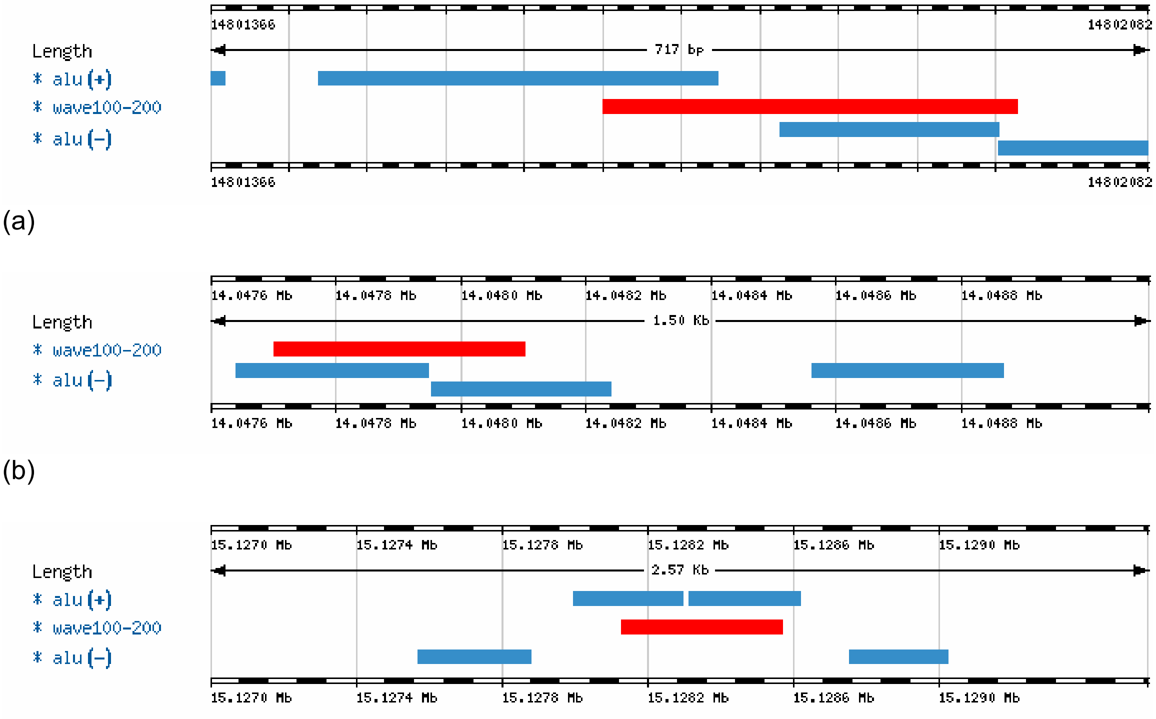
4.3.4 Why Alu repeats were frequently associated with periodic flexibility patterns

The results, discussed above (Section 4.3.3), showed that the flexibility patterns that were observed contained a large proportion of Alu repeats. The structure of Alu repeats themselves (Batzer *et al.*, 1996), as well as their recently outlined insertion patterns (HGSC, 2001), could explain why they had been detected using the wavelet transform.

- **Alu structure**

Firstly, Alu repeats are dimers of two roughly 100 bp-long 7SL-RNA derived fragments (Batzer *et al.*, 1996); however, the left and right monomers do not share any sequence similarity. Alu sequences also contain a poly [A] linker region separating the 2 RNA fragments and a poly [A] tail at the 3' end. The tetranucleotide parameter set that was used for converting DNA sequences into flexibility (Section 2.3.1) and indeed most of the other DNA flexibility parameter sets (Bolshoy *et al.*, 1991; Brukner *et al.*, 1995; Goodsell & Dickerson, 1994; Olson *et al.*, 1998; Packer *et al.*, 2000a; Packer *et al.*, 2000b) all model poly [A] motifs as being rigid in conformation (Section 1.4.1). Therefore, in the flexibility sequences, which were supplied as input to the wavelet algorithm, the 100 bp–spaced poly [A] motifs were becoming recoded as 100 bp-spaced rigid motifs. However, the wavelet transform only yields strong co-efficients when there are locally periodic patterns. A more detailed view of such locally periodic regions (>28dB co-efficient strength) showed that Alu repeats, which were in a very close arrangement, accounted for the regions of high flexibility (Figure 4.7). This would explain the periodic patterns that were observed. The fact that Alu repeats could represent the major class of poly [A] sequences in human was indicated in much earlier work (Lustig & Petes, 1984).

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.



(a)

(b)

(c)

- **Alu retention biases**

Alu repeats have been reported to be preferentially retained in GC rich regions (HGSC, 2001). Although it is thought that Alu insertion is more or less random, it appears that they tend to remain fixed in GC rich regions (Smit, 1999). It had also been reported that most of the preferred GC rich regions were mostly occupied by the older[19] AluS. Younger Alu repeats were reported to be found in similar proportions in AT rich regions as GC rich regions possibly due to saturation of the GC sites by the older Alus (HGSC, 2001). Since genes also display a bias towards GC rich regions in the genome, it was apparent why the locations of strong periodic flexibility patterns and gene dense regions appeared correlated (Section 4.3.3). These results could also

[19] AluY are estimated to be 20 million years old; the middle aged Alus (aluS) 35 million years old; the oldest Alus (aluJ) 50 million years old (Batzer & Deininger, 2002)

explain Arneodo's observation that long range correlations in human DNA were related to GC content (Arneodo *et al.*, 1998).

- **Percentages of repeat families picked up by the wavelet transform**

To estimate whether the patterns picked up by the wavelet transformation were representative of the whole population of Alu sequences, the percentages of different repeat families associated with periodic flexibility were compared (Table 4.1(a)). As seen in Table 4.1(a), only 2.06 - 2.67% of any of the Alu age categories were detected as strongly periodic flexibility patterns. However, roughly 82.07% of the regions classified as highly periodic were associated with Alu sequences of all ages. Therefore, although the wavelet transform itself was strongly biased towards picking up Alu sequences, the total Alu population which they had picked up represented only a small fraction of the total Alu population (presumably only the ones whose positions were very close to each other). L1 repeats were also represented as highly periodic flexibility regions (25% in Table 4.1(a)); this could once again be due to the wavelet transform picking up clustered Alu repeats, which are thought to rely on the endonuclease activity of L1 repeats for their own replication.

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).

(a) Human chromosome 20

	% repeat	proportion of total periodic flexibility
aluY	2.67	18.35
aluJ	2.17	31.01
alu	2.16	82.07
aluS	2.06	62.66
LTR	1.36	3.59
MIR	1.99	5.49
L1	1.13	25.95
MST	1.02	1.05
7SL RNA	1.02	0.42
MER	0.85	9.49
MLT	0.51	3.59

(b) Mouse chromosome 19

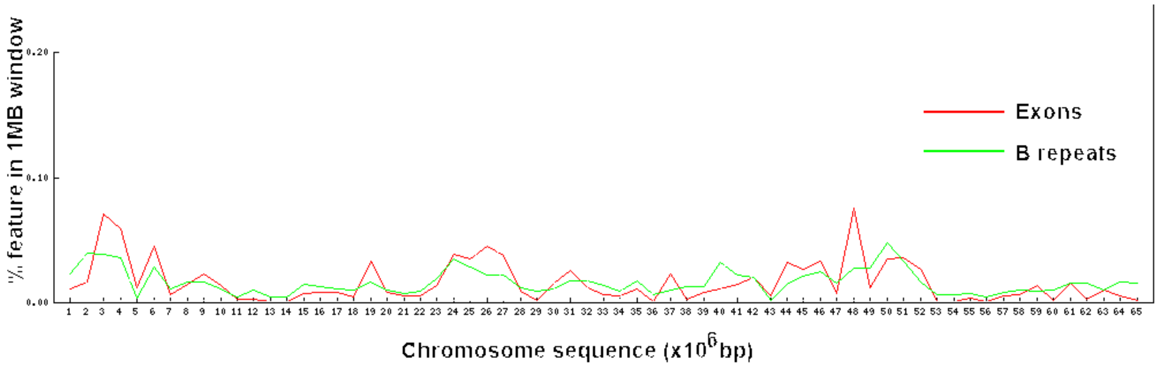
	% repeat	proportion of total periodic flexibility
Simple sequence repeats	1.98	63.01
MER	1.40	1.83
RMER	1.38	2.74
B-type	1.38	44.00
PB1	1.21	9.36
Lx	0.92	10.50
L1	0.89	11.87
ID-based	0.81	5.94

4.3.5 Conservation of periodic flexibility patterns in eukaryotic genomes

An important feature of a nucleosome positioning pattern may be that it is highly or at least moderately conserved between 2 species. To investigate this, a similar investigation, using wavelet transformation of flexibility sequences, was performed on mouse genomic contigs as was done for the human contig data. The data for the mouse genome was available only during the latter stages of this analysis; the data was, therefore, in its infancy and not as refined as the analysed human contigs. The only high quality alignment between human and mouse available at the time was the BRCA2 syntenic region (a 1.2 MB sequence alignment). Similar flexibility patterns were not observed between human and mouse in the BRCA2 syntenic region though (data not shown).

Figure 4.3(c) (page 4-124) shows the results of applying the wavelet transform on flexibility sequences in a syntenic region in human and mouse. The densities of periodic patterns that were observed in mouse were much lower compared to those in human. The locations of such patterns also did not show any kind of similarity with any corresponding locations in human. Furthermore, the probabilities of observing the different periodic patterns were not similar to what was seen in human (Figure 4.4(c), page 4-126). The peak periodicities in human could be grouped into 3 distinct classes but in mouse, there was only a single broad peak with a maximum of around 600 bp. These results indicated that the periodic flexibility patterns, which were seen in human (and which largely resulted from the clustering of Alu repeats), were not conserved in mouse.

Figure 4.8: Correlation of B repeat density and gene density in a region of mouse chromosome 2.



Genomic features, frequently associated with these periodic flexibility patterns, were found to be mainly simple repeats and B1 repeats in mouse (Figure 4.5(c); Table 4.1(b)). Whereas in mouse, simple repeats accounted for roughly 63% of the total periodic patterns represented (Table 4.1(b)), in human, simple repeats were only marginally picked up by the wavelet transform: these peaked at 50 bp periodicity and there were 2 such regions near the telomeric regions of both human chromosome 20 and 22 (data not shown). B1 repeats are the lineage specific SINE family in mouse, which are monomers of roughly 100 bp and similar in sequence to the left monomer of Alu repeats (Quentin, 1994). They also show a bias towards being retained in GC rich regions (alongside gene dense regions) (Figure 4.8), a pattern which was pointed out in the recent analysis of the mouse genome (IMGSC, 2002). Therefore, B1 repeats, although they show the same biased retention patterns as their human counterpart, do not represent the same contribution of periodic rigidity in mouse. This result is expected from the inherent structure of B1 repeats, which are monomers and do not share the poly [A] linker and poly [A] tail motifs of their human counterparts. Similar to the lack of periodic flexibility observed in human exons, mouse exons also lacked periodic flexibility behaviour (Figure 4.5(c), page 4-128).

4.3.6 Re-examination of the hypothesis of nucleosome translational positioning with respect to Alu repeats

The current research has raised a fundamental question: “Is it likely that Alu sequences direct the translational positioning of nucleosomes?”. Although a conclusive answer cannot be provided owing to the limits of the methodology outlined in this chapter, the evidence obtained using independent methods which link Alu repeats with nucleosome positioning can be considered. Secondly, there is also significant evidence in the literature that suggests that Alu sequences have acquired important functional roles in the human genome. Although these functional roles may not be directly related to nucleosome positioning themselves, the critical nature of the functions themselves may influence opinion on whether Alu sequences have developed a close enough symbiotic relationship in the host genome that could include effects such as nucleosome positioning.

- **Other evidence linking Alu sequences and nucleosome positioning**

The only recent computational work, which had connected Alu repeats and nucleosome positioning, was using the measurement of dinucleotide relative abundance distance discussed earlier (Section 1.9.4). This concluded that Alu repeats had the highest nucleosome formation potential but the nucleosome model used was itself questionable.

Fox *et al* (Fox, 1992) reported that large-scale isolation of genomic poly [A] clones (containing a large amount of Alu sequences) and reconstitution onto nucleosome core particles did not show significant aversion to nucleosome binding compared to random DNA fragments. This result was contradicted later by Englander (Englander *et al.*, 1993), who showed that Alu sequences showed rotational and translational nucleosome positioning capacity using an *in vitro* nucleosome reconstitution experiment. They showed that transcription in the *in vitro* DNA construct was blocked by nucleosomes; these nucleosomes were thought to be translationally positioned over the Alu elements. DNase I digestion indicated that the poly [A] linker region and poly [A] tails of the Alu sequences were probably directing this positioning. Englander *et al* later estimated that the left monomer of Alu repeats probably also had rotational positioning capacity (using DNase I digestion and software analysis) (Englander & Howard, 1995).

Englander *et al*'s results, particularly in (Englander & Howard, 1995), have interesting implications for the observations made in this chapter. Firstly, they estimated a rotational component in only the left monomer of the Alu sequences; this sequence is a homolog of B1 repeats in mouse (Quentin, 1994). This could suggest that clustering of Alu repeats and B1 repeats in GC rich regions ensures a significant quantity of rotational positioning signals in the upstream regions of genes in human and mouse respectively. This feature would not have been picked up in the current wavelet approach since the software they had used, for measuring curvature, was based on scanning for curved DNA; the wavelet tool used here was used to detect periodic flexibility of the scale order of 50–1000 bp. However, according to the signal which was picked up by the wavelet transform, namely the poly [A] motifs of the Alu repeats, it was highly unlikely that translational positioning was a conserved mechanism between human and mouse. The conclusion from linking the wavelet results from to Englander *et al*'s work is, therefore, an interesting one: rotational nucleosome positioning could be conserved between mouse and human but translational positioning is unlikely.

- **Alu repeats have taken on important functional roles in the cell**

One theory suggests that "Alu elements integrate randomly but those that are actively transcribed (and are therefore more likely to reside in G+C rich regions of the genome) are more likely to become fixed in the population " (Smit, 1999). This suggests that Alu repeats may play some functional roles due to their retention near gene dense regions (G+C regions). And indeed a number of recent experiments have shown that Alu sequences have adopted roles in important cellular functions.

Firstly, 1/3rd of CpG islands have been estimated to be contained within Alu repeats (Rubin *et al.*, 1994; Schmid, 1991). This could suggest that Alu repeats have an effect on the expression pattern of downstream genes due to mutations that alter the CpG methylation patterns. Alus are also known to directly insert into coding sequences and 0.1% of all genetic disorders are known to be caused by such unfavourable insertions (Deininger & Batzer, 1999).

In many organisms, SINE expression levels also increase under conditions of stress (Chu *et al.*, 1998; Li *et al.*, 1999; Liu *et al.*, 1995). Under such conditions, SINE RNA transcript has been reported to bind a specific protein inhibitor, and thereby block its activity. Therefore, under conditions of stress, Alu repeats may be specifically induced to upregulate the expression of many genes. This increase in Alu expression has also been linked with a rise in DNase I hypersensitivity in chromatin indicating possible Alu-mediated reshuffling of chromatin arrangement (Kim *et al.*, 2001).

Some recent work has provided the first indications of common functional roles between Alu and B1 repeats in human and mouse respectively (Zhou *et al.*, 2000; Zhou *et al.*, 2002). Zhou *et al* showed that the strongly evolutionarily conserved *Pax6* transcription factor, which is critical in the development of the eye, pancreas and central nervous system, exhibits more than 1 kind of preferred binding site in both human and mouse. However, the transcription factor binding sites included several Alu repeats in human and B1 repeats in mouse. An interesting twist was that the binding sites in the 2 lineage-specific SINE families did not share any sequence similarity! This suggests that the evolution of *PAX6* function may have been aided or merely influenced by simultaneous SINE evolution.

4.4 Conclusion

The wavelet transformation of flexibility sequences showed that the clustering of Alu repeats resulted in locally periodic rigidity patterns. On account of such clustering, two classes of periodicity could be seen: 80–120 bp and 120-200 bp respectively. These were observed near gene dense regions, which was expected from the biased retention property of Alu repeats in GC rich regions. Similar flexibility patterns were not seen in analysis of mouse contigs. SINE repeats may have simultaneously evolved to serve some common functions in their respective host genomes. But according to the results presented in this chapter, it is unlikely that nucleosome translational positioning is one such conserved function.

5 Modelling DNA Sequence Motifs from Known Nucleosome Datasets

5.1 Introduction

Rotational positioning signals have been described for both of the nucleosome datasets available so far but it has not yet been clarified what proportion of the sequences in either dataset exhibit this property (Section 1.11.3). This formed the need to analyse these sequence datasets using a classification-based approach. The approach would be to partition the dataset into 2 parts: a training set and a test set. The aim would be to learn models from the training set and analyse them on the test set to understand if the models truly represented the respective nucleosome datasets. A powerful classification software for numerical datasets, *Eponine* (Down & Hubbard, 2002), was available to carry out this procedure.

A similarly motivated approach was described earlier where a dinucleotide-based system was used to classify mouse nucleosome sequences from mouse non-nucleosome sequences (Section 1.9.4). However, as mentioned earlier, the positive dataset, used in that study, contained mainly centromeric repeats and were, therefore, unlikely to represent the vast majority of nucleosome-forming DNA in genomic sequences (centromeric nucleosomes exhibit specialised structures in eukaryotes (Smith, 2002)).

5.1.1 The *Eponine* Tool

Eponine was developed by Thomas Down and its initial and major application has been in modelling transcription start sites (Down & Hubbard, 2002); this yielded a model with an estimated prediction specificity of >70%. The software uses a Bayesian machine learning method to learn complex models comprised of one or more DNA weight matrices. DNA weight matrices are "weighted" short, un-gapped sequence motifs, which contain a series of column distributions over the DNA alphabet. An *Eponine* model is a linear combination of the weights of these matrices. These weights have to be trained iteratively to optimise their values.

Eponine uses an implementation of the relevance vector machine (RVM) technique for training the weight parameters. It takes as argument (a) a positive dataset containing the feature of interest and (b) a negative dataset which lacks the feature of interest. The RVM algorithm works by initializing a model with a set of suggested weight matrices and iteratively selecting only those subsets which are most "relevant" in classifying the positive training dataset from the negative training dataset.

Eponine has the option of learning 2 kinds of models: *"anchored"* or *"unanchored"*. In an anchored model, each DNA weight matrix is further compounded with a probability distribution over distance; this distribution describes the distance relative to a reference or *"anchor point"* in the model (for example, Figure 5.3). Conversely, *"unanchored"* models do not have distance constraints.

This software tool was an appealing option to learn models representing important sequence motifs in the 2 available nucleosome datasets (Section 1.8). Particularly, anchored models, with their anchor points set to the approximate mid-points of the sequences, could be useful to learn rotational positioning motifs, which are expected to be symmetrical about the midpoints of the sequences (Section 1.9.2).

However, it could also be expected that weight matrices, additional to the previously described rotational positioning motifs, could be learnt. For example, multiple expansions of the [CTG] motif was shown to bind nucleosomes 9 times more strongly than an intrinsically curved DNA (Wang & Griffith, 1995); this same motif did not show preferential rotational positioning in the analysis of the chicken sequences (Satchwell *et al.*, 1986). Therefore, it was not essential for the learnt weight matrices to represent the rotational positioning motif which has been described before; the important thing was that the learnt weight matrices should represent properties of the dataset which could help to classify its sequence members from other DNA sequences. Also, it was reported recently that the signals which affected translational positioning were not the same as the signals which affected rotational positioning in an artificial DNA sequence (Negri *et al.*, 2001). Therefore, there was potential for learning both rotational and translational positioning motifs using *Eponine*.

5.2 Methods

5.2.1 Selection of positive and negative datasets

Positive datasets were quite easily defined for the nucleosome classification problem. These were of course the chicken nucleosome dataset and Levitsky *et al*'s nucleosome dataset (Section 1.8).

In Levitsky *et al*'s data, however, 16 of the mouse sequences differed from each other by only a few bases; these close variants were removed (Section 1.8.2). Furthermore, sequences less than 144 bp in length in this dataset were not considered; this was because a model roughly the size of core DNA was desired. This resulted in a final dataset size of 160 sequences.

Finding an appropriate negative training set was a much more difficult problem. This was because an appropriate collection of nucleosome-repelling sequences was not available. Therefore, initial studies were performed using randomized versions of the 2 datasets as negative data.

However, for the positive chicken nucleosome data, a better negative set was to use background chicken genomic DNA. Two chicken genomic clones were available for this purpose (Section 1.8.1). Genomic sequences for the negative datasets were obtained by randomly selecting 146 bp length fragments from these 2 clones. An assurance of randomly selecting genomic fragments as negative data was that rotational positioning signals were unlikely to be present symmetrically about the centre of the sequences as they have been described previously for the positive nucleosome data (Section 1.4.2).

Table 5.1: Summary of classification categories used.

POSITIVE DATA	NEGATIVE DATA
177 sequences of Levitsky et al's data	Levitsky et al's data randomized
177 chicken nucleosome sequences	Chicken nucleosome sequences randomized
177 chicken nucleosome sequences	Chicken background genomic sequences

Therefore, 3 kinds of classification categories were finally used (Table 5.1). Both kinds of training, anchored and unanchored, were performed on each of these classification categories. For anchored training, the models were anchored at sequence co-ordinate 73, which was close to the midpoint of most sequences. Sequences, which were much longer than 146 bp (Section 1.8.2), had ambiguous midpoints and were treated differently (discussed subsequently; Section 5.2.3).

Roughly 20-25 training attempts were made on each classification category to assess whether consistent models could be learnt. Each training run involved randomly partitioning 25 sequences from both the positive and negative datasets to form respective "jack-knifed" test sets. 15,000 cycles of training were performed per training run. Models were dumped every 500 cycles and their predictive power assessed on the test sets (discussed below).

5.2.2 Estimation of a model's predictive power

The accuracy and coverage of the dumped models were calculated to assess how well they could correctly classify the positive test samples from the negative test samples. Accuracy was calculated as the total number of correct predictions over the total number of predictions made. Coverage was calculated as the total number of correct predictions over the total number of true data samples (25 such samples in this case). The output was analysed using ROC (receiver operating characteristic) curves, for example in Figure 5.1; the points on the ROC curve were obtained using different scoring thresholds in *Eponine*. Only models that scored with >80% accuracy and

>50% coverage in the test set were considered useful representatives of a nucleosome dataset and were analysed further.

5.2.3 A modified approach to find rotational positioning motifs

In the initial training attempts using anchored training, an anchor point approximating the midpoints of the sequences was used. This anchor point, 73, was reasonable for the chicken data as the sequence lengths did not vary that greatly: 142 to 149 bp with an average length of 145 (±1.5) bp. However, many of the sequences in Levitsky *et al*'s dataset were around 200 bp and had ambiguous midpoints. Therefore, to enhance the chances of learning rotational positioning signals, which are thought to occur symmetrically about the mid-point of core DNA (Section 1.4.2), the following modified training approach was also tried: After each round of training, each of the training sequences was shifted a few times within a range of a few bps. This led to a set of 'offset' sequences for each training sequence. For each round of training, each of the offset versions of a training sequence was scored with *Eponine* and the highest scoring offset sequence stored for the next round of training. Offset values of 6-20 bp were tried.

5.2.4 Model prediction using *Eponine*

Models, which were trained from chicken nucleosome sequences, were used to predict nucleosome sites in a 92,863 bp chicken locus (Genbank accession ID: AL023516). The *Eponine* scoring threshold, which yielded the best accuracy and best coverage (a point approximating to the middle of the ROC curve) for a respective model, was used. The scoring threshold, which gave the least number of false predictions was also used. For a cross-species comparison, the BLASTN alignment tool (Altschul *et al.*, 1990) was used to find the homolog of this locus in the mouse genome.

Predictions were made on this homologous segment separately and compared to the predictions in the chicken locus.

5.2.5 Principal components analysis of trinucleotide background distributions

The background trinucleotide distributions of different eukaryotic genomes and the 2 mapped nucleosome datasets were also investigated. The aim was to see if either of the nucleosome background distributions could be classed along with the background distributions of other eukaryotic genomes. To investigate this, principal components analysis was performed on the relative frequencies of the 64 trinucleotides in the different genomic samples. As a negative control, the positions of the background distributions of *E. coli* and a human codon table were also plotted along the principal component axes.

5.3 Results and Discussion

5.3.1 Unanchored training results

Out of 25 unachored training attempts on each of the 3 classification categories (Table 5.1), only 2 models with accuracy and coverage greater than the desired thresholds (80% and 50% respectively) were learnt. Both of these models were learnt from different training runs on Levitsky *et al*'s data (Table 5.2). As seen in Figure 5.1, the midpoint of the ROC curve for both models was at 85% and 60% respectively using the jack-knife test.

Table 5.2: Unanchored models learnt using Levitsky *et al*'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.

MOTIFS			Weight
ttatagt	gaacaat	tacgcgg	-5.70
ttacccgtg	tacgcg		-4.64
tttacgatcg	agtgtgtct	ctgacta	-2.92
aggatcc	tgctcgc		-0.48
ctcaa	atcaa		1.80
ctggaaac	tggaa	gtgatt	2.66
atgcagc	gcatcat	aaggtc	5.00

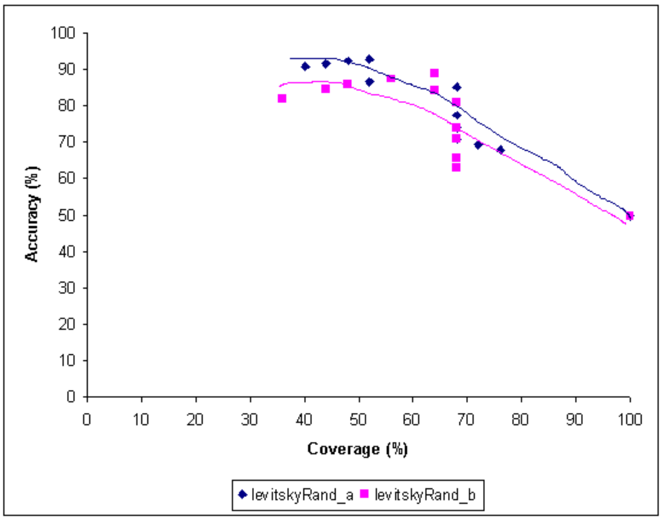
(a) Model levitskyRand_a

MOTIFS			WEIGHT
ctagg	agagtc		-7.83
ttatgcg	ccgtgg	ggtagggt	-5.49
atgtaagg	aacga	acagt	-4.93
acggg	acggg		-1.32
acaaag	agcaaag		2.33
ttcctaaatt	gcatct		3.06
ttgaggag	gttggg		3.76

(b) Model levitskyRand_b

It was not apparent why good predictive models could not be learnt using the unanchored approach on the chicken data. Only 2 out of 25 runs learnt models with good predictive power from the Levitsky data. However, the 2 models did not show any obvious similarity in the weight matrices they had learnt (Table 5.2).

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).



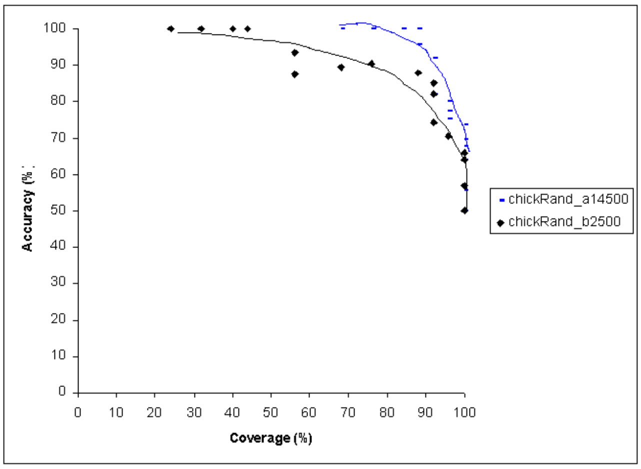
However, it was observed that the models had learnt multiple CpG motifs in the negatively-weighted matrices; these are highlighted yellow in Table 5.2. An important fact known about long runs of CpG motifs is that they occur very rarely in eukaryotic genomes (Cooper & Gerber-Huber, 1985; Sved & Bird, 1990). Therefore, the fact that randomized sequences were being used as negative training data explained why CpG appeared as negative weight matrices in the learnt models. The predictive power of the models was biased by the negatively-weighted CpG-containing matrices since CpG appears rarely in the positive nucleosome test set but has a random probability of occurrence in the negative test set. The conclusion from these results was, therefore, that using randomized sequences as negative data either for testing or training was unsuitable. It would only learn motifs which represented the background sequence composition of the positive dataset rather than any significant weight-matrices. The problem was that a more appropriate negative dataset for the Levitsky data was not available. This ruled out analysis of the

Levitsky nucleosome dataset any further. For the chicken nucleosome data, using a negative dataset of background chicken genomic sequences was more suitable.

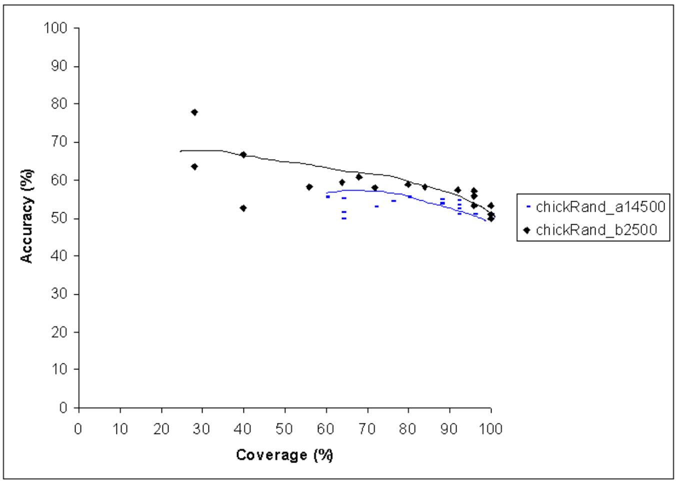
5.3.2 Anchored training results using randomized chicken nucleosome sequences as negative data

Although the use of randomized sequences was considered inappropriate, they had already been used as negative data for anchored training from the chicken nucleosome dataset. This yielded some interesting observations about the background distribution of the chicken nucleosome sequences, which could be linked to the cyclical HMM results (Chapter 3).

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.



(a)

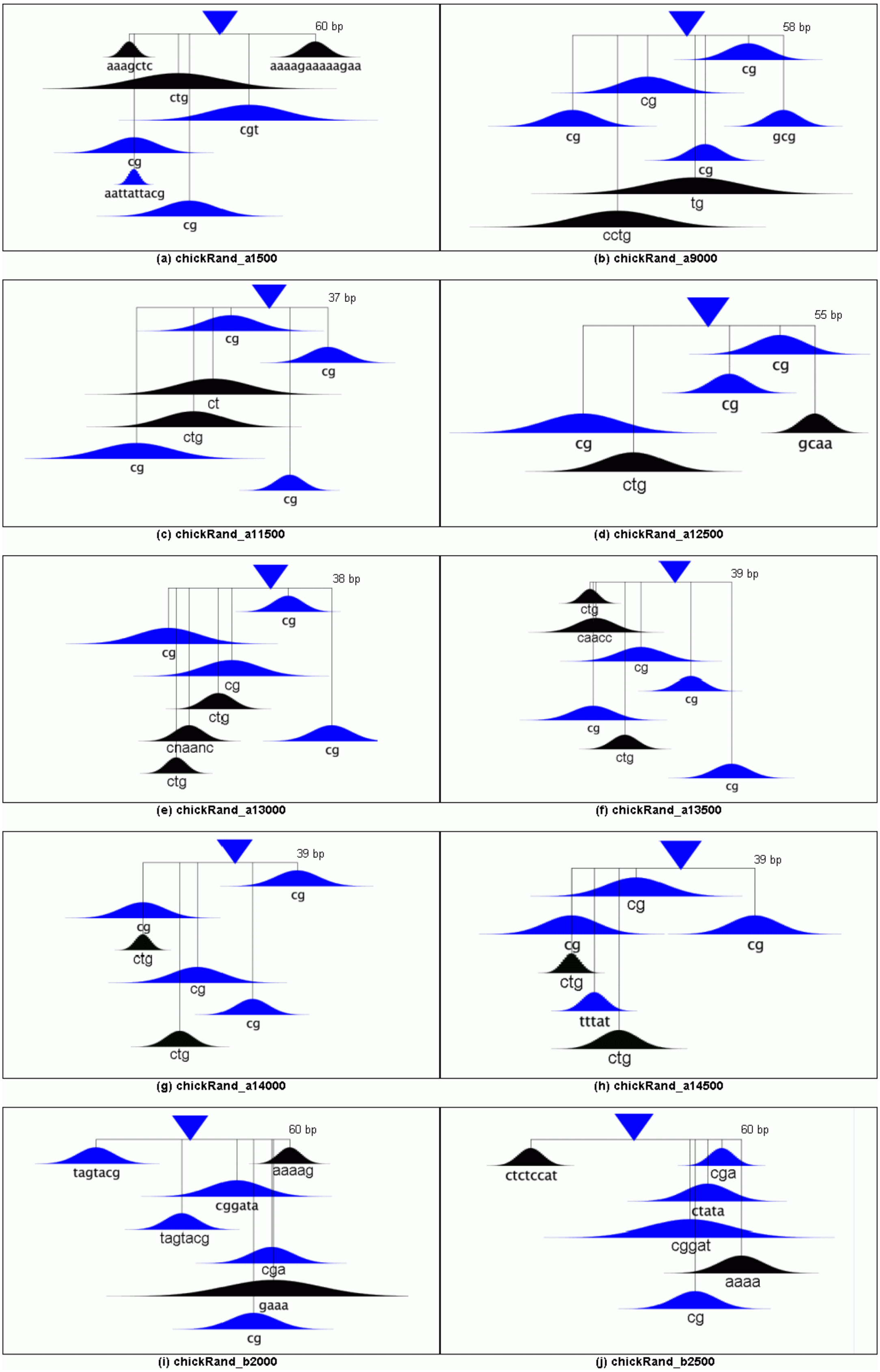


(b)

The results of this were 10 models having good predictive power in the jack-knife test (Figure 5.2(a)). The midpoints of the ROC curves were around 88% accuracy and 88% coverage respectively. However, the models were not accurate in correctly classifying the chicken nucleosome DNA from background chicken genomic DNA (Figure 5.2(b)); in this test, the accuracy of these models were <80%, which was less than the threshold being used for selecting good predictive models.

Most of the models learnt positively-weighted [CTG] motifs (Figure 5.3), the pattern which had been seen most often using the cyclical HMM learning in human genomic sequences; this outcome is discussed in the next section, 5.3.3. The models were also enriched in negatively-weighted CpG motifs which, as mentioned in the previous section, are a consequence of using randomized sequences as negative data (Figure 5.3). 8 of these models were dumped from different cycles of 1 training attempt (Figure 5.3(a)-(h)) whereas 2 were from cycles of another training attempt (Figure 5.3(i)-(j)). A total of 25 training attempts were made. The positively-weighted motifs learnt in the 2 successful training attempts did not appear similar.

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 *training run a* and models (i)-(j) were learnt in different cycles from *training run b*. The inverted blue triangle represents the "anchor point".



(a) chickRand_a1500

(b) chickRand_a9000

(c) chickRand_a11500

(d) chickRand_a12500

(e) chickRand_a13000

(f) chickRand_a13500

(g) chickRand_a14000

(h) chickRand_a14500

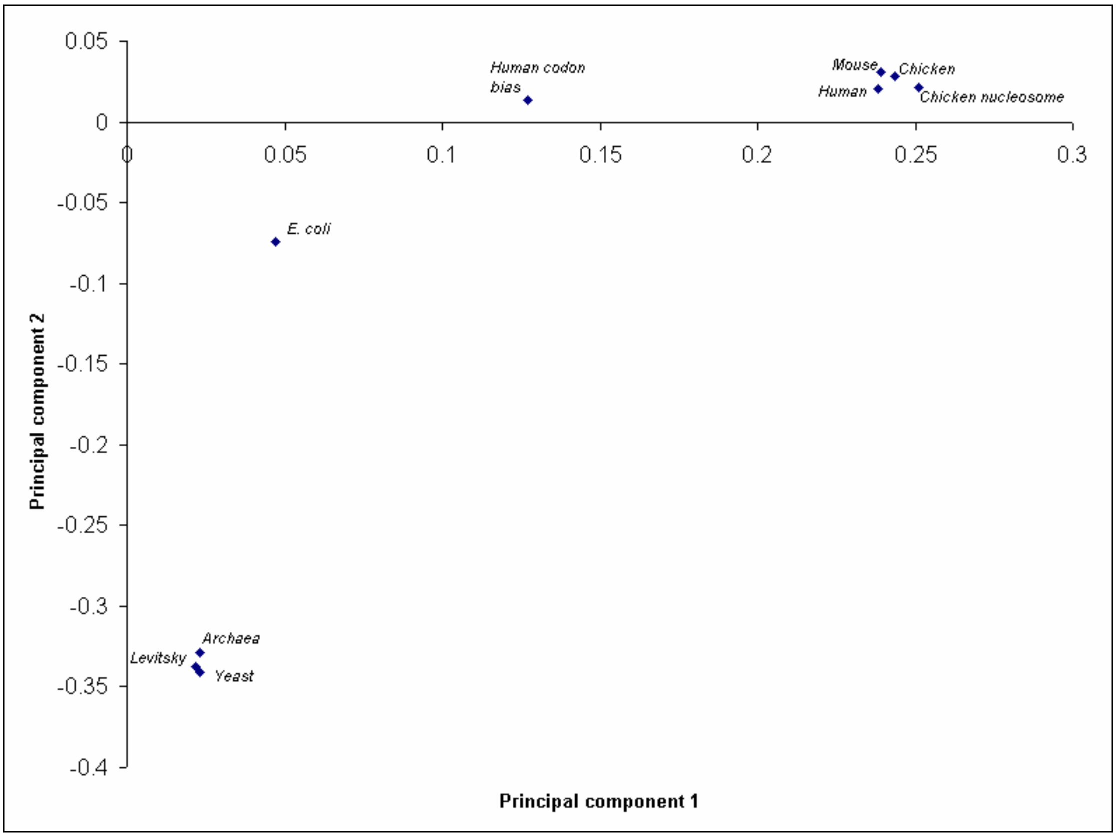
(i) chickRand_b2000

(j) chickRand_b2500

5.3.3 Could the background trinucleotide distribution in different genomes affect nucleosome positioning?

The motif [CTG], which is also a member of the ambiguity set [CWG], was learnt using *Eponine* training from the chicken data and was also learnt using cyclical HMM training from human sequence data (Chapter 3). In addition, the labelling of the [CWG]-learnt HMM models was seen to be related to the background density of [CWG] in human (Sections 3.3.4, 3.3.8). Therefore, it was interesting to assess whether the background trinucleotide distributions were similar amongst different eukaryotic organisms and the nucleosome datasets (Figure 5.4).

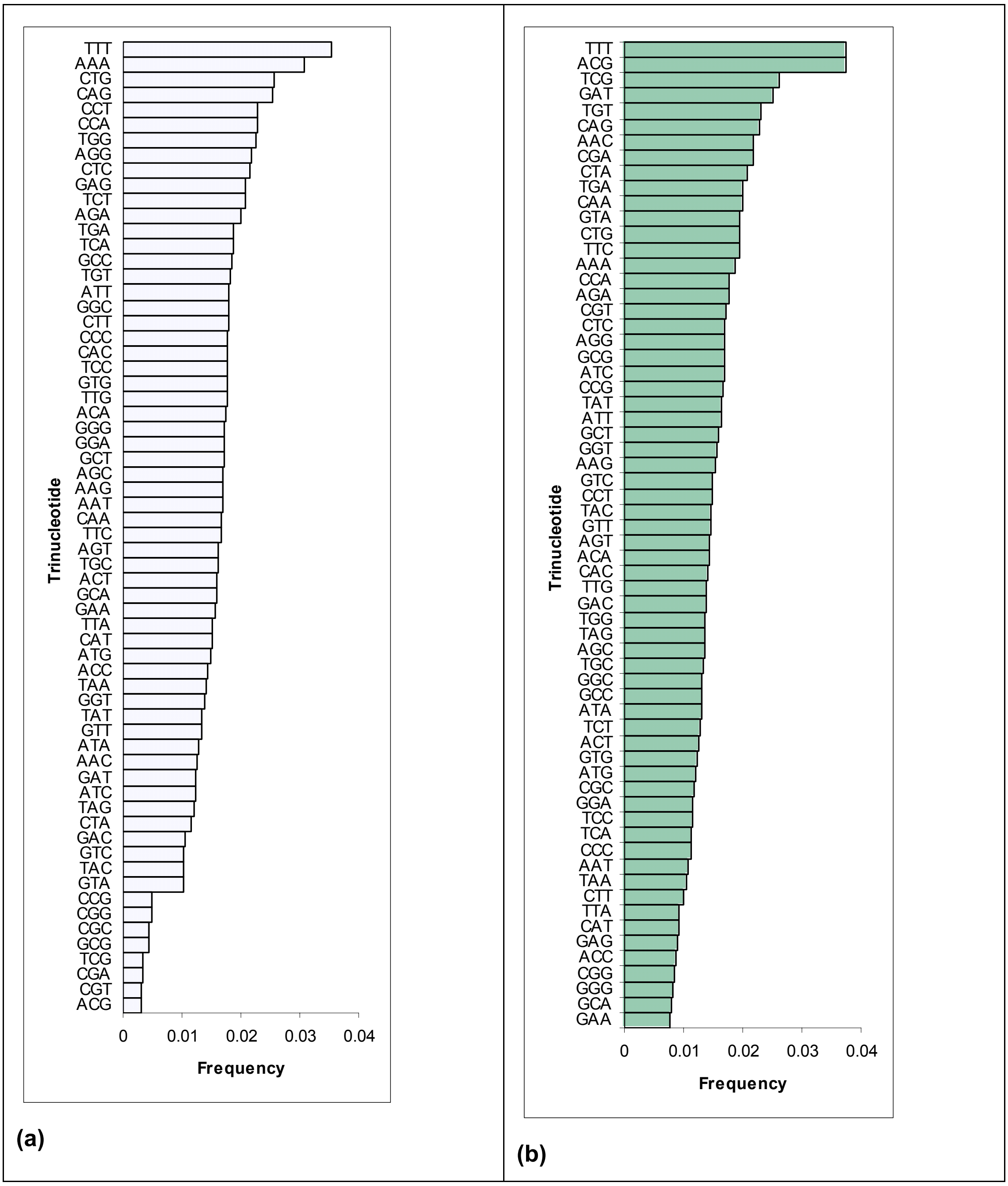
Figure 5.4: Principal components analysis of the background trinucleotide distributions of different genomes and the 2 nucleosome datasets.



The higher eukaryotes, human, mouse, and chicken were seen to have similar background trinucleotide distributions (Figure 5.4, Figure 5.5(a)); the correlation co-efficient between the human and mouse distributions was 0.82. A similar distribution was apparent in the chicken nucleosome dataset. As seen in Figure 5.5(a), the most frequent trinucleotides in human were [AAA/TTT] followed by [CWG] (note it was earlier observed that in mouse, [AAA/TTT] was most frequent but not [CWG]; Section 3.3.4). The human and mouse background distributions do not differ significantly about their means as a two-sample t-test at the significance level of 0.05 showed that the means were equal.

The location of a human codon bias table was also plotted on the principal components scale (Figure 5.4); this showed that the plotted trinucleotide background distributions did not represent a contribution of codon bias. In the same table, the co-efficients against the *E. coli* data shows that none of the eukaryotic backgrounds were similar to the prokaryotic negative control.

Figure 5.5: Background trinucleotide composition in descending order in (a) the human genome and (b) the Levitsky nucleosome data.



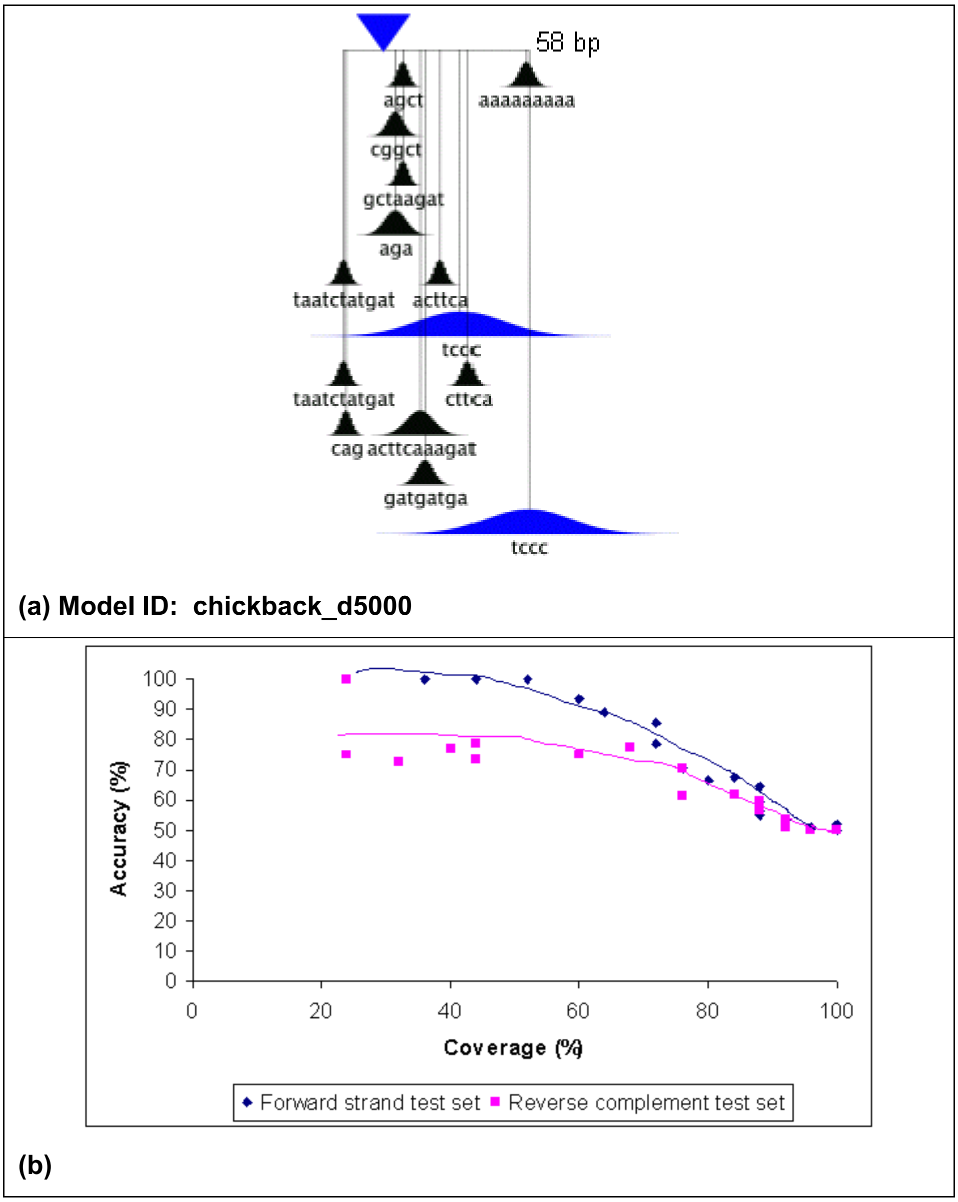
The background trinucleotide distribution for the Levitsky data was quite far from the distribution of the higher organisms along the principal components axes (Figure 5.4, Figure 5.5(b)); the correlation co-efficient between the human and Levitsky distributions was 0.02. The means of the distributions did not differ between the human and Levitsky data as a 2-sample t-test at the significance level of 0.05 showed the means to be the same. On the principal components axes, this distribution was much closer to the lower eukaryotes, archaea and yeast, and contained a high proportion of [TTT] and [ACG] (Figure 5.5(b)). The similarity to archaea and yeast could be expected as both these organisms were represented in the Levitsky data (Section 1.8.2).

Taken together, the 2 kinds of background distributions (Figure 5.5) suggest that if the background trinucleotide distribution is important for nucleosome positioning, then the pattern is maintained differently between higher eukaryotic organisms and lower eukaryotic organisms. For certain higher organisms, both [AAA/TTT] and [CWG] may play a role in nucleosome positioning (the relevance of either motif in nucleosome positioning was discussed previously in Sections 1.4, 1.5.1 and 1.5.2). On the other hand, in lower organisms such as yeast and archaea, only [AAA/TTT] may be important for nucleosome positioning as has been suggested from previous studies of their genomic sequences (Bailey *et al.*, 2000; Widom, 1996). The background trinucleotide distributions may also account for the differences in rotational positioning analysis of the 2 nucleosome datasets. Specifically, in the chicken data, [GC/GC] was seen to occur in anti-phase with [AA/TT] whereas [TT] was seen to occur in anti-phase with [AA] in the Levitsky data (Section 1.9.2).

5.3.4 Anchored training results using background chicken genomic DNA as negative data

Using background chicken genomic sequences as negative data was perhaps the best available option of finding motifs that separated the chicken nucleosome sequences from their genomic background. Unfortunately, the alternate training method, designed to find symmetric rotational-positioning weight matrices about the sequence midpoints (Section 5.2.3), did not yield good predictive models using the jack-knife test (data not shown). The rotational positioning motifs previously described were perhaps too weak to be picked up by *Eponine*.

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.



(a) Model ID: chickback_d5000

(b)

Only one model with good predictive power was learnt from 25 training attempts using the regular training method (Figure 5.6(a)). The midpoint of this model's ROC curve was around 85% accuracy and 75% coverage; also around 40% coverage, there were no false predictions ("Forward strand test set" in Figure 5.6(b)).

A separate test was performed to see if this model could classify positive sequences from the Levitsky data from negative chicken genomic sequences: it failed to do so (data not shown). As from the observations of the trinucleotide backgrounds, it was again clear that the chicken nucleosome dataset and the Levitsky data were quite different.

One notable observation about the model was that it had learnt a poly [A] weight matrix +58 bp from the anchor point. This poly [A] tail could be the same signal which was mentioned before in the initial assessment of the chicken nucleosome sequences (Drew & Travers, 1985; Satchwell *et al.*, 1986); it had been suggested that poly [A] tails were present towards the ends of the sequences and could possibly help to direct nucleosome translational positioning. The test sequences were later examined by eye to assess if they had poly [A] tails at their right ends, which could have biased the ROC analysis. Such a bias was not observed in the test sequences.

Another analysis was performed to see if such a poly [A] motif appeared symmetrically towards both ends of the sequences. This procedure involved reverse-complementing the test set and testing it (Figure 5.6(b)). The results showed that at 20% coverage, there were no false positives. This was a much lower accuracy than the forward strand test set (40%) suggesting that poly [A] tails did not occur symmetrically in these nucleosome sequences. This observation was interesting as it suggested that there might be some bias to having poly [A] tails at one end rather than at both.

However, the positions of each of the weight matrices were themselves not placed symmetrically or repetitively about the anchor point. Therefore, rotational positioning motifs were not featured in this model. The other positive weight matrices in the model, with the exclusion of one [CAG] motif (-15 bp from the anchor point), were not consistent with any other kinds of motifs that have been reported previously to be involved in nucleosome positioning. This approach was therefore made difficult, mainly due to the limited number of sequences available. However, it did show that a good model could be learnt.

- **Prediction analysis**

Although the weight matrices in this model did not represent a rotational positioning motif, it did have good predictive power in the jack-knife test against a reasonable negative test set. Therefore, it was used to make some comparative predictions on a 192 kbp-long chicken genomic locus and its homologous region in mouse (Figure 5.7;Figure 5.8). The BLASTN search found a 5,000 bp alignment in mouse chromosome 17 (Figure 5.7); however, upon examining the annotations, it was apparent that the aligning pairs were all coding DNA. The evolutionary distance between mouse and chicken, estimated to be 200 Myr[20], was probably too great for any potential regulatory regions to be found using BLASTN. This was unfortunate as potential regulatory regions could not be assessed. The predictions, within the coding DNA, did not appear to be conserved (Figure 5.8).

[20] Compare with 80 Myr between mouse and human (Burt *et al.*, 1999)

Figure 5.7: Locations of high-scoring BLAST segment pairs between the GGB locus in chicken and in mouse.

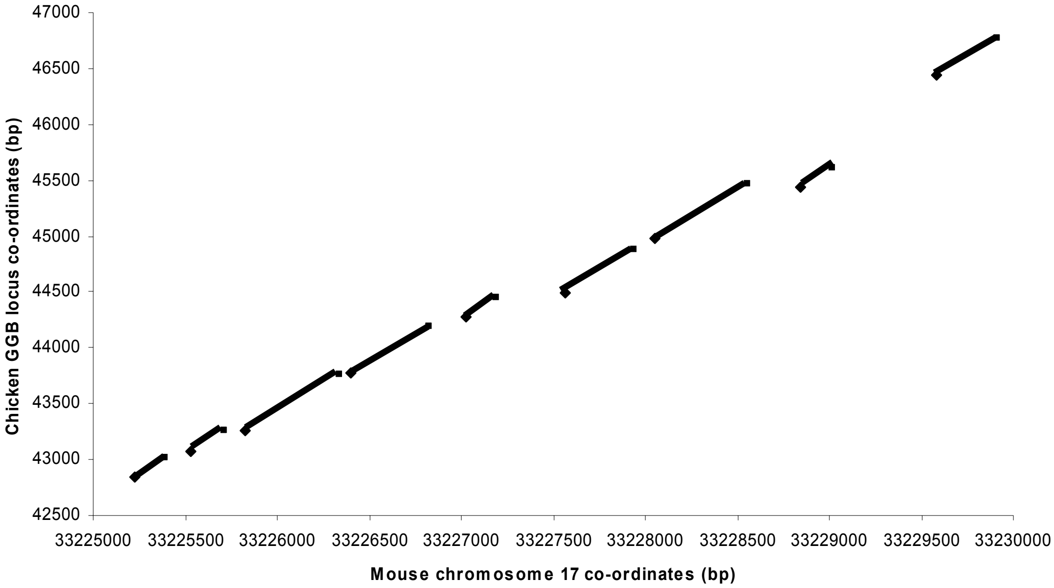
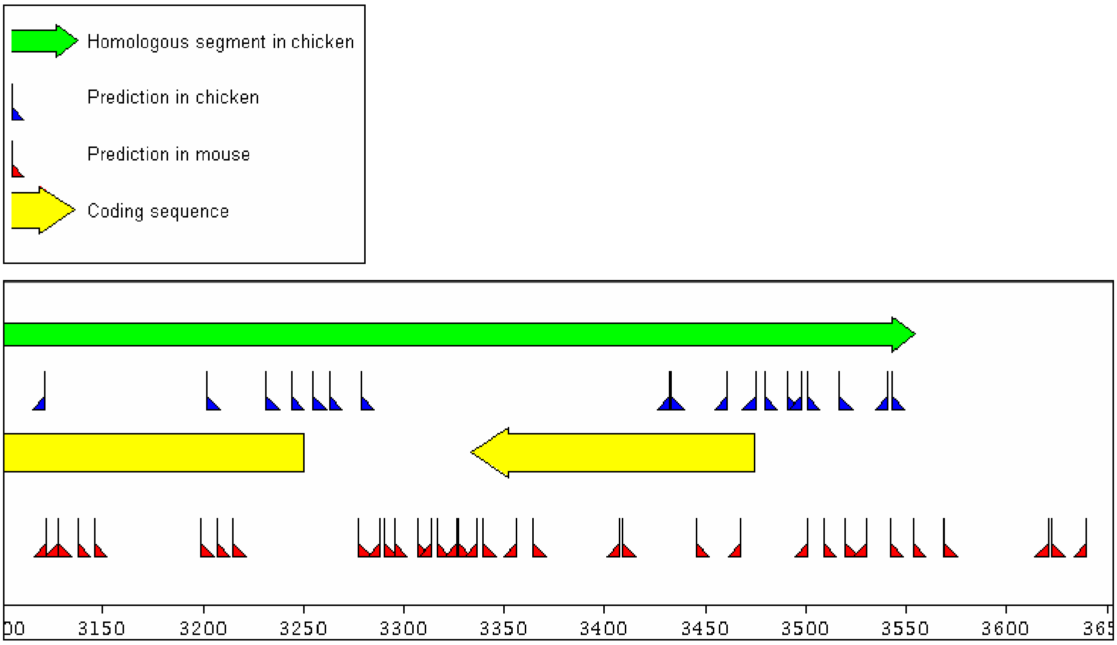


Figure 5.8: Prediction using model *chickBack_d5000* (Figure 5.6(a)) on the chicken GGB locus and homologous regions in mouse. The sequence co-ordinate axis represents the mouse sequence.

Key:



5.4 Conclusion

Overall, the approach from using *Eponine* to analyse the nucleosome datasets was met with the difficulty of finding an appropriate negative dataset. Also, only a minority of the total training attempts produced models that had good predictive power. This could be due to the small number of sequences in either dataset. Definitely, a much larger set of nucleosome-binding and nucleosome-repelling sequences respectively is required for a machine-learning tool like *Eponine* to identify important nucleosome positioning motifs. But it did show that predictive models could be learnt; the best trained model showed 100% accuracy at 40% coverage.

In this study, using *Eponine* led to the further analysis of the background trinucleotide compositions in different genomes. This in turn provided some useful insights into the way higher and lower eukaryotes differ in their trinucleotide compositions. The results showed that the most frequent trinucleotides in human and in lower eukaryotes, [CWG, AAA/TTT] and [TTT] respectively, had been previously implicated in nucleosome positioning.

6 Summary

6.1 A difficult area to research

The work, carried out in this thesis, highlighted one important truth: nucleosome prediction is not easy to study either computationally or using experimental means (Section 1.10.1). Experimental protocols are difficult as indicated from the small sizes of the nucleosome datasets. The differences noted between the 2 mapped nucleosome datasets indicate that the genomic background sequences of the source organisms are important. At the current level of understanding, the differences in the 2 datasets could be described largely as biases of the background sequence distributions of the represented species. This could mean that higher and lower eukaryotes differ in the way they position nucleosomes.

Despite the lack of a full understanding of how nucleosome positioning occurs, the mechanism itself is plausible. Proteins are known to recognise and bind to specific structural motifs in DNA. For example, binding of TATA boxes by TBP proteins is well studied and thought to involve recognition of specific kinks within this motif (Kim *et al.*, 1993). The difference with nucleosomes is that they are ubiquitously distributed in eukaryotic genomes so it is difficult to judge how many positions in genomic sequences could represent nucleosome positioning signals. Lowary *et al* estimated this value to be 5% of genomic sequences in mouse (Section 1.7). However, as was evident from the comparison of [CWG]-learnt model labelling between mouse and human (Section 3.3.4), the density of this model's labelling differed significantly between mouse and human. This highlights the importance of nucleosome positioning prediction in relation to the species being investigated. It mostly appears that the results from one species cannot be extrapolated readily to another species, even between human and mouse, which share large amounts of syntenic regions (IMGSC, 2002).

6.1.1 The sensitivity of different methods used to detect nucleosome positioning

The nature of what is understood about nucleosome positioning *in vivo* (Sections 1.10.1, 1.5.3, 1.10.1) has some important consequences for the ability to computationally map such positions with high accuracy. This is especially true for methods which approach the problem using whole genome analysis (Section 1.4.3, Chapter 3).

As an example from this thesis, the cyclical HMM analysis was able to learn a pattern [CWG], which appeared to have a weak 9, 10 bp – periodicity associated with it. The pattern could be learnt from various fragments of genomic sequences both coding and non-coding. To learn this pattern required a large number of genomic training sequences (Section 3.2.5). However, as discussed earlier, the number of precisely positioned nucleosomes should be expected to be quite few (Section 1.10.1) mainly as it would be energetically unfavourable to have an overall large density of positioned nucleosomes. Therefore, combining this view with the results of Chapter 3 suggests that the results may not reflect ‘positioned nucleosomes’ *per se*. At the same time, this does not refute the property that [CWG] could have enriched periodicities at 9,10 bp. The overall impression is that the weakly periodic [CWG] may have some effect on nucleosome positioning but it is unlikely that it will result in specifically-positioned nucleosomes, which could be involved in targeted regulation. To overcome such limitations will once again require compilation and analysis of much larger datasets of mapped nucleosome sequences.

6.1.2 Properties of the [CWG] motif

The [CWG] motif is interesting partly as multiple expansions of it have been described to position nucleosomes (Section 1.5.2). Although the [CWG]-model labelling properties were different in human and mouse[21], the most dense occurrences of the motif were often seen to be in coding DNA in both human and mouse (Section 3.3.8). This suggested that some aspect of [CWG] could be conserved. Another interesting feature of the motif is that it is trinucleotide-based. Given 10 emission states within the wheel models, there was potential for di-, quad-, penta- nucleotide motifs to be learnt. This suggests that [CWG] could have some importance in chromatin structure in higher eukaryotes such as human and mouse – it is a prospect which should be assessed further.

The opposing [W] model labelling to the labelling of [CWG] models (Chapter 3) was also interesting. Firstly, it could be guessed by intuition that the [W] motif models would label areas of the genome, which were also labelled by [CWG] ([W] appears in both motifs). This did not explain the opposing style of wheel-state labelling that was observed. Both motifs have also been suggested previously to have an influence on nucleosome positioning: [CWG] and long runs of [W] having positive and negative influences respectively (Sections 1.4, 1.5.1, 1.5.2). The analysis, using cyclical model labelling, however, indicated that the proportions of either motif were different in human and mouse. This contended the plausibility for [CWG] vs. [W] density to act as a universal positioning property in higher eukaryotes.

6.1.3 Possible influence of repeats in nucleosome positioning

Much of the results, in this thesis, suggest that repeats may have an influence on nucleosome positioning. The wavelet results showed that Alu repeats accounted for

[21] However, it was seen that the same motif could be learnt from training sequences from either species.

previously reported periodic flexibility in human (Chapter 4). Also, both Chapters 3 and 5 indicate that the background distribution of specific trinucleotides, especially densities of [CWG] and [AAA/TTT], may have some effect on nucleosome positioning as these motifs have previously been implicated in nucleosome positioning (Sections 1.4, 1.5.1, 1.5.2). The background trinucleotide distribution is in turn affected by the distribution of ancient repeats in the specific genome. However, as discussed earlier, it is difficult to detect highly diverged repeats or fragments of repeats, which have become dispersed in genomes (Smit, 1999). This makes it difficult to appreciate what contribution ancient repeats may have in affecting nucleosome positioning.

6.1.4 Concluding remarks

Although the lack of data made it difficult to build and validate strong predictive models, the observations taken together suggest that there is evidence of weak nucleosome positioning signals. A model was learnt from the chicken nucleosome dataset which showed 100% accuracy at 40% coverage (Section 5.3.4). It also appeared that the [CWG] models tended to fit a 9 as well as a 10-wheel model in intergenic sequences (Section 3.3.7).

7 References

Aiyar A. (2000). The use of CLUSTAL W and CLUSTAL X for multiple sequence alignment. *Methods Mol Biol* **132:** 221-41.

Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215:** 403-10.

Arneodo A., Bacry E., Graves P. V., and Muzy J. F. (1995). Characterizing long-range correlations in DNA sequences from wavelet analysis. *Physical Review Letters* **74:** 3293-3296.

Arneodo A., D'Aubenton-Carafa Y., Audit B., Bacry E., Muzy J. F., and Thermes C. (1998). What can we learn with wavelets about DNA sequences? *Physica A* **249:** 439 - 448.

Arnott S., Chandrasekaran R., Hall I. H., and Puigjaner L. C. (1983). Heteronomous DNA. *Nucleic Acids Res* **11:** 4141-55.

Arnott S., and Selsing E. (1974). Structures for the polynucleotide complexes poly(dA) with poly (dT) and poly(dT) with poly(dA) with poly (dT). *J Mol Biol* **88:** 509-21.

Audit B., Thermes C., Vaillant C., d'Aubenton-Carafa Y., Muzy J. F., and Arneodo A. (2001). Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Phys Rev Lett* **86:** 2471-4.

Audit B., Vaillant C., Arneodo A., d'Aubenton-Carafa Y., and Thermes C. (2002). Long-range correlations between DNA bending sites: relation to the structure and dynamics of nucleosomes. *J Mol Biol* **316:** 903-18.

Bailey K. A., Marc F., Sandman K., and Reeve J. N. (2002). Both DNA and histone fold sequences contribute to archaeal nucleosome stability. *J Biol Chem* **277:** 9293-301.

Bailey K. A., Pereira S. L., Widom J., and Reeve J. N. (2000). Archaeal histone selection of nucleosome positioning sequences and the procaryotic origin of histone-dependent genome evolution. *J Mol Biol* **303:** 25-34.

Baldi P., Brunak S., Chauvin Y., and Krogh A. (1996). Naturally occurring nucleosome positioning signals in human exons and introns. *J Mol Biol* **263:** 503-10.

Bash R. C., Vargason J. M., Cornejo S., Ho P. S., and Lohr D. (2001). Intrinsically bent DNA in the promoter regions of the yeast GAAL1-10 and GAL80 genes. *J Biol Chem* **276:** 861-6.

Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S. R., Griffiths-Jones S., Howe K. L., Marshall M., and Sonnhammer E. L. (2002). The Pfam protein families database. *Nucleic Acids Res* **30:** 276-80.

Batzer M. A., and Deininger P. L. (2002). Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370-9.

Batzer M. A., Deininger P. L., Hellmann-Blumberg U., Jurka J., Labuda D., Rubin C. M., Schmid C. W., Zietkiewicz E., and Zuckerkandl E. (1996). Standardized nomenclature for Alu repeats. *J Mol Evol* **42:** 3-6.

Bellard M., Dretzen G., Giangrande A., and Ramain P. (1989). Nuclease digestion of transcriptionally active chromatin. *Methods Enzymol* **170:** 317-46.

Blomquist P., Belikov S., and Wrange O. (1999). Increased nuclear factor 1 binding to its nucleosomal site mediated by sequence-dependent DNA structure. *Nucleic Acids Res* **27:** 517-25.

Bolshoy A., McNamara P., Harrington R. E., and Trifonov E. N. (1991). Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc Natl Acad Sci U S A* **88:** 2312-6.

Brown N. P., Leroy C., and Sander C. (1998). MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics* **14:** 380-1.

Brukner I., Sanchez R., Suck D., and Pongor S. (1995). Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *Embo J* **14:** 1812-8.

Buldyrev S. V., Dokholyan N. V., Goldberger A. L., Havlin S., Peng C. K., Stanley H. E., and Viswanathan G. M. (1998). Analysis of DNA sequences using methods of statistical physics. *Physica A* **249:** 430 - 438.

Burt D. W., Bruley C., Dunn I. C., Jones C. T., Ramage A., Law A. S., Morrice D. R., Paton I. R., Smith J., Windsor D., Sazanov A., Fries R., and Waddington D. (1999). The dynamics of chromosome evolution in birds and mammals. *Nature* **402:** 411-3.

Calladine C., and Drew H. R. (1992). "Understanding DNA: The Molecule & How it Works," Academic Press.

Calladine C. R., and Drew H. R. (1986). Principles of sequence-dependent flexure of DNA. *J Mol Biol* **192:** 907-18.

Calladine C. R., Drew H. R., and McCall M. J. (1988). The intrinsic curvature of DNA in solution. *J Mol Biol* **201:** 127-37.

Cao H., Widlund H. R., Simonsson T., and Kubista M. (1998). TGGA repeats impair nucleosome formation. *J Mol Biol* **281:** 253-60.

Chu W. M., Ballard R., Carpick B. W., Williams B. R., and Schmid C. W. (1998). Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR. *Mol Cell Biol* **18:** 58-68.

Clamp M., Andrews D., Barker D., Bevan P., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T., Durbin R., Eyras E., Gilbert J., Hammond M., Hubbard T., Kasprzyk A., Keefe D., Lehvaslaiho H., Iyer V., Melsopp C., Mongin E., Pettett R., Potter S., Rust A., Schmidt E., Searle S., Slater G., Smith J., Spooner W., Stabenau A., Stalker J., Stupka E., Ureta-Vidal A., Vastrik I., and Birney E. (2003). Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res* **31:** 38-42.

Clecom (1999). AutoSignal - spectral and time domain signal analysis and processing software, http://www.clecom.co.uk/science/autosignal/details.html.

Cooper D. N., and Gerber-Huber S. (1985). DNA methylation and CpG suppression. *Cell Differ* **17:** 199-205.

Davey C. A., Sargent D. F., Luger K., Maeder A. W., and Richmond T. J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J Mol Biol* **319:** 1097-113.

Deininger P. L., and Batzer M. A. (1999). Alu repeats and human disease. *Mol Genet Metab* **67:** 183-93.

Denisov D. A., Shpigelman E. S., and Trifonov E. N. (1997). Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. *Gene* **205:** 145-9.

Dowell R. D., Jokerst R. M., Day A., Eddy S. R., and Stein L. (2001). The Distributed Annotation System. *BMC Bioinformatics* **2:** 7.

Down T., and Pocock M. (1999). The Biojava Project.

Down T. A., and Hubbard T. J. (2002). Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res* **12:** 458-61.

Drew H. R., and Travers A. A. (1985). DNA bending and its relation to nucleosome positioning. *J Mol Biol* **186:** 773-90.

El Hassan MA, and Calladine C. R. (1997). Conformational characteristics of DNA: Empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY OF LONDON SERIES A-MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES* **355:** 43-100.

Englander E. W., and Howard B. H. (1995). Nucleosome positioning by human Alu elements in chromatin. *J Biol Chem* **270:** 10091-6.

Englander E. W., Wolffe A. P., and Howard B. H. (1993). Nucleosome interactions with a human Alu element. Transcriptional repression and effects of template methylation. *J Biol Chem* **268:** 19565-73.

Fiorini A., Basso L. R., Jr., Paco-Larson M. L., and Fernandez M. A. (2001). Mapping of intrinsic bent DNA sites in the upstream region of DNA puff BhC4-1 amplified gene. *J Cell Biochem* **83:** 1-13.

Fox K. R. (1992). Wrapping of genomic polydA.polydT tracts around nucleosome core particles. *Nucleic Acids Res* **20:** 1235-42.

Gonzalez P. J., and Palacian E. (1989). Interaction of RNA polymerase II with structurally altered nucleosomal particles. Transcription is facilitated by loss of one H2A.H2B dimer. *J Biol Chem* **264:** 18457-62.

Goodsell D. S., and Dickerson R. E. (1994). Bending and curvature calculations in B-DNA. *Nucleic Acids Res* **22:** 5497-503.

Grate L., Hughey R., Karplus K., Moeri N, and K H. (1996). Tutorial: Stochastic Modeling Techniques: Understanding and using hidden Markov models.

Hager G. L., and Fragoso G. (1999). Analysis of nucleosome positioning in mammalian cells. *Methods Enzymol* **304:** 626-38.

Hamiche A., Sandaltzopoulos R., Gdula D. A., and Wu C. (1999). ATP-dependent histone octamer sliding mediated by the chromatin remodeling complex NURF. *Cell* **97:** 833-42.

Hartl D. L., and Jones E. W. (1998). Eukaryotic Chromosomes. *In* "Essential Genetics", pp. 81 - 88, Boston Jones & Bartlett.

Havlin S., Buldyrev S. V., Bunde A., Goldberger A. L., Ivanov P., Peng C. K., and Stanley H. E. (1999). Scaling in nature: from DNA through heartbeats to weather. *Physica A* **273:** 46-69.

Hendrich B., and Bickmore W. (2001). Human diseases with underlying defects in chromatin structure and modification. *Hum Mol Genet* **10:** 2233-42.

HGSC (2001). The International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Higgins D. G., Thompson J. D., and Gibson T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266:** 383-402.

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T., Durbin R., Eyras E., Gilbert J., Hammond M., Huminiecki L., Kasprzyk A., Lehvaslaiho H., Lijnzaad P., Melsopp C., Mongin E., Pettett R., Pocock M., Potter S., Rust A., Schmidt E., Searle S., Slater G., Smith J., Spooner W., Stabenau A., Stalker J., Stupka E., Ureta-Vidal A., Vastrik I., and Clamp M. (2002). The Ensembl genome database project. *Nucleic Acids Res* **30:** 38-41.

Hunter C. A., and Lu X. J. (1997). DNA base-stacking interactions: a comparison of theoretical calculations with oligonucleotide X-ray crystal structures. *J Mol Biol* **265:** 603-19.

IMGSC (2002). The International Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-62.

Ioshikhes I., Bolshoy A., Derenshteyn K., Borodovsky M., and Trifonov E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J Mol Biol* **262:** 129-39.

Ioshikhes I., and Trifonov E. N. (1993). Nucleosomal DNA sequence database. *Nucleic Acids Res* **21:** 4857-9.

Jenuwein T., and Allis C. D. (2001). Translating the histone code. *Science* **293:** 1074-80.

Kang J. G., Hamiche A., and Wu C. (2002). GAL4 directs nucleosome sliding induced by NURF. *Embo J* **21:** 1406-13.

Karchin R. (1999). Hidden Markov Models and Protein Sequence Analysis.

Karrer K. M., and VanNuland T. A. (1999). Nucleosome positioning is independent of histone H1 in vivo. *J Biol Chem* **274:** 33020-4.

Kim C., Rubin C. M., and Schmid C. W. (2001). Genome-wide chromatin remodeling modulates the Alu heat shock response. *Gene* **276:** 127-33.

Kim J. L., Nikolov D. B., and Burley S. K. (1993). Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365:** 520-7.

Kiyama R., and Trifonov E. N. (2002). What positions nucleosomes?--A model. *FEBS Lett* **523:** 7-11.

Klug A., Rhodes D., Smith J., Finch J. T., and Thomas J. O. (1980). A low resolution structure for the histone core of the nucleosome. *Nature* **287:** 509-16.

Koo H. S., Wu H. M., and Crothers D. M. (1986). DNA bending at adenine . thymine tracts. *Nature* **320:** 501-6.

Kornberg R. D., and Lorch Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98:** 285-94.

Levitsky V. G., Podkolodnaya O. A., Kolchanov N. A., and Podkolodny N. L. (2001a). Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis. *Bioinformatics* **17:** 998-1010.

Levitsky V. G., Podkolodnaya O. A., Kolchanov N. A., and Podkolodny N. L. (2001b). Nucleosome formation potential of exons, introns, and Alu repeats. *Bioinformatics* **17:** 1062-4.

Levitsky V. G., Ponomarenko M. P., Ponomarenko J. V., Frolov A. S., and Kolchanov N. A. (1999). Nucleosomal DNA property database. *Bioinformatics* **15:** 582-92.

Lewin B. (2000). Chapter 19: Nucleosomes. *In* "Genes VII", pp. 567-606, Oxford University Press, Cambridge, Massachusetts.

Li T., Spearow J., Rubin C. M., and Schmid C. W. (1999). Physiological stresses increase mouse short interspersed element (SINE) RNA expression in vivo. *Gene* **239:** 367-72.

Liu K., and Stein A. (1997). DNA sequence encodes information for nucleosome array formation. *J Mol Biol* **270:** 559-73.

Liu W. M., Chu W. M., Choudary P. V., and Schmid C. W. (1995). Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res* **23:** 1758-65.

Lorch Y., Cairns B. R., Zhang M., and Kornberg R. D. (1998). Activated RSC-nucleosome complex and persistently altered form of the nucleosome. *Cell* **94:** 29-34.

Lowary P. T., and Widom J. (1997). Nucleosome packaging and nucleosome positioning of genomic DNA. *Proc Natl Acad Sci U S A* **94:** 1183-8.

Luger K., Mader A. W., Richmond R. K., Sargent D. F., and Richmond T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389:** 251-60.

Lustig A. J., and Petes T. D. (1984). Long poly(A) tracts in the human genome are associated with the Alu family of repeated elements. *J Mol Biol* **180:** 753-9.

MacDonald D., Herbert K., Zhang X., Pologruto T., Lu P., and Polgruto T. (2001). Solution structure of an A-tract DNA bend. *J Mol Biol* **306:** 1081-98.

Mahadevan M., Tsilfidis C., Sabourin L., Shutler G., Amemiya C., Jansen G., Neville C., Narang M., Barcelo J., O'Hoy K., and et al. (1992). Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255:** 1253-5.

Marini J. C., Levene S. D., Crothers D. M., and Englund P. T. (1983). A bent helix in kinetoplast DNA. *Cold Spring Harb Symp Quant Biol* **47 Pt 1:** 279-83.

Meyer I. M., and Durbin R. (2002). Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics* **18:** 1309-18.

Nair T. M. (1998). Evidence for intrinsic DNA bends within the human cdc2 promoter. *FEBS Lett* **422:** 94-8.

Negri R., Buttinelli M., Panetta G., De Arcangelis V., Di Mauro E., and Travers A. (2001). Sequence dependence of translational positioning of core nucleosomes. *J Mol Biol* **307:** 987-99.

Nelson H. C., Finch J. T., Luisi B. F., and Klug A. (1987). The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature* **330:** 221-6.

Olson W. K., Gorin A. A., Lu X. J., Hock L. M., and Zhurkin V. B. (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* **95:** 11163-8.

Packer M. J., Dauncey M. P., and Hunter C. A. (2000a). Sequence-dependent DNA structure: dinucleotide conformational maps. *J Mol Biol* **295:** 71-83.

Packer M. J., Dauncey M. P., and Hunter C. A. (2000b). Sequence-dependent DNA structure: tetranucleotide conformational maps. *J Mol Biol* **295:** 85-103.

Pattini L C. S. (2001). Evaluation of long term correlation in different regions of a gene sequence. *Biosignal Processing***:** 394 - 396.

Pina B., Bruggemeier U., and Beato M. (1990). Nucleosome positioning modulates accessibility of regulatory proteins to the mouse mammary tumor virus promoter. *Cell* **60:** 719-31.

Pocock MR, Down T, and TJP H. (2000). ISMB 2000 Poster Presentations, http://ismb2000.sdsc.edu/posters/poster-list.html.

Polikar R. (2000). The Wavelet Tutorial.

Pruss D., Bushman F. D., and Wolffe A. P. (1994). Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc Natl Acad Sci U S A* **91:** 5913-7.

Puhl H. L., and Behe M. J. (1993). Structure of nucleosomal DNA at high salt concentration as probed by hydroxyl radical. *J Mol Biol* **229:** 827-32.

Quentin Y. (1994). A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. *Nucleic Acids Res* **22:** 2222-7.

Rhodes D. (1979). Nucleosome cores reconstituted from poly (dA-dT) and the octamer of histones. *Nucleic Acids Res* **6:** 1805-16.

Rhodes D. (1997). Chromatin structure. The nucleosome core all wrapped up. *Nature* **389:** 231, 233.

Richmond T. J., Finch J. T., Rushton B., Rhodes D., and Klug A. (1984). Structure of the nucleosome core particle at 7 A resolution. *Nature* **311:** 532-7.

Rubin C. M., VandeVoort C. A., Teplitz R. L., and Schmid C. W. (1994). Alu repeated DNAs are differentially methylated in primate germ cells. *Nucleic Acids Res* **22:** 5121-7.

Satchwell S. C., Drew H. R., and Travers A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* **191:** 659-75.

Satchwell S. C., and Travers A. A. (1989). Asymmetry and polarity of nucleosomes in chicken erythrocyte chromatin. *Embo J* **8:** 229-38.

Schmid C. W. (1991). Human Alu subfamilies and their methylation revealed by blot hybridization. *Nucleic Acids Res* **19:** 5613-7.

Shamir R. (2001). Algorithms for Molecular Biology: Course Archive.

Simpson R. T., and Shindo H. (1979). Conformation of DNA in chromatin core particles containing poly(dAdT)-poly(dAdT) studied by 31 P NMR spectroscopy. *Nucleic Acids Res* **7:** 481-92.

Sinden R. S. (1994). "DNA Structure and Function," Academic Press.

Smit A., and Green P. (1997). RepeatMasker, http://ftp.genome.washington.edu/RM/RepeatMasker.html.

Smit A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9:** 657-63.

Smith M. M. (2002). Centromeres and variant histones: what, where, when and why? *Curr Opin Cell Biol* **14:** 279-85.

Sprous D. (1996). Force Fields: definition and overview.

Staynov D. Z. (2000). DNase I digestion reveals alternating asymmetrical protection of the nucleosome by the higher order chromatin structure. *Nucleic Acids Res* **28:** 3092-9.

Stein A., and Bina M. (1999). A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucleic Acids Res* **27:** 848-53.

Sved J., and Bird A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci U S A* **87:** 4692-6.

Thastrom A., Lowary P. T., Widlund H. R., Cao H., Kubista M., and Widom J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J Mol Biol* **288:** 213-29.

Thompson J. D., Higgins D. G., and Gibson T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-80.

Tomaszewski R., and Jerzmanowski A. (1997). The AT-rich flanks of the oocyte-type 5S RNA gene of Xenopus laevis act as a strong local signal for histone H1-mediated chromatin reorganization in vitro. *Nucleic Acids Res* **25:** 458-66.

Tsukiyama T. (2002). The in vivo functions of ATP-dependent chromatin-remodelling factors. *Nat Rev Mol Cell Biol* **3:** 422-9.

Turner B. M. (2000). Histone acetylation and an epigenetic code. *Bioessays* **22:** 836-45.

Uberbacher E. C., and Bunick G. J. (1985). X-ray structure of the nucleosome core particle. *J Biomol Struct Dyn* **2:** 1033-55.

Ulanovsky L., Bodner M., Trifonov E. N., and Choder M. (1986). Curved DNA: design, synthesis, and circularization. *Proc Natl Acad Sci U S A* **83:** 862-6.

Wada-Kiyama Y., and Kiyama R. (1996). An intrachromosomal repeating unit based on DNA bending. *Mol Cell Biol* **16:** 5664-73.

Wada-Kiyama Y., Kuwabara K., Sakuma Y., Onishi Y., Trifonov E. N., and Kiyama R. (1999). Localization of curved DNA and its association with nucleosome phasing in the promoter region of the human estrogen receptor alpha gene. *FEBS Lett* **444:** 117-24.

Wang J. C. (1982). The path of DNA in the nucleosome. *Cell* **29:** 724-6.

Wang Y. H., Gellibolian R., Shimizu M., Wells R. D., and Griffith J. (1996). Long CCG triplet repeat blocks exclude nucleosomes: a possible mechanism for the nature of fragile sites in chromosomes. *J Mol Biol* **263:** 511-6.

Wang Y. H., and Griffith J. (1995). Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements. *Genomics* **25:** 570-3.

Widlund H. R., Cao H., Simonsson S., Magnusson E., Simonsson T., Nielsen P. E., Kahn J. D., Crothers D. M., and Kubista M. (1997). Identification and characterization of genomic nucleosome-positioning sequences. *J Mol Biol* **267:** 807-17.

Widlund H. R., Kuduvalli P. N., Bengtsson M., Cao H., Tullius T. D., and Kubista M. (1999). Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. *J Biol Chem* **274:** 31847-52.

Widlund H. R., Vitolo J. M., Thiriet C., and Hayes J. J. (2000). DNA sequence-dependent contributions of core histone tails to nucleosome stability: differential effects of acetylation and proteolytic tail removal. *Biochemistry* **39:** 3835-41.

Widom J. (1996). Short-range order in two eukaryotic genomes: relation to chromosome structure. *J Mol Biol* **259:** 579-88.

Wolffe A. (1998). "Chromatin Structure and Function," Academic Press.

Wolffe A. P., and Guschin D. (2000). Review: chromatin structural features and targets that regulate transcription. *J Struct Biol* **129:** 102-22.

Wu H. M., and Crothers D. M. (1984). The locus of sequence-directed and protein-induced DNA bending. *Nature* **308:** 509-13.

Zhou Y., Zheng J. B., Gu X., Li W., and Saunders G. F. (2000). A novel Pax-6 binding site in rodent B1 repetitive elements: coevolution between developmental regulation and repeated elements? *Gene* **245:** 319-28.

Zhou Y. H., Zheng J. B., Gu X., Saunders G. F., and Yung W. K. (2002). Novel PAX6 Binding Sites in the Human Genome and the Role of Repetitive Elements in the Evolution of Gene Regulation. *Genome Res* **12:** 1716-22.

8 Appendix

A. Multiple Sequence Alignments of Experimentally-Mapped Nucleosome Datasets

The sequence datasets were aligned using *Clustal W* (Aiyar, 2000; Higgins *et al.*, 1996; Thompson *et al.*, 1994) and coloured using the *MView* multiple sequence alignment viewer application (Brown *et al.*, 1998). The results are also sorted by pair wise sequence identity.

CHICKEN CORE DNA DATASET

```
Identities computed with respect to: (1) CC56_145/1-183
Colored by: consensus/65.0% and property
                              1 [        .         .         .         .         :         .         .         .         .         1         .         .         .         .         :         .         .         .  ] 183
 1 CC56_145/1-183               ---------------TGCTGCATCCAGGGCTTGCGTTCTTTACGTGTCTGTAAATTAGAATTACACAACAAAATATTATTAGCCCAAACAGATCT--ATTTGAC-CATGAATATTTTTTTTCAGAAACAGAGAACAAACACACTATGTTACAAGTTATAGAGA--------------------
 2 CC129_143/1-183              ----------------TACCATTTTTCTATTCCTCTCTATAAAAATGTAGTCTGAGGATTCTAA-TTGGGGAATTTAGTTAGCTGATCCTGGTGGATGC-TTTCTTCCCCATGCTCCC--ATTAGCATGTATATTTTGCTTTAAAAAAAAAAAAGAAAAAAAT--------------------
 3 CC239_147_TRIMMED/1-183      ----------------GTATTAATGTTCATTTTAAACTTATAT-TATCATTAAATGACAGAATC-ATATATATACATATATATATATATATATGTATGTGTATTTATGTA---TGCATAAAACAAGTGATACACAAAGCAATTGCTCACCATCCACCAACTGATGCT----------------
 4 CC11_148_TRIMMED/1-183       ----------------CCCCAAACATCTGAGAA---ATTTCATTCATTGACAATAAAGGAATAATACTGTGTTAATTAGATGACAAACTGAATGACATAAATAGTGAATATATAATAATTAAAAGAAGAAGTTCACATAATAGTTCTGATTAAAAAAAAAAATCG------------------
 5 CC21_142/1-183               --------------------TGGCATTCTTCTGCTTTCTAAGACTTTATAAATCTATAGGAGATATTCCACTGTTTCCTTCTCTGAT-CTTTGAGGATCTAAGGC-AAATTATCATTGCTTCAAATCTCCATCACTTCTA--CTTCTCTAATCCTATGCAAATGAG-----------------
 6 CC24_146/1-183               --------------TGTTGTGATGAACAGATTCTTGACAGTTACCAGTTCTCTGACAGTAATTTCATGACTAAATGTCCTTATCTTTGTATATAGCCTGCAAAATTAATGTCACTGTGTACCTCCCAGAGA--ATAAAGTGAAACCCATTATGATGTGGGGA---------------------
 7 CC186_142/1-183              -------------------TTGCTTTAAAGTTTCTAAAGTATTTTAATGTTATCTCAGATTTC---AACCTCTCTTTCCATTCCTCAGTCAAATGGCACGATATGTAACAACATACATTATTAAG---AGAAGTTATATA-AAACTCTTGAAAACCACCATATAAAGC---------------
 8 CC72_143/1-183               ----------------TGTTGTTTACTTGGTAATAGCCGATGTCTGAGATTTTACTG-TGTGTACCTGTGTGTGTACAAACCTGCTCTGATA-CAGATCTCCCTACTGCTGAAAGAGCAGATAAAGAAGATACTCAAAGGCA--GAATTAAAATGTAAAAAAA--------------------
 9 CC124_145/1-183              ------------CCCTTTCTTGTGTTTATTACTGT-TTTTACACTCATGTATATGAGTACTTCTTACAAACGAATGAATTTACAATCACAGTTCACCTGTGACATAGAGGTGATTTTGCATTTCCATTTGG--GAGAA-TGGAAAACAAACCAATGTTGGA----------------------
10 CC143_145/1-183              -----------------TCCTGGAGATCAGTTTACACTTTGACACTATGTCTGGTAGCTGGGTC-ACTCGTGCTGCTCTTGGATCTGCTTCCATGATTTGTGTATGCCAAGCTTCTATGATCTAAGAAACAGTT--TAATTAATGATTTAAACCTATTAAAAAAA------------------
11 CC234_145/1-183              ------------------TGTTTTTGTGGGCCTTGACAGTTGAAAAGATGATCATGCTTCCTCTGCCCTTATCTAAATTAACAATAGACAACTGAAACAACAGAGAAAATCACAAAATCTTTAGAAGTTGTGGACCTTT-TTCCTGAAACAGACAATGTTACTG-------------------
12 CC170_146/1-183              ----------------TCCCAGTTTGTACCTGTTTCATCTGAAATATGTCTAAATCAGATTCC---AGTTTCACTTT--GTGCAGAGGTTTAAAGGGCCAGGTTGTAGCAGTGATTATTTTTAATTATATTATTTCTTCATAAACCCCCGAAAAATTAAATAGGAAA----------------
13 CC68_144/1-183               ----------------TGGTGTTTTGAAACAATGTAGTCCTGAGAAACATTATTTTTAGTTCTGCTAATTTTACTGTGTTAGCCTGGGGCAAATGATTCAATCTTCAATAACAAAGA--AACAAT---AACAATAATACTTTGAAGTCAGAAAAGTCTACTATCA------------------
14 CC133_144/1-183              -----------------TTGTCTCTCTCCTTCCCCTGCCTCCCCTCTTCATCCAAGCGACAGGGAGTCATAGGGTTGCACTCCATTCAAAAAATAGCTGCGTCACATCACGTAGTCCAATCCCTCCAGCAGTGGGGAAAGGCACTGTAGCTCGAACAAATG----------------------
15 CC131_148_TRIMMED/1-183      ---------------------GTCCACTGCAAACGCAGTTTTTCATAAGTTCTCCCAAAACCTTCGTGCAGATGGTCAGAGAAGAGCTGGCAGGATACAAGGAATGCCTTGAAAGAGGAAATGAAAAGGAAAAAGATAAAGATAAAAAGAAAAAAGAAGAGAAAGAC----------------
16 CC45_145/1-183               ------------------TTGCATTTTTGTTGAAAGGAGCCATCTCCTCCTTCCCTGATC-AGGCAAGTGATGGTTCAATTGAAGATAATTCT--TGTTGCCTGACATTGCACTTAAGATACTTCAAGGCAGATTCTTTACTTCAGTAAAATAGCACGAGGAAGAT-----------------
17 CC04_142/1-183               --------------TGCACCATTT--TCAAAACCTGCCTGTTT-TCAGAGGACGCCTATGATTTGATGCTTCAATAATCACCCATTTATGTTTTCACTATTATTGCATCTTATGTG--GAAAAAAAACATTTACCAGC---TACACCATGTGAGATGAATGAAG-------------------
18 CC183_147_TRIMMED/1-183      ------------GCAAAATCATCATGGAGAGGCTCAGCACCTCATTAGGTTGTATGCAATTCATTCAGTTGGAGATTGTA---CGTTCATTAATGAAAT-GCTATCACTGATTATGTAGTAATCAATTTTTTCCTTTATCGTA-GAAATACTCTGACCATGCG--------------------
19 CC132_146/1-183              ----------------------CTGCTCTTTGCACAGTCTTGGACAATTTT-TTTCACCTTATT-TTTATGAAATGTGTCAAGGACTTGAGTTCTCTGTGTGATCCATCTCTCCATGCTGATAAACCTTGACACTAATCTGTTACAGAAAGAGTCCACAGCATCCAGA-------------
20 CC123_145/1-183              ------------------TGAATCAGTGAAATTCTCCATGCGAAATGGAGGTTTATTCACCTTTATATATTGTCATTTAATAACCTCATAAGTGAAGAATTATG--AAACCAAGGTAATATTGAACCATGTTTTCCTAT-TTCCTGTCACACATGGTTTTTCTGCT-----------------
21 CC14_143/1-183               ----------------TGCCTTGGATCCTCTTGTTTCTATAAAATTGGAGT-TATTTATCTCAT-CTTTGGTTTGTTGCCTGTAGATCAGTGAGGTTTG-TGGCCCCAGTGTGCTTCTG-AGCTGCAACTGTGTACTACGTAAAATTATTATATCATTAAAGA--------------------
22 CC200_145/1-183              ---------------TTGGATATTTGTTTTAATTGAAAGAAGAATTGCAAATGGAAGGCTGACATATTCTTGATGAAATG-ACAAGGCAAGCAGAGAGCCATACAGAGCTGACAATGTGAGACAGAAAAAGGGGACTATTCTTTAAAGCACTAAATGTGAA----------------------
23 CC69_145/1-183               ------------TGTGTTGAGCTGTTTCAGCACATTAAACTGTTAAACTCACCCTACAGGAGTATTTTTTCCATATGAAAAAAATGGTTCCATATGG-TTAGAGGG----TCCAATAGAGAGAACAGGTAACACCTAAGACGGAATGGAAATGTATGAGGGA---------------------
24 CC156_143/1-183              -------------CTTCTGTGCACTCTGAACCGTGTAACAACAAACAAGT----TTGAGAATTTTGGATGTGAACAGCTTCTAATTTAGAGAGTAGAGTGACCAGTAAAAGCTACCAGTGTTGCCTACAAG--ACCAGGCAAAAAGCCAGAGATGCTCTGCA---------------------
25 CC77_145/1-183               ---------------TTGTATTGTTCCAAAATTTTCTCCTCTAAACATTTAATCCGTGCCTTTGCTTAATCGAGATGTTCTCCCTCAGTGCCAGATTTGGTGCAGAAGATTTCGGGCTCTGTGCCGCCAGCTGGGCAC-TCAGTGTCTGTCAGAATGGGAA----------------------
26 CC26_144/1-183               -------------------TCCCGTTTTTGGGAGCCAGGGGGCCCATGGTCCTACTACTCCCCTGTCACAGGCATAGATTTATCTA-GATTGATCTAGATAACTCCGTGACAACGTATAACC-TGAGTTTGAAAAGGATAGATATTGAAATCATTGTTTAATGCT------------------
27 CC17_146/1-183               --------------ATCCTTTATCAAACCTGAAAGCAAGGCTTCACAGTTTTTGCTG-----TTCATAGGACTGTGCTTAGCCAAAGTAGCAAAATGATTGTACCAAGATACCAAGATAAA-GAGAAGGAAAGAGAGAAAGAGAGAGAGACAGAGGAAGAAAGAAA-----------------
28 CC173_146/1-183              -------------GAGAAGTACACAGCTGGAACAGAGTAGAAAAACAGTCTTGCTTATAT-TTCTCACAACTAAAACTCCAGTAAATCTCATGGAGTTTG--AATCAGAC---TTCCAAGCCTGGAAGATATGAATCATAA-ATCAAAGAAGAAGGTATAGGAGAA-----------------
29 CC52_145/1-183               ----------------TTGTTTAAGCGGCTCCAGCAGCAAACTAGGTCCTTT-CAAGTTGCTGTGTACTTTGGTCTGTGTCACAAAT-TGACAGTTTCATCAGAAAACAAAGTGAAGACTGTTGAAAGCTAGC-CATGTTTTATGGAG-AAATAGCTCCAGCACA------------------
30 CC35_146/1-183               ---------------CAGCTTTGTTGACTATAAATTGAATCACAATCGTAACTGACAAGCTGAG-ATAC--TTTCTGGTTACTTGTACTGGGCAGAGAAATGTAGGCCCTCTGATGCTA-TGGGGAAATTATGCAGGAATCAGGTTGAAAAAAAATCAAAAGAT------------------
31 CC07_145/1-183               -------------TGACA-TAGGGGCCTCAGAATGAAAGTCTTGCTCAGGGCATGTGTTTCTTCCCTGATCTGCTGCATCCCCCTGGGAGTCTTGCTTTCCTCCTTGACTCCATTTTCTTCCCAGTAGCATGACTA-GCAGTAACAAATACTTCATTGCA-----------------------
32 CC33_145/1-183               --------------------TTTTTAATTGCTCAGCTCTAGAGGTGAAATGTCTCACAGGATACTTTGGATCACATGAGCTAATAAAGTGCTTCCTGCTAAACTGAGAGGAACAGCCTTACTACCTTTACATTAACTCCTTGCTTTCTGTCTAAAAATGTAGGGA------------------
33 CC27_145/1-183               --------------------TGTTTTAAGACTCTCGACTTCTATGGCTTACAGAGCTGGGAAGGAAAGGTTGAGATGAGCAGCAGCC-TCAAACTCTTCTCAACCAACAAGC-GAGTAATTTTAAGGCTCCACAGCTAATATGTTCACTGGGAACTCAAATAAAGTA----------------
34 CC159_145/1-183              -----------------CTCTGTTCAGCATCCATACCTTGACCTGAACAAGATCCTGACACTAAGTACTCACATTTCTCAACTTTCAGATCGCTACAGGTCTGATCCATGCCT--TAGACTCACTCCAGCTATCATTGATTTTTCTTTTTTTAAAAAAGAAAA-------------------
35 CC99_145/1-183               ---------------TTCCATGCTTATGTTATTTGCTGAACATTGATAGACATTGAGAAGAATAGATAATATCTTCCCCTTGAGTTTTGAATTTAAAAGTAGTTTTACTACTTCAAGTTCATCTTAGCCAATAAAGTTTTGGGTTTTGGACTACACATCT-----------------------
36 CC225_144/1-183              -------------------TAAATTTTCCAGTACAAAGTACAAACCGTTTA-AAATTAGGAAATACAAACTTCTGGACCAAGG-ATAAATTAAAGATGTTAACATGAAG-CTGTTTATTTCT-TTGTAAGAAATGTGAACGTCTCCTGCTTTACCTGTATTGTCATG----------------
37 CC118_142/1-183              ----------------CGGAAGTCTTGATTTCAACTGGATGACATTTTCCACTGCAAAACTGCCAGGGAAAGGTTT---TGTTCCTGCTAGTTGT-AATCTGATGTTATCAAGTACAGCAGCACTCGGAGT-TAAAAATTACAGCCCATTTTGAAGTCTTGCG--------------------
38 CC208_147_TRIMMED/1-183      ---------------TCATTATGGCTCAAACAC-ATGTTGCCTTTAATACAAGTGCATGATCTTGAATATTATTAAGGTTTAAATAAATTTAAATGTGACAAGCACAAGATTAAATGTGGCACATAAGCAACTTTAAACAGCTAATTAGGTTTTGTTGGTCC---------------------
39 CC178_142/1-183              ----------------TGCTGAAGTGATGGCTC-CCACATAACCTTCGCCTCCCTGATATCAAAACGACATTTAGTTAAGATTTTACAAATGAAAACCTTAACGATGATGAGAAAGAGAACT-----AAAAATATCGTTGTGGGTGTTAAATCCTACCAGGGCT-------------------
40 CC107_144/1-183              ----------------TGTTGTTCTCTACCTTTTTAATG-ACCTCAGCTCTCTGACTGCATCAGTTTACTTTTCTGGGTCTGCCTGGATTTCCTCTTCTTCCTCCTGTCCACTTAATTCCTG--CAACTATCAAACAC-CAGCTGCTGAAATCCAGGTTATGCA-------------------
41 CC125_145/1-183              ------------------TCTTTTGGAATGTGAATTAAGTTTAAAGCAAGAAAGGACAGATCAGAAAAGGAGCCTAGGGAATCTGCACAACAGAGGAAAAACAGCCAGGGATAATGCTGCCATACACTAAGCAGCAATGGCTACGTACTTCCAGAAATAAGGA--------------------
42 CC85_144/1-183               ---------------TGCTTATTGTGATATGAAAAAAAAAATCTCACTGTGACCTAAAAGGCTGATTAATTAGACAAACAAAAATGAAGCAGTAGAAGTCTT-TAGACATTG-TAGGCATTGCTGTGTGCTTTCACAGAAAGGAATGTCTGGGGGAGAGTA----------------------
43 CC84_145/1-183               --------------AGCATTCTCATGTGGTTCTGTCTCCTACCTTTTCTTTTTTGTCCTATTTACTTACTGCTCTCCAGTT--CCCCCCCTGGGCTTGCTGTTGCCTTCCAGTCCTATCCCCTTTATCCTTGAAATTC-AAGGGAATCGATGACAACAGGCT---------------------
44 CC216_145/1-183              -----------GTAATTTGAATTACCTTATTTCTTAACTCTGCAAAGCTGCACATTAAGAGTCCAGTGCATCTTGGAACAGG--CAAAGAGCCTCTACTGTAAATAA--CTCATCT-TGGGATATTCAAGGTGGTTAACAGAGTTTATGCTCTTTTGAACA----------------------
45 CC182_143/1-183              ---------ATGTCTGGTTTTCCCTCTTTTTAATGCAACATCCACTGTTAGCCCTTTGA---TCATCTATGAAGTAGGCAGAAC-ATAAAATATGTCTTGGCTTTATTTTAG--------TAGCTGAAAAATTAATAAACTAGAAAATCTAAACTAGAGAATTA-------------------
46 CC142_147_TRIMMED/1-183      ----------------GTATTTCTACCACTCATACTGCACCACAGCACTTGTGCACGGTGAGAGGTGGCCAGGAAGGATTAACA----TGACCTTTTC-TTAATAAACAGACTATAAGCTTTTAGAGACTGG---GTGGTAAACAAAGCAAAGTGCTTGAAAAGAGCAGA-------------
47 CC36_146/1-183               ---------------TGCATTTAAAGTTGATAAATGGGTGATGAAAGTGTTATTTTAAA-----ATAAGAAAATCACACAGAGA--AGCCATGAAATGTCATGTTGATGTTGCTAATAATTTCTGAATGCTGTGACTGTAAAACAGG-CTGTGGAATATAAAAATAGAA--------------
48 CC203_146/1-183              -------------TCCTGGTACTCAGCTCACAGAATCTTTGGGAGAGGACTTGCAC------TGTGAGATCTGATGCATGTGTTGGGCTTTTAGAGCTTTGCAATGAAACAGTTCTTAATAATAAGAAAAATGAATTGAAGGACTGATTTAGAAGGTAGGAGTGA------------------
49 CC53_143/1-183               -------------TTTTTTTTTTTTTTTTTTCTTTGGTTTTTACAGTCTTCTGGATTGAAGACCTCTATTAGTTGAAACACACAATGCTGCCTGTCAAAGTAAGGGAAGTTCCA-TGTGCCCCAGTTACCTGGGGTT--CCTTTGGCGCAGATACAGCT------------------------
50 CC176_144/1-183              ------------TTCCTTCCTTCAGTACTCCTTGCTCTTCAGTAGAAATATT-TTCTTATTACTTGTACTAGTTGAGATTATGGACCAAGCATTAGGAAAAACTTCATTTGCACA--AAAGCAGAACTCCTT-AGAAAGGCACAGTGGT-GCATCTCTGTG----------------------
51 CC48_145/1-183               -----------CGCTATCCAATCATCCGGA----CCACGCTTCTGTAGGTTACCTGCAAA-CATGCTGTGTGAGACAGTGACGAGTCCTTCACTGAAATCAATATAACCA--CATGCACTGCTTCACCGTGAGCTCTCCAGCACGACGTCATCATAAAAAGCT--------------------
52 CC168_143/1-183              -------------------TTTTTTAATGTCTCTCCATTCTCTGGGGATTTTTGTTTTAACTCTTCCA---GCACACCATGC-AGAAGATACTGGTGAGTTGGTCCCATATTAGGGA-CACTCAGCATTACAGCACCTGAGTTTTCCCAAGAAGCTCAAATCATTA----------------
53 CC29_147_TRIMMED/1-183       -------------------CCTTGGTATTTCCTCCTAATCTGACAAGAACTGTTGACAATATTTTTTCTGAGTCATGAAGCACTGAAGCACCAAGGGAAGTCAGG---CTCTTGACTGGAGAGAAGCGTTGCTCCCAGCTGCTGACAGTAGTGCAAGCATAAACAGCT---------------
54 CC110_147_TRIMMED/1-183      -------------------GCTTTCTACCTCCTGCAGGAAGAAAGATTCATTTACATTAACTCACGTGACGCCAGATCCTGAGAACCACTCAGATGCAATTGCAGTCAAT-TAAGACAGTCTGGGGAATCAGGCAAAACTTCATCAAGATAAGCTACTAGAA-AGCA----------------
```

```
 55 CC73_145/1-183               --------------CTCTTGCTCTCCAAATTTCTTTCCTCCTGTTTACTG-GCACAACTACTTC--TGCATCAAGGAACTGATCCAGATGGAAAATAACCCAGAAAA--AACATCAGCTGGCTTCTGGACTGGGCTCACAATGTAGCTGCACAACAGAAGGGCA-------------------
 56 CC137_145/1-183              --------------------TGGGAATTGTGGATTTTTGAAGAA--AGCAAAGTCTCTACC-TGTTTCCTTAC----GTGTACATTTCCTTCCAC--AGGGATGTTGTAAGCACTCCGGATCAAGTTCTTAGCAATGAAATAATAATATACTGAAATGACAGACAGAGGAATGA---------
 57 CC41_142/1-183               ---------------CCATATGTAAATAAATCTCTGGGTGCGTTCTATAAATAGCCACAACTCCTTCATTACAAATGGCCTTTTGCTCTCAGCGCAGTTA-CACAGAACCTGAACCTTGTCCAGGTGGAGGTCATAGCATC-TTATGAGCGTACACCT------------------------
 58 CC139_145/1-183              ----------------CTGGAAACTCTTACAACAAATAGAAAAATAACTGCAAATCATCAT-GGATTACACAGCACCACTGTTACAAATAGTTTTTAATATATT-AATGAGCGATGAAGAGATGTTTTAATAAATAGTTAATGAGTGATGCTATTGGTATACA--------------------
 59 CC141_143/1-183              ---------------TGCTTAGCACCACTGTTTTCAACAGAAATCCCAAACCTAACTCTATCCCAGGCAAAACCATGACAGTCTGATTTGTCACATTTCTAAAGCCGAGGTTCAGA--AGTCACAGCGCAGCAGTTTGTTAGCACTGGTGTGCAAAAGCT-----------------------
 60 CC106_144/1-183              --------------------TAAACTTCTGAGGACAATTCTGTCTTCTTTCACTCCACTTCTTT-TTTTTCTTCTTTTTACTGTGCTTGTAATTCATGAGTGAAGAATT----CAGGGTGATAAACAAAAAACCA-TGTTTAGACTCACACACTCTTCTTAAAACAAGTG-------------
 61 CC163_143/1-183              ---------------------TGATCCAAGTGCACACTCATGTTTGGTG----GTCCTTCTTCAGCCTTTTCTAGGGGAACATTTATTAACCTTTTGTCTCTGGAAAAGTTTGACCCCACCAAAATTGCACTAATTATAAAAATACAGCTTCAGAAACACACAAAGTA---------------
 62 CC195_142/1-183              ------------------CTCCTCAGTTTCTCTCTAGTTCATCTGTGTTGCATGCCA--ACTGTAGGAATAGGAAGGAAAGCAGCCTCAGTTTGCTGAAGGGAACATGACACCCTAAGTGTTCGGGCTAGGGAGGAGTTTAGCTGTGTATTTCATTTTTTTT---------------------
 63 CC103_143/1-183              -----------------ATTTTTCATTTACCGATGCAAGACCTTTGAGTAAAATGGAAGGAGATAATCATCTGCT--CAACTCTGAAAACATAAAGCGAGGTTGT-ACATAAAAAGTGAC---GGGCTGTTTCTCTTCTTGGTTTGGAATTTAGTTTGTCTTTGGT-----------------
 64 CC148_147_TRIMMED/1-183      ----------TGAGTATTCTCAAAAATGTGATATT-ACTCTTACTTTTCATGGGTAGTTCATTAGAGAATATAAAGACTACAAAACCATATAAA--GATATGCAATGTTAAGTACATGCAACTGCAGCTGT--ATAAA-TGACAGACATTCAATTTTAAACA---------------------
 65 CC194_145/1-183              ---------------TTCTTGCAGACCTGTCTGAGTGATTTACAAAACCATGAGGGAGTCCTAATTAATAAATGAAAAAGGTTAGACACAGCTTAATACAGAGGGGGATGCGCTAGAACAAAA--TGAAGAATGTCTGAAAAT-TGCTGGGTACTTTAAAGGA--------------------
 66 CC161_145/1-183              -------------GGCCATATGTATAACTGTACCCTGATCTTT-AATGGCTTCTCCCCTTTTGACTGCTGAAATTATATGCATTCAAACGTGTATTGAAATGACATACCTGTGGTCAGCTCCAGAGTTTTCTTCTCTTTGAGTAGTGAAGGAGGAAACA------------------------
 67 CC01_142/1-183               -----------------------AGTTTCTGTCCATTCTT---TTTCATAAAAACATCAAAGAGTTTCATCTATTGGCAGCTTTGGTACCCCGTAGCCTTAGACT-GCTTTCATGTTGATCCAAGGAAGGTTCACAACCAGTTTTCTCTAGTCGGTGGCCCATGAGACA--------------
 68 CC236_144/1-183              -------------------CTGTGAAACTTGCAATGCTTGCTCTGCAGAGACCACTGACTCTAACTTCTAACTTCTGTAAAAT---AGAT-GCTATAAAAATGTTTTAGTCTT--TGGCTTTTAGAAGGTCACTATTAAGTTTAGAAAACACAGAGGAAAAATCTTAGA--------------
 69 CC188_143/1-183              ---------------TTTTCAGTGATGTGTTTGCATATATTGTGTT-CCAAAATACAATTTCAA---ATGAAGATTTAAAGTTGTCTGCCTTTTTAGATGGAATAGGATTGGAAGAGACCTTC--AAAGATCGTTCGGTCCAACTGCCTGAGCCCTCCAGGGTG-------------------
 70 CC199_145/1-183              ------ACCA-GATTTCTGCTCTATAGGGCAGCTGGAAGTGAATTTCATTATC-TTGAGTCACAAT-TAGAAAAATGTTTGATCTCTTTACTTGT--ATAATCTTATGTGTGGCTGAGCAAAAAGAAAAAA--TGCAAAAGAAAGGTAGTGGAAAGAT-------------------------
 71 CC128_146/1-183              -------------TCACTGATATTAAATTTGCTGACATCACTCTTCAACAAGCA--GATAGGGATTACATGGAGCCCAGGGAGAA-AAATGCAGACTGTCATTCTGTAGCCTCCGTGGCTTGAAAAAACTTGTCATAAAGGGGAGATTATGAAAAAAATGAT---------------------
 72 CC120_147_TRIMMED/1-183      --------------------CCCAGTTGATAGAGAAAGAAGTTCTCTAGAGAGGA-ATACAGCACTCACAGGCATAGATGTATTTTGGATACTCTGAGTACCTCTTTTGGTCTCTGTGTTCATCAGGCAGACCCAGTGTGGTGTCTCATTCTGCCTGGAAACCCTG----------------
 73 CC81_145/1-183               -----------AGGTATTCTGGAAATTATATCTCT-CCTGATAATCTA-ATGCTTCATATATGTTTTAAGTGAAATACTGCTTATTCCTATAAAACGATTTACAGTGTCAGAGGTGT----TGGCAATACC--ATACC-TTACAGTGTCTCCATCAGGAAACTGG------------------
 74 CC119_146/1-183              ---------------TGCTTCTC-CCGCTGTCCCCTTTCCTTCTGTCTGTTTTAACAGTGACCTTGCCA---CCACCAGAGTC-AAGCCACCACCACTCTCAAGCCTAATTTAATACTATTAAGGAAGAGTCAGCTTTTTCCTTATGCTGGAAGTATGCAATAACA-----------------
 75 CC145_147_TRIMMED/1-183      -----------------GATTTGGAATCAAATAATCC--CAATTTCTGTGTCATTTGTGTTGTGATAAA-TGAAGCAGTTTTACACTAAGTGATTTGATGCTTCT-GAAGTTGCATTGCATTTGTTGGGTTTCTGTTTTGGGGTGGGTGATCTGGGGATGTTTTGGT----------------
 76 CC65_144/1-183               --------------TACATTTTCGTCTTTTAATTTCACTCCTGGCTGTTGCATCTCAGTCTACC---ATGGCTCATATCT-GCCTTAACAACATCTTTATTCTTCCAGTTGAATATACCATACCTTTTAAGTATTGT-----CTGTGAA-TAGCAAGAGATTTTTACT---------------
 77 CC220_143/1-183              --------------------------TGTAGATTAGCCTTAATTTTTGGCCTAGCTTGATTAACCTATTTTAAAAGCCTCCTTCTGTATGTTTTCTTGCAGCATATGTAAGAAAAAAATATAAAGCA-AGTATTTAAGCTACAGTGATGAGAAAACTCAGCTGGTTAACA-------------
 78 CC184_143/1-183              ----------------TTCGTGTTGTAGGGGTGACAAAGTCACCAGAGATTTGCATGTGATCTA-ACACATCCTCCTA------CTCCATCCATAAGCTGTGAATGTAAAATCACTCTT-TCCAGGAATCTATTCATAAATAAAGAGAGAAATAGAGTGCAAGTTCA----------------
 79 CC114_144/1-183              ---------------TGCAACAC--AGGAGCAGCAATAAAGCAAGACTGACAAAAGAACGTCAAACAACAAGGAAAACAGAGCAAAGCAAAACAAAGACACGAAATACAAGAGTCCAGCACATGAACTCAAGACAGCATGATC-GCAAAGCCAAACACAGGA---------------------
 80 CC202_146/1-183              ------------TGCTGTAATCCCAACTATTCAAGCAAATTGCAGAAGCATCTTCTTAA---ACATTGCATGTTGGGGCAGAGG-GGAATATTCGCAATGCCTATGACAGAGCTCAGCCTTGGCTGAAGCCATCGCAGACTGCCCAATGCAGACAGGAGTAG---------------------
 81 CC187_146/1-183              ----------TGTCTTTTATCTCTGATTAAATGATAACGGTAAGTCTGATTGAGTTCTGTTTGGTTGA--TTCTACTGTTGTGCTTCACTGCCTTATAGGCAGAT-GA-ATCAAGTTGGATGTTCTTGATAGGCTTAACTG---TGAAAACCTGATCAATGTT--------------------
 82 CC217_145/1-183              ---------------TTATGATAGAGATAACAGGATGTATCCATTTGCGATAGCGCAGGAT--TAGGCATTATGATCCAGGAAGTTTGGTTCAGTTCTTGCCATAAAAGCCAAAATAAAGATGAAAATATAACCTATACACCAAAC--AGCCTCATTCATCAAA-------------------
 83 CC130_142/1-183              ------------CTACTGAGAGAGAAACAACC-ACCGCTGTGAAGCTTCATTGCAGACAG--TAAGAATGAGAGAATTCTGCCAACTCACTTATTTCAAGGTTTTCCAATCTTAACAG----CCTATTACAATGATTAATGTGAGCA--AACACGTTTTAGCA--------------------
 84 CC08_145/1-183               --------------TTTTTTTTTTTTTTTTTCCTGGGATAGAGGCTGTCA-GTCTTATATACAGCT---TGGGAAACAATGAAGAATGTTTCTAGCAGATTATGTATTACAGAAGGCAGTTGAGGTTCCTAC-ACAGGCTCTCATCC----AATTCTGTATTTTAGCA---------------
 85 CC42_145/1-183               -----------TGCTTCTGTCCTTATAACACATTCAGGAATGCAGCCTTCATGTTCAAAGGACAGCCAAACTGTTTGCTTGGTCAATACACTTGTTAATAGTTTGGTAAACAGTACAAGAACACGTGGATTATTGCAGTTTGAAAGTGTTTTGAAG---------------------------
 86 CC22_145/1-183               ----------------AGAAAAAGGGAGTCTATGTCCTCT--AGAAGAAGGACAGGCTACTTGGGGAGATCACAAGGAAGTTGCTAAGGTATGCAGGGAGGAAGTTAGGAAGTCAAAAGTCCAACTTGAATCAGATTGGCCATAGCAGTAAAAGAGAATAAG--------------------
 87 CC149_142/1-183              ------------------AGGGACAGGCATCCTGCCACTGGTGAG----GCCACCAAGAA--TGTGTGTCTGTGTAGCTTCCCAGCTACCATCAAGGGAAGATGCTGCACCATCACAGTATTCCTGAAAGAATGCTGAAAATGCTCATGAAAATGATGTAGTTGAG-----------------
 88 CC224_144/1-183              ---------------------CTCTGGAGAAGAGCTAATGGAGAGCAGCCCTGTGGAGGGGGACTTGGAGTTTCATGTCAATGAAGAGCTTGAATGAGCCATAAGTGCCTTCATGCAGCCCCGATGGACGCATGAAACCTTGGCTCCATCAAAAAAGAGTTGGCT------------------
 89 CC15_147_TRIMMED/1-183       ------------------GCTGCATTTCCTCACACACCTATATACATCTGTGACTGTGTGCAT--GTGTGTACTCATATAGAGCTTATCAGGGAGCTCTGCATTTTTATT---TATGTTGAGCAAGACTCTCAGACAGTCCTTACACATC-CAGACCTACTGTTTGTAAG-------------
 90 CC112_145/1-183              ---------CAATGCACATGTGTGTTACTGCCCTGGTAACCTG-TTGAGCCGCTACCGGTGTGAATGCTCAGAGCTGTTGAAGAGTCAC-TGTTTTGCTCTAATA-AGCTGCGATC--TTTAGGAGCATCCTGCATTGTGATATTGCACGTTAATAGCA------------------------
 91 CC30_144/1-183               --------------------------TCCAAAAAGGCATTTCCTCTCTATCTG--TAAACT-CTGGAATGTTAATGCATTTCTTGAGACTCTCTCATGCAACACCTGACTAAGAAAAAATCCATAAAA-GGGGAACACTCAAGATATGCAGAAACAATTAAAAAAACCCCAAACA---------
 92 CC167_147_TRIMMED/1-183      ------------------TATGAAGATTAATAAATAGGTGCAGTTGGCTGCCTCTAACAAGTTGCTCTTTAT---TTGCTTCCAAGCTGCAGGCTGAGTAAATTCAGGATCTGAGC-TCCTCCATCTCCATGACAATGCAG--TGATGTGCAGACACGATGCTAAATGAA-------------
 93 CC39_148_TRIMMED/1-183       -------GCTTTACTGGGGCTTTTCATTGCTTATTTGGTTAGCCCTGAATCTATAATATGTTCAATCTTATGCCTTAGGGCTTCTGCTA---GTCACCTGCAGGGGACAAG--AGTAATTT-TCTTTC--ACTGCTTGATGATTTATTTGAGCAGCCAAGG----------------------
 94 CC75_145/1-183               ----------CTACAAACTTATGTTGGCTGTCTGGAGAAACATGGCCAGTGGTTATTTCACCCCTTGCA---CCATGCCATG---AATGACAGAGGTAGCACAGATGCATATAAGTCAGTGCTCAGAGAGGCTGGGCACAGCAATTCTCCAAAGAGAAGAA----------------------
 95 CC108_147_TRIMMED/1-183      ------------------------TTCTTGCTTCCTGATCCACTACCAT----ACACTAGGGCAAACA-AGGTAGAGATATCCCTCCTCATTCCCCTGTGCAAACCAGTTCTCTCTTCAACCAGCATTTAGC-CAGAAAGCACTGCAATAAAACTCACAAGTGTGTAATGAACAGG-------
 96 CC136_145/1-183              --------AATGAAAAGCTATGAAAGA-GTACACAGGCAATCCTTACTCGCTTATCTGATTATGTAACTCTGAA--CGCAAGCAAATACTTGTAC-TGTGAAAGCGGG-AGCAGCTGCCTTCCAGAGCTGCAAG---AACTGTTTACACCTATTAGTGTGA----------------------
 97 CC153_146/1-183              -------------ATCCCCCTTGTGCTGCAGTTCGGTTGTGGAGGCCATTAAACCATGTCTCTG--TGGTCCAAACAAGA--CCTCAG-GCCTGCAACTGTACA-AACATTTCTAGCTCAGCATTTGCTGCTGTTCTCATTAGTGTAAGAACACATAAGAAAACA------------------
 98 CC96_145/1-183               ---------------------TCTGCTATATGC-TGCCCAGAAGAAATGCCTGTTTGGGTGACAGATTTTGTGCTAGTTACAGACACAAAAA-CCTGACTGAGCTCCTACAGAAAAGACTCTTGCCAGTTCTGGCTTTGATTTGAGCTCTTGAGGAACTGGAAAAGCA---------------
 99 CC231_146/1-183              --------------TCCAGATTTATGCTGTGCAGCACTTGGGAATAATTCATGGCAACCCTTCC---TCAAAAAGCAACT-CAACCTACAACCCCACCTTTGCATTTCCTGCTGCTACTTTCACGCCCAACCATTTTA--TTCCATCAT--GACCTGAGCATTAAAGA---------------
100 CC215_144/1-183              --------------------CTGGTGGAAGGCTGAGTTCGGCTTTTGTGTAGGCTGAAAAAGACTCGGAGTGGGACTGTGCTTGGCTTTCATCTTTAGTAAATATAAATCACCA-GCAGTGCTTTGGCTGAAATGTAGGGGTGGTGGCTTCTGAGGTGTAAGAGG------------------
101 CC31_144/1-183               -----------AGTTTTGTTACAGTTTTAACATATTAAGTTGGGGT---TGTCCCAACT----------TTCTGTTGCATAGGAGAGAGCTCTCTTGCCAAGGAGAGCTTGTAACCTAATCTGAAATTGGATGTGAGCTAGGCTGATGAGAGGAGAAGGGAGTGAGGA---------------
102 CC162_143/1-183              ----------------CTGCAGTCTATGCAAATATCCTTTTGTTCAAGAATGG--TAGACCACTAGGATGTTCTGTTACTTCTGGAAACAGTAGCA-GCATCTGGGGACAAAGATTATTACAGGGAA-TTA-ATTGTCCAAGAGGGGAAAATGCATTTGTTGAA-------------------
103 CC212_145/1-183              ---------------------TGCTTTGAGCACACAATAGAGGATCATGTTGAGTTCCTCATCAACCAATGCTCCAAGTCCGCCTCCATAGGGTTCTCCTTCAGCCA--TTCTCCTTCAGCTGAACTGGAAGTGTTAAACATAGTGCCATTCAGAGTCTCTGAAAGCT---------------
104 CC155_144/1-183              ----------------------------TTTCTAAACCATATAACTTATAGACCCTTGGAAATCTGTGATTGCAACATCATTCAGGTTTGGATTTGCTGTAGTAAGTGGTTACCTGAGTTGCCACTGGACCACAGGGTCAGTTTTGAAAGTCAAGGATCTCACTAACTTACG-----------
105 CC34_144/1-183               ------------TCCTTTAGTTGAAGCCTAATGCAAGCAGTTAA---GGTGGATCTCAGATTTTGTGT--TATTAGGATTAAATTATTCCTGGTTTTCACCATGT-TA-GTGTTGTTCCTTTTCTGTGGTGGTTTGCACTG---TGAAAGTCTGAGAAGTAGTGCA-----------------
106 CC100_142/1-183              ---------TTTTGTTGGCATTCTGACTGTCTGTTTTGCCTTTCCACAAGCAATAATGGGCTGGTTCTAAAGC-----AGATTTTGCTT---GAACACACAGGATTTCAAA--TGAAATTT-TACTGCAGACTGAATAAGGAGGAATAATGGCAGGTAAACA---------------------
107 CC192_146/1-183              ------------CGGGTTGCGCTGT---ACCTCCTTGGAATGCCGTACCCTCCCCCGAAGCCTGCTTATCAGCCCTGCAGGAGATGTTACCACATTC-CTCCAAGGACCTTCCCCCCCCCTCCCCGACTGACTTCAGTGAAGGATATAAGCCTTGAGAGGGA---------------------
108 CC177_145/1-183              ------TTGTCAACTCCACTGTCAGCAGGTTTCTATTTGCAACTGGGTTTTGTTGTTGTTGTTGTTTGTTTGTTTGAAGTC--CTAACTTGGCAGAGGTCACTTTCATCCATTTCTCTCCTG-----CCTTGAAAATT-GAATTTTGCAAGCTCAGCAG------------------------
109 CC190_147_TRIMMED/1-183      ----------GTTCTAATTCAATTGATGGACT-TCTTGTATTCAGTGAGCTTTGAATGAGATGAAAGAAAGGC----ATTGTAGTGCTA---TAACCATCGTGGTGCCAGG--TTAAATCTCTAGCTCACACCACAGAATCAATATGAACAGCATGTTGTCATTGG-----------------
110 CC238_145/1-183              --------------------------TGGTTGTTTTCCCACAATCTTCTTTATCCTGCTGTTTAAATTAT--TAAACATGCA-AAAGACAGAACATACTTTTGCCT--TGAGGCTTGCATCTCTATCAGTTAATCTAATTTTTCATTTGCATTC-ATCATAGATGTAACAACTACAG------
111 CC80_145/1-183               ---------------------CTACTCCTTTTTTCACTGG-CAGAAGTTCTTGGA--GGTTACAGACTTCCCCATCTGCC---TAATACACATTCCCCCTTGTATGTGACCTATTAGGCCGTGCAGGAGCCTGCCCACCACAGGTGAAGCAGGGAAATCCA--AGCTCAGTCCT---------
112 CC147_146/1-183              ---------------------CCCATTCCCGGGGATGCGATGTGGCAACAGTAACTGCTGCCTTCCTCCTTCCCTCCAACCCCAATGC-----CACACCTTTACACCATTAACACAACAC-CACAGAGGACTGCAAGGGGAACCAAATTAAAGAGTAAAGAACCAAACAAAAA----------
113 CC205_145/1-183              ------------------------CTGTTGCCAGTATCAACATACTGCCT-TTTATGTAAAAGAGA---TGATAGATTCTATTGTGCATCTGCATGGGTGTAGGTGTCCAACATGTGTGTGCAATATGACAGCCAGAAAGCACTCTATGCATATAGGAGAGGCTGTTTAAAGA----------
114 CC63_145/1-183               ---------------ACCATCTAGTACACTGTCTTAAATCTA--CTGCCA-TTTAGGTAATTATCA---AGGGCAACG-TAAGAGATGTATATGATGTACTCTGCATATTTAAAGCAAGGCTAGGTTCTTGCTGTATTGTCTTAACTCTTAAATTCCTTACTTTCCT----------------
```

```
115 CC78_145/1-183                 -----------------------TGCCTTGTCAGC-TCAACAGACAGGCA-TTGGGATGGGAAAGAACTTGGATGAGGCTAAAAGGGGGAGTTCTCATCACCAGTGTTTGTGATGAGGGAACAGGAAGTGCTTCACTGATGTCTTTTATTGGATGCCACAATATTTTCCG-------------
116 CC23_144/1-183                 -------------TGCTTCTTTCTATCTTTTTTCCTGGTCAGTTACCCTGGAGGCCCCAATGCTCCCGCAGGGATGGGCATCCCGCCGCACACCAGGCCACCAGCCGATTTCACC--CAGCCAGCAGCTGCTGCTGCCGCCGCTGCAGTTGCAGCTGCA------------------------
117 CC179_143/1-183                -------CAAGCTTCTCATTAAATTACATGATTTAAAGGGAATGTAACTGG-CTAACATTTAATAATAAGACGTCCCTTTTTTTCTGGGTCTTTGTATGCCTTTATTAC-CTACTTAAGGTT-TAATATTGAAACC--TTGCAATAATTTTTGAA----------------------------
118 CC82_145/1-183                 -----------------------AGCTAAAAGGGAAAGGGGGGGCGGGGGGGAAC----AACCTATCAGGCTGTTTAATGGACCCATGGAGATCTTGAAAA---CTGCATGCATTTTGGGCCACCGTATAAGAAGGAAGATATAGCAGAATTGGGATAAAAA-GCTGAAAACTGGA-------
119 CC70_144/1-183                 ----------------TCTTGGGCAGAGGGAAAAAGGAGCCTACTCAGTCTGCTTCTTCCACAGGTCACCTACTGCCTGTCTGCTGGACATGGGGAAGCTCAGAAATGATCCATAATGTTTGTATGAGCTTCATTTTGTGTGCCCTACATTTTGTCTGCA-----------------------
120 CC05_146/1-183                 ------------------------AGCCAAACCATGTATAAACGTTGCCAAACTGCACTACTCTAGGAATCGCAGTGTTAGCAACATC--TTCTTAGGACAGCACTTGACTTCATGAGCATTAAGTCTTCGCTGTTATAAGCCTTCATCCTCCAAAACTTAAGCATACAACA-----------
121 CC206_144/1-183                CTTTCTACGACAGCAGCAGCACCATCAGGTAAACAAAACAGGTTTCTAATGTTATGGAATGACAGTATTTAAAATTTATTTATTTATTTATTTAT--TTATTTATTTAATACAATCAATGTCATTAGGACA--GGGAAAAAAAAACAA-----------------------------------
122 CC98_143/1-183                 ---------------TCCCAGTCCAACAGATTTTGGTCTGTTTTTGAA----CTGAGAAGTCT---CCATCCATTCCAAATACCTGTTGCATGTGTACAC-GTTTTCCTCCTTTCCATTCCCAGTCTTCACTATGGATTTGTCCCTTG--CTACACGTCCTCTACTCT---------------
123 CC116_146/1-183                ----------------CGGGCATGGCCTTGCTCAGCAAGG--GGAGGAAAAGCCTGAAGGGCACGAGAAGGGGATAGAGAAGGGCAGGCAGTAGGTGTGAAGCTGCGAGCAGAGTAGAGAGCAACTTGGCAGAGCAGCAAGCAGAGTAGGCAAGTGTTGAGGTT-------------------
124 CC55_145/1-183                 --------AGCTGTTTTTTTTTTTTTCCAGAGCTCTGCTAATTTACATTTTCCCTCCAAGAGCCATCTTGCAGGATAGAA----GTTGTGCGGT-GTTTCTTGCCTTGTTTCTGAG--GCTTGTGGTGACTGT---GGTATTTGTCCTGATAAAA-GATCTGCT-------------------
125 CC94_146/1-183                 TGTGAGATGTGAATCTTTATGTCTTCTTCATAG-GTTTTTCTATATCAGT----CAGAGTATCTTCAGTGTG----GCTGTTCCCTTA---AGTTGTATAAATACTTATATCTATGAAAAATGCCCATAGGGAATCAAGAAAAAAAATAAATATTCTC-------------------------
126 CC196_145/1-183                -----------TTCCTCCAACTTTAAGAACCAGCTTTAAAG--GTTTTTCCAAAGGCTTGTCTA--CATAATCTAAATAAATAAAA--AAGTCAAAGATGCT---TATGAAAGTG-ACAGCATGCTATCATTAAAGCATATAA-ATTATGTTAGTTCTAGAAATGCT----------------
127 CC211_143/1-183                ----------------CCGGAGAATTCCAACAGCTCCCACCTGGGTTCAGAGCAGGGGA-TGGGATCGGGGTTCTGGCCTCACAACCTGCAAAGAACATCACCAGGAC-TGCAGGT-GACCCCACAGCCCTGCATGGGATGCTCAAAGAGTTTGGTTTCCCA---------------------
128 CC25_143/1-183                 --------------CTCTGACCACCTCTCCTATAAGGAAAGG-------CTGACAGAATG-GGCCTTGAATAGCTTGCAAACAAGAAAGCTCTGATGAGACCTCACTGTGGCCTTCTGGTACTTGAAGGGAGCATATAAACAGGAGGGGGGAACGGCTGTTTTCA------------------
129 CC12_145/1-183                 ----------------AGCTTCTTCTCGTGCTCCCAAGGATTAACACTATT-ACACTGTG--TCCATTACATCTGTTCCAAGACATCCATGACCACAGTTTACCTGAGGTCTACTTAAGGCT-TATTATCATCCC---ATGGTTATGGTTGCAGCTGTGTTTCCATGG---------------
130 CC20_146/1-183                 ---------------ACCAACTTCTCCTGACCACCCACACCACGGCCAAGGTAGAGGGACTTACACTGTGGATACATCATACCTGCCTGAAGAGAGCA---CCAGGAC-CACAGTG-GACAGCACAAGCAGA---GAGGACCGCTGAAAATCAGGATCCGAAAGCACGG--------------
131 CC46_146/1-183                 ---------CCTCCTCTCCCATTGTCACAG-----CCCGCCATGCTGTGAGAATTGGAAG-CTGGCTGGGTGGATTTGTGACATCAAGGTCAA-AAACTTTTTATGACCAACCATAGAGC---TAGTTCAAATGCAAGTGAAACGTCATTCTTTGAAGTTATGCA------------------
132 CC127_144/1-183                ----------------TCCTAACATCCAAAGTGACCCTCCCCTGATGCAGCTCCATGCCATTTCCCTGGGTCTATCTCTAACAGATTGAG-GATGCCAGCCTTGAAGGCTACCTTGGATTCCTCCACACCCTTCTGATACAGGCTGGGGCTCAGGCCAGCA----------------------
133 CC06_145/1-183                 -----------ACCAACGTCAAA-ACCCAAAGAGCCGAGCAGAGCGCTGCAAGTGAGCAGAGTGCT---ACAAGTGGTTGACCGACCTG--AGCAAAGTGTTTAGGTGAGGACTGCAGCTCACCAAAACATTCCCAAGCAGCTCTATGC--ACGAACAGAAGTA-------------------
134 CC223_146/1-183                ---------------CGCCCTGAGGGCCGTGCCTGGGCACTGCT------CTGGGCTGGGAGTGTGGGTTTTAC--AGCCCACATTTAATTGTCCCTGCGGCCGCGAGTGGTTGGAGCGCTCCCGAGCATTAAGGTTAAATGATGATATCACAACTTACGATGTTTGCA--------------
135 CC59_146/1-183                 ----------------AAGTTGATCGTAAAAAG--GTGCATACCTAACACA---GGGAAACTCTAAACTTCA--GCAATCAGTCACAAGAAGAGAATTCCCTGCATGCTCTAAGGAAGAGTA-GTTCTGAACTGATTGTTCTCAGAACAAGAAAAGGAAGGTTGATGGCA-------------
136 CC219_148_TRIMMED/1-183        --------GCTGACGGTGTTGCTCAGCCATTGGGCTGGTGTGTGAGCTGGAAGGGAGAAGGCCTGC---AGCTGTGCCTGGGTAAGCCAC-AGCCGTAGGTTTGTCAAACACGTGAGTCGTACCAGTGCAT-CCCTTGCAGCTGCA-GC--AAGAGGAGAAG---------------------
137 CC235_147_TRIMMED/1-183        ------GTGGGATTTGCTGGGAACTCCTAGTTCCTTTTTATAGGGCCGCTATTCCTGGGGGATTTCTGTGAGGGGATTTCTGACCTTT----GCACAAGAAGCCATCAGGTCGCGGAGCGG--CACAGAAA--TTGAACAGAAAAGCTTAAGTGACTTGG-----------------------
138 CC122_148_TRIMMED/1-183        ------------CCGGTCCTGTTCTGCCAAAGGGATGGAGAAAAAACAATCCTTAAGGTCAATGAC---ACAAGGGGCCATCCCCTGGGA-ACAGCTGCAAGTGATGGACCGCCTTGCTGAACGGGACCGAACTGGACGAGCTGTGCATTAACAGCCCGAAG---------------------
139 CC229_146/1-183                ------------CTATATGCGAGCCCCCAGGTGTCTCCAGGAGTTCACTCTCTTTATTCTACAGGTCAG--AATGAGGAATGATGTTGCATTACGATCATGATA-----TTTGCATTGCTGATAAGTTTCTTGATTGAAAAGCTCTGTGTACTGCTTTTTCAGCA------------------
140 CC89_143/1-183                 ---------GCTTCAAATTACCTGAGAGACTTATCTCCTGAAGACAGACCTATTCGGAACAAATGGCAAAATCTCTCGGGGCTTCAAAAGCCGACATGCAAAGATGTCTGGCATTCAAATGTTAGCCACTACTGAGCATGCTGCATGAGGCT-------------------------------
141 CC191_144/1-183                ------------TGATTGTACTCTCGCAGCATATCGAATTCCAGCCAGG--GCTGCTTCCAC---TTCTGTGCTTCCTTCATTTTTTTTCTCGGGTAGCTCAAATTTTATTCCTTTTATGT---TGTATTTCGGCTTCT---CCCACCGTTTTGACCATGGTTGAGG----------------
142 CC54_147_TRIMMED/1-183         -----------------------TCTCAAGTTACCACAGGGCAATAATTCTTACA--TAATAACAACATGCCTGGATTCA---AGATTAAAAAGTGAGCTTTCATGTCAT-TCTGATTGTGTTAAATGGAAAGCCTACAAAGCACCCCACACATAGGTGAC--AGAGAGGTTATTCG------
143 CC180_147_TRIMMED/1-183        ----------------GAAAGGTGGGAAGAGGAATTGGGA--AGAGAAAAAGTGGAAGAGAGAGGAGAGGAGAGGGGAGGAGAGGAGAGGAGAGGAGAGGAGAGGAGAGGAGAGGAGGGGATGGGAGAGGAGAGGAGAGGAGAAAGGAGGGAAGAGGAA-------------------
144 CC109_149_TRIMMED/1-183        ------------CAAAATAGGGGGAAAAAACTGTTTGT-GCTGTTTAGGATCCTTCTAATCCAGGGCAACAAATGCACTCTCCAGTCACCCTGCCATCTTGATGA-G--TTTCTGTGGGTGGTTGGTTGTTTTTTTTTGGACTTTTGACCTGGTATTGTTTG---------------------
145 CC228_145/1-183                --------------AGCATTTGGTCTTATTGAATCTCATCCCATTGGCTTCAGCCCAGCTATCC---AGTCTACCCAGAT-CCCTCTGTAGGGCCTCGCTACCCTCAGGC----AGATAGACACTTCCAGTCAGCCT----GCTGTCAT-CTGCAAACTTACTGAGGGTGCA-----------
146 CC71_145/1-183                 -----------GAATACTCAAGTTTGGAAAAGCTC----AGATTGCACTAAAATGCCAAGATAAACATGAAAAAATAACA----TGCTTCCACTTGCCCCTCCAGCTGTTTCTTAGCTGCTTTGAGCCAGTGTCCAAGCAGCCATGTTGGTGAGATGACATGCA-------------------
147 CC74_145/1-183                 ------------TCCCTCCTTAACATCCTCATTCCCTAAGAATTTCAT----GCAACTTCTTGCAGTAACACTCCTCTCATTTTCTGTGTTAGCTGTTCCCAGATGGAC-ATATCTATTCCT-TATGAGGGAAAAC--CTGATCAAAGTTGGGCACATTTTTTGA------------------
148 CC61_144/1-183                 --------------AGCTTCATGGCCTAGTAAAACTT-TTAAACATGTTTGAG----ACTGTCA---AGGC-ACACAGAT-TTCTAGAAGAACATTTCACACACTCAAGGCTGAAGATAGAAGTTTTCATCTAGTCTT--AACTGCCTTTCTGCATCATTACTTTATTAA-------------
149 CC237_145/1-183                ------------AGCTCATTCACAGCCCTATGCAAATATGGTGGCTTGTGGTATG----CATCTACTACACAGTGTAATGTTCTGCAGAAATACCTGAG-----CTATCCTTGAGTAGGAACAGT-TTTAGGATGAAAAACATGACAGCTGTGGCATTGCAATGCTG----------------
150 CC221_145/1-183                -----------TGGTGGAGTTTTTGCCCAGAGCATTCGTGACTGTCAGAGTCACCTTGTACTTCACTGTTGAGAAGAGC-----TGGAGGTATCTGATATGACATCTGTTTTTGGG--GAAGATGTCCTTCTCACAGACCATCTCTCTCGTGCCATGTCTGGA--------------------
151 CC91_144/1-183                 --------------------TGGTTTAGAGGAGACTCCTTCTCCTCCCTGCACAA----GACATGGAAGCTGTATGGGGCAGCCAC--TGAAGCGAGGCTCCAGCAG-AAGC-AGACGAGCCCCAGACTCCAT-GTTACTAAGCCGACTCCATGAGCAGATCCAATTCAAGCT----------
152 CC104_146/1-183                -------------------TGCATTGATGGACT-CCAAAGCCCCATTGAACTCCTTATTCTCAA--GAGGCTTGGTGGAGGA-----AGGTGTACCGCCTAGGGGAGATGAGGGAAATTCCTTGTAAAAGAACCTTGTAAGAAGGGGTAGATGTGCACATAGTTACAAAAGAA----------
153 CC115_145/1-183                ------------TCAAATGAAGGAGAAAAAATGTCATTTT----CATACTCAGTAGTCACTAGTGTGAAAATGTAGGAAGGCTCTGAGG--TGACCCTTCTAAAGGACCATCTTTAAGATGATGGTT-CTGTTGTGGAATCAACATTAGTTTTGGTTTCAAGGT-------------------
154 CC204_144/1-183                ---------------CTCAAGCAGAGCTTGGACCTCAGTGTGTGGT---CCTCCCCAGTG-AGGTCTGCTCATGTTCCTGGGCCGGGAAATACCATGCCAACTGGATCCC---ACTTGGCAGAGAA--GAAACAATGCCAAGCCAGAGAGCAGGGATGGAAGCGTAGA---------------
155 CC169_146/1-183                --------TGCACAGACAGCAAACGACTCAGCGGTACGTGCTTGGAAGAAAACAGCTGTCCTGCTCAGGAGGTGTGTGGCTGAACAAAGAACTGATGGTCACTGTTTCCAACACACACTTCCCAGAAG-ATGACTAAGCTTTAACAAATTGTGCT----------------------------
156 CC43_146/1-183                 ---------TTCTCTATTCTTCCTGATCGTGTGATCTGATGAACCCAGCACCCTCTTTGTTCTCTTGC--TGGGACCCTTCCTTACTGAAGGCTTT--TTCTTCT-GC-GTTGCATT-TATTTTTGGAGCATAGTTATT----TCTTTTAGCTGAGGATAATTGCT-----------------
157 CC67_145/1-183                 ----------CAGTGGTCCTGGC-CGTAGAGCAACCATGGAGCAACCTAGCTC-AGGAACTTCTTCTGTTTCAAAGAC--ACAGCCTTCT-GGGGCTAGCCTTG--GCCCACACTGGGTCATCTCACATGCCTCTGAGGTGGCACGAGCCAAGGGACAGTGG---------------------
158 CC111_142/1-183                ---------------------TCTCTTCTGTTGCCACCAG-CGAAAATAACTGGC--AGTTTCAGGCCTCGTGCTTTG-----TGATA-ACCTACAATATTGTTGGCAA--TATTCCTCATGGTGGTAGTGGGTGCACCGAGGGAGGAAATAGCAAATGAATCAGCACGGGCAT---------
159 CC193_148_TRIMMED/1-183        ---------------CCTTCTCCCCAGAGCTCCTCTCCAGCAGGTCATGCCCCAGCCTGTACTGATACTTGTGGTTGTTCCTTCCCATGTGCAAGACTCCACGTTTGCTTTTGTTAAACCT---CACCTGGTTTCTTGCTGCCCAGCTCTCCA---GTCTGTCCAGG----------------
160 CC105_146/1-183                -----------------TAGCTGTCACCCAGCCCTAGCAGGGAGGTGTC--------CAGCCCAGGACTAGCAAAAGGCAGAGAAACATTCAGCAGAAGTTGGAAGTA---TGAGA---TTTGGAGCTTCAG-CATCTTTTGATTCGAGGAGGAAAGAACAG-CTTACATGTGTCAGGT----
161 CC117_145/1-183                ----CGTCGACGTGCAACTTAGCTGATGTAACTTATGGGAGGAGTAGG--CTCCTAAATGAGCTGCTCCTGTGTGCTCTGAAGATGGTTCATTTGAACCATTTTTACCTACTCTAAGGTGTT--TGGTCAGCCAAGTGCTGTCTGACTGAAGA------------------------------
162 CC175_145/1-183                ---------------------------ACTCCAACAGCTTCACAGTAACAATTCTAATGAAAAAGC---TTCTTCAGAACATATTCAGTAAATGACAGACTGAGAATGGCT--TGGCTATGCAACCACTCAC-GAAGGCCAGGAACACTCACTGCAAAAATTTCCAGCAAGCTCTTTG-----
163 CC157_142/1-183                ---------------CTTTCAGGCCACATGGCCCCAACACCAGATATTGCA--AGGCAGCTATGCAACACCC--CTAACCCTTCCCCAG-ACAGCCCTTCATCCATGCTTTTATCTCACCG---TAGCGCTTATTAAGGAAACAGGCTGTGAAAATGTGTTTCCA------------------
164 CC210_146/1-183                ---------CCACCAACAAGAGCCTGACTCTGCCCTCACTGTACTTTCCT-TCAGTGATTCATTGACACGACTGGGATC---CCCTGAGCCTCCTTTTCTCCAGGCTGAACAGCCTCAGCTCTGTCAGCTTCTCCTCATAGCAGAGGCATTTCAGGCCC------------------------
165 CC160_148_TRIMMED/1-183        ------------------GAGGAGAGGTGTAGAGCTGGGGAGAATTATCCCATCCTTGATGATGTGCACTTGCC--AGTGCACACCTGCATCACC--TGGGGCTCAGCAGGCTCC-----TCCATTTCTTCGTG-TGCTCCCTTTCTTCCCCCAGTTTAGAGTGGGATTACTTG---------
166 CC37_144/1-183                 ---------------------AGGCAAAAGCCCAGCTTGAGCT-CAACCTGGCTGCTGGGGTAAAAGGGAACAAGAAACTCTTTTACAAGTATAT--CTAGAGTAAGAGGAGGA-CCAGGGAGAATCTCCATTCTCTACTGGATGA--GGCTGGGAACGTGGCCACTGAGG------------
167 CC57_143/1-183                 ------------------------------CTGGTCACTTGCAGCCAACTCCTTTTCGTACCCACCTACAACTCTTACACCAG-TCCTATCTTATCCAGTTCAACAAGAACACTTC--CCATTTGGTGTGGTGTGGAGGTACTGTAGCATACGTAGAAAAAACCTCACGCTCACCT-------
168 CC38_147_TRIMMED/1-183         ----------TCCAACACCCTCTGAACAAGCAGGATGGCATGCTGTGAACAGACCTGTTTATTAAGCACAAA--GGAGTTTCCAGGATTTGCTCTCAGACAGGTCAGGAGGTTTCCTCCAGAAGAGCCAGAAACATGCTGTTCTGGTGAGACTGTGCA-------------------------
169 CC214_142/1-183                ---------------TGCCTTTAGCAGTCTCCTTCCAGCCTGCAGCCCTGCAGTGGGG-CACGGTG-GGCTCGG--GGCTGCAGACCCACAGGCAGGCACTCCCTAGTCTGCCAAATAATAGCCTGG---AAGAAGGGGGGAGTGAGGCTGCAGCTATTGCACG-------------------
170 CC44_144/1-183                 -----------CACAGCAGGATATCT-CCAGATACACACAATGCTCTC---TCTGGGTCCTCTTCCTGCGCCTCTCACAAGAGACTGGGCCACCGCTCCCACCGCCGCCACAGGGGTGTAGGTGGCAACAACTGTGCATCACACCGAGGAGTGTCACCT------------------------
171 CC10_146/1-183                 --------------CGATGCGCTCCATGAACGGCTCGCGGAGGAACAGCGGCTCCTCGT--TGGTCTCCAGCTCTTCGGCCTCCTCCAACCTCAGCCACCTGGAGGAGGACACCTGGATCCTCTGGGGCCGCATCGTCAACGAGTGGGACGAGTGGAGGAAG---------------------
172 CC28_145/1-183                 --------------------------TGTTTGACAAGCCTGAGCTGGCTC---CCAGATGTTTAAATTGTCCTCAGCCTGTG-ATAATTTATTTATCTGTTTATGTG-TGTTACCAGAAAAAGTATCTCTCAGACTCTTGACTCATTTGCAGGGGAATCCAGAGGGAACAGTAGCA-------
173 CC146_144/1-183                -------------------TGTAAAATAATCTG-GTGGTATGC-----TGTGATGGAAAACCAG---ATGAAAGAAGATGGCT-----CACGTTCATTTAGGGGGGGATGGACTTACAGAAAC--TGGACTTGGGGGGCATGTATGATGGTTTTTTGATAGGCACCGAGTGCTGAAGGT----
174 CC49_147_TRIMMED/1-183         --------------GCGCCAGCTTCTG-CTCCTGGAGGCGCTTCTGCCTCTGCCGGTC-CATGTTGCGGCTCAGCAGGTTGGTGACGGTCATGCGGGGCCCGGCCAGCTCG--AAGTGGTAGCCCAGCTTGAGCAGCGTGGTGTTCTCCTTCAGCAGCTTGGCG-------------------
```

```
175 CC126_146/1-183             --------------------------TTTTCCCACTAAATCAAAGTGACA-TTCCCCTAAGGATGA---AATTATACAACCTA------AATTACTTGAATCATGATGGCTCCTAGTTTCACCACTGCTTGC-CAAATTTTGTACAAATGCAGTCATTTAACTTCCTCAGCCTTTCTTAAGGA
176 CC158_144/1-183             ------------TGCCCAGTTGGTCTCTAAAAACCGTGGGTACGTAG-GCA---GGCAGCTTCTCTACCC--------TCAGCTTTGCGCCTGGCACTCCCCAAAACCTGCCAAGCTCCGC---TTTCTCTCAGAGTTATCTCAGAG-AGCTTTAGAATGGCTGAGGTCTCA----------
177 CC60_145/1-183              --TCCCCACCCCACGATGCTAGCCCC-CCATGTGCAGGCAGTGCCATGCGGTGTGGGCATTATTCCTGCCGACACCACA--------------CTCTCCAACTGAGAGAACTGGAGAACCCATCCCTACTGACCCATCCCGTGCAAAGCGCGGCCCCATCAG--------------------
    consensus/100%              ......................................................................................................................................................................................
    consensus/90%                        ...................................................................................................................................................................
    consensus/80%                          ..............................................................................................................................................................
    consensus/70%                            ...........................................................................................................................................r..............
MView 1.47.3, Copyright © Nigel P. Brown, 1997-2002.
```

LEVITSKY NUCLEOSOME DATASET

```
Identities computed with respect to: (1) NM0014/1-331
Colored by: consensus/65.0% and property
                     1 [        .         .         .         .         :         .         .         .         .         1         .         .         .         .         :         .         .         .         .         2         .         .         .
  1 NM0014/1-331       --------------------------------------------------------------------------------------------TGAAGGACCTGGAA-TATGGCGA--GAGAACTGAAAATCACCGAAAATG--AGA-AATACACACTTTAGGACG-----TGAAATATGGCG---AGGAAAACTGAAAAAGATGGAAAATTTAGAAATATCCACTGTAGGA
  2 NM0006/1-331       ----------------------------------------------------------------------------------------------AGGGACATGGAA-TATGGAGA--GAAAACTGAAAATCACGGAAAATG--AAA-AATACACACTTTAGGACG-----TGAAATATAGCG---AGGAAAACTGAAAAAGGTGGAATATTTAGAAATGTCCACTGTAGGA
  3 NM0004/1-331       --------------------------------------------------------------------------------------------TGAAGGACCTGGAA-TATGGCGACGGAAAACTGAAAATCACGGAAAATG--AGA-AATACACACTTTACGACG-----TGAAATATGGCG---AGGAAAACTGATAAAGGTGGAATATTTAGAAACGTCCACTGT----
  4 NM0034/1-331       ----------------------------------------------------------------------------------------------TAGGACCTGGAG-TATGGCGA--GAAAACTGAAAATCACAGAAAATG--AGA-AATACACACTTTAGGATG-----TGAAATATGGCG---AGGAAAACTGAAAAAGGTGGAAAATTCAGAAATGTCCACTGTAGG-
  5 NM0016/1-331       ----------------------------------------------------------------------------------------------GAGGACCTGGAA-TATGGTGA--GAAGACTGAAAATCACGGAAAATG--AGA-AATACACACTTTTGGACG-----TGAAATATGGCG---AGGAAAACTGAAAAAGGTGGAAAATTTAGAAATGTCCACTGTAGGA
  6 NM0033/1-331       ---------------------------------------------------------------------------------------------AGAGGACCTGGAA-TATGGCGA--GAAAACTGAAAATCACGGAAAATG--AGA-AATACACACTTTAGGACA-----TGAAATATGGCG---AGGAAAACTGAAAAAGGTGGAAAATTGAGAAATG-CCACTGTAGGA
  7 NM0010/1-331       ------------------------------------------------------------------------------------------------TGACCTGGAA-TATGGCGA--GAAACCTGAAAATCACGCAAAATG--AGA-AATACACACTTTAGGACA-----TGAAATATGGTG---AGGAAAATTGAAAAAGGTGGAATATTAAGAAATGTCCACTGTAGGA
  8 NM0019/1-331       -----------------------------------------------------------------------------------------------AGGACCTGGAA-CATGGCGA--GAAAACTGAAGATCACGGAAAATG--AGA-AATACACACTTTAGGGCG-----TGAAATATGACG---AGGAAAACTGAAAAAGGTGGAGAATTTAGAAATGTCCACTGTAGGA
  9 NM0038/1-331       -----------------------------------------------------------------------------------------------AGGACCTGGAA-TATGGCGA--GAAAACTGAAAATCACGGAAAATG--AGA-AACACGCGCTTAAGGACA-----TGAAATATGGCG---AGAAAAACTGAAAAAGGTGGAATATATAGAAATGTCCACTGTA---
 10 NM0054/1-331       -------------------------------------------------------------------------------------------------GACCTGGAA-AATGGCGA--GAAAACTGAAAATCACGGAAAATG--TGA-AATACACACTTTAGGACA-----TGAAATATGGCG---AGGAAAATTGAAAAAGTTGGATAATTTAGAAATGTCCACTGTAGGA
 11 NM0009/1-331       ----------------------------------------------------------------------------------------------------CCGGAA-TATGGCGA--GAAAACTGAAAATCACGTAAAATG--AGA-AATACACACTTTAGGACG-----TGAGATATCGCG---AGGAAACCTGAAAAAGGTGGAAAATTTAGAAATG-TCACAGTAGGA
 12 NM0042/1-331       ----------------------------------------------------------------------------------------------AAGGACCTGGAA-TATGACGA--GAAAACTGAAAATCACGGAAAATG--AGA-GATACACACTTTAGGACG-----TGAAATATGGCG---AAGAAAACTGAAAAGGTCGGAAAATTTAGAAATGTCCACCGTAGAC
 13 NM0049/1-331       -----------------------------------------------------------------------------------------------AGGACCTGGAA-CATGGTGA--GAAAACTGAAAATCACAGAAAATG--AGA-AATAGACACTTTAGGACG-----TGAAATATGACG---AGGAAAACAGAAAAAGTTGGAAAATTTAGAAATGTCTAACGTAGGA
 14 NM0041/1-331       ------------------------------------------------------------------------------------ATGCACACTGTAGGACCTGGAA-TATGGCGA--GAAAACTGAAAATTAAGGAAAATG--AGA-AATATACACTTTAGGACG-----TGAAATATGGCG---AGGAGGACTGAAAGAGGTGGAAAATTTAGATACGTCCAC-------
 15 NM0022/1-331       ------------------------------------------------------------------------------------------------------GGAA-TATGGCGA--GAAAACAGAAAATCACGGGAAATG--AGA-AATACACACTTTAGGACG-----TGAAATATAGCG---AGGGGAACTGAAAAAGGTGGAAAATTTAGAAATGTCCGCTGTAGGA
 16 NM0015/1-331       ---------------------------------------------------------------------------------------------------------AA-TGTGGCGA--GAAAAGTGAAAATCACGGAAAATG--AGA-AATAAACACTTTAGGAAG-----TGAAATATGGCG---AGGAAAACTGAAAAGGATGGAAAATTTAGAAATGTCCACTGTAGGA
 17 NM0017/1-331       ----------------------------------------------------------------------------------------------------------------AGA--GAAAACCGAAAATCACGGAAAATG--AGA-AATACGCACTTTAGGACG-----TGAAATATGGCG---AGGAAAACTGAAAAAGGTGGAAAATTTAGGGATGTCCACTGTAGGA
 18 NM0013/1-331       ----------------------------------------------------------------------------------------------TAGGACGTGGAA-TATGGCAA--GAAAACTGAAAATCATGGAAAATG--AGA-AACATCCACTTGATGACT-----TGAAAAATGACG----AAATCACTAAAAAACGTGAAAAAT-GAGAAATGCACACTGAAGGA
 19 NM0039/1-331       ------------------------------------------------------------------------------------------------------------ATGGCGA--GAAAACTGAAAATCACGGAAAATG--AAA-AATACACACTTCAGGACG-----TGAAATATGGCG---AGGAAAACTGAAAAAGGTGGAAAATTTAGAAATGTCCACTGTAGGA
 20 NM0070/1-331       -----------------------------------------------------------------------------------------------AGGACGTGGAA-TACGGCAA--GAAAACTGAAAATCATGGAAAATG--AGA-AACATCCACTTGACGACT-----TGAAAAATGACG----AAATCACTAAAAAACGTGAAAAAT-GAGAGATACACACTGAAGGC
 21 NM0046/1-331       ------------------------------------------------------------------------------------------------------------------A--GAGAACTGAAAATCACCGAAAATG--AGA-AATACACGCTTTAGGACG-----TGAAATATGGCG---AGGAAAACTGAAAAAG-TGGAAAATTTAGAAATGTCCACTGTAGGA
 22 NM0012/1-331       ----------------------------------------------------------------------------------------------TAGGACGTGGAA-TATGGCAA--GAAAACTGAAAATCATGGAAAATG--ATA-AGCATCCACTTGACGACT-----TGAAAAATGACG----AAATCACTAAAAAACGTGAAAAAT-GAGAAATGCACACTGGAGGA
 23 NM0053/1-331       ------------------------------------------------------------------------------------------------------------ATGGCGA--GAAAACTGCAAATCACGGAGAATG--AGA-AATACACACTTTAGGACG-----TGAAGTATGGCG---AGGAAAACTGAAAAAGGTGGAAAATTTAGAAATGTCCGCTGTAGGA
 24 NM0069/1-331       ----------------------------------------------------------------------------------------------TAGGACGTGGAA-TATGGCAA--GAAATCTGAAAATCATGGAAAATG--AGA-AGCATCCACTTGATGACT-----TGAAAA-TGACG----AAATCACTAAAAAACGTGAAAAAT-GAGAAATGCACACCGTAGAC
 25 NM0037/1-331       -----------------------------------------------------------------------------------------------AGGACGTGGGAGTATGGCAA--GAAAACTGAAAGTCATGGAAAGTG--AGA-AACATCCACTTGATGACT-----TGAAAAATGACG----AAATCACTGAAAGACGTGAAAAAT-GAGA-ATGCACACTGTAGGA
 26 NM0011/1-331       ------------------------------------------------------------------------------------------------------------ATGGCGA--GAAAACTGAAGTTCACGGAAAATGG-AGA-AATACACACTTTAGGATGCAG--TGAAATATTGAGCGAAGGAAAACTGAAAAAGGTGGAAAATTTAGAAATGTCCACTGTAGGA
 27 NM0031/1-331       ------------------------------------------------------------------------------------------------------------------A--GAACACTGAAAATCACGGAAAATG--AGA-AATACACACTTTAGGACG-----TGAAATATGGCG---AGGAAAACTAAAAAAGGTGGAAAATTTAGAAATGTCCACTGTAGGA
 28 NM0024/1-331       ---------------------------------------------------------------------------------------------TGACGACTTGAAA-AATGACGA--GATCACTAAAAAATGTGAAAAATG--AGA-AATGCACACTGAAGGAGC-----TGGAATATGGTG----GGAAAACTGGAAATTACGGAAAAT-GAGAAATACACACTTTAGGA
 29 NM0002/1-331       -----------------------------------------------------------------------------------------------AGAACGTGGAA-TAAGGCAA--GGAAACGGAAAATCATGGAAAATG--AGG-AACATCCAATTGACGACT-----GGAAAAATGACG----AAATCACTAAAAAACGTGAAAAAT-GAGAAATGGACACTGAAGGA
 30 NM0005/1-331       -----------------------------------------------------------------------------------------------AGGACGTGAAA--TTGGCGA--GGAAACTGGAAAG-GTGGAATATTT-AGA-GATGTGCACTGTAGGACG-----TGGAATATGGCA-----AGAAAGTGAAAAGCATGGAAAAAT-GAGAAACATCCACTTGACGA
 31 NM0048/1-331       ------------------------------------------------------------------------------------------------------------------A--GAAAACTGAAAATCACGGAAAATT--AGA-AATACACACTTTAGGACG-----TGAAATATGGCG---AGGAAAATGGAAAAAGGTGGAAAATTTAGCAGTGTCCACTGTAGGA
 32 NM0021/1-331       ------------------------------------------------------------------------------------------------------------------A--GAAAACTGAAAATCACGGAAAATG--AGG-AATACACACTTTAGGACG-----TGAAATATGGCG---AGGAAAACTGAGAAAGTGGGAAAATTTAGAAATGTCCACTGTAGGA
 33 NM0072/1-331       -----------------------------------------------------------------------------------------------AGGACGTGCAA-TAAGGCAA--GAAAACTGAAAATCATGGAAAATG--AGA-AACATCCACTTGACGACT-----TGAAAAATGACA----AAATCACTAAAATACGTGAAAAAT-GAGAAATGCACCCTGAAGGA
 34 NM0050/1-331       -----------------------------------------------------------------------------------------------AGGACGTGCAA-TAAGGCAA--GAAAACTGAAAATCATGGAAAATG--AGA-AACATCCACTTGACGACT-----TGAAAAATGACA----AAATCACTAAAATACGTGAAAAAT-GAGAGATGCACCCTGAAGGA
 35 NM0086/1-331       ------------------------------------------------------------------------------------------------------------ATGGCAA--GAAAACTGAAAATCATGGAAAATG--AGA-AACATCCACTCGACGACT-----TGAAAGATGACG----AAATCACTAGAAAACGTGAAAAAT-GAGAAATGCTCACTGAAGGA
 36 NM0023/1-331       --------------------------------------------------------------------ACGTGAAAAATGAGAAATGCACACTGAAGGACCTGAAA-TATGGCGA--GAAAACTGAAAATCACGGAAAATG--AGA-AATACACACTTTAGGACG-----TGAAATATGGCG---AGGAAAACTGACAAAGGCGG--------------------------
 37 NM0003/1-331       ------------------------------------------------------------------------------------------------------------ATGGCTA--GAAAACTGAAAATCATGGAAAAAG--AGA-AACATCCACTTGACGACT-----TGAAAAATGTCG----AAATCACTAAAAAACGTGAAAAAT-GAGAAATGCACACTGAGGGA
 38 NM0080/1-331       -----------------------------------------------------------------------TGGAAAATTAGAAACATCCACTTGATGACTTGAAT-AATGACGA--AATCACTAAAAAACGTGAAAGATG--AGA-AATGCACACTGAAGGACC-----TGGAATATGGCG----AGAAAACTGAAAATCACGGA-------------------------
 39 NM0047/1-331       -----------------------------------------------------------------------------------------ATTTGTAGAACAGTGTATATCAATGAGTTACAATGAAAAAACATGGAGAATGATAAA--TACCACACTGTAGAACATA---TTAG----ATGAGT--GAGTTACGCTGAAAAACACATACGTTGGAAACCGGCATTG-----
 40 NM0043/1-331       -------------------------------------------------------------------------------------------ATGTAGGACAGTGTATATCAACGAGTTACAATGAGAAA-CATGGAAAATGATAAA--AACCACACTGTAGAACAGA---TTAG----ATGAGT--GAGTTACACTGAGAAACACATTCGTTGGAAACGGC----------
 41 NM0008/1-331       -----------------------------------------------------------------------------------------ATTTGTAGAACAGTGTATATCAATGAGCTACAATGAAAAT-CATGGAAAATGATAAA--AACCACACTGTAGAACATA---TTAG----ATGAGT--GAGTTACACTGAAAAACACATCCGTTGGAAACCGGCAT-------
 42 NM0020/1-331       -----------------------------------------------------------------------------------AACAGGATTTGTAGAACAGTGTATATCAATGAGTTACAATGAGAAA-CGTGGAAAATGATAGA--AACCACACTGTAGAACATA---TTAT----CTGAGT--GAGTTACACAGAAAAACACATTCGTTGGAAACGGG----------
 43 NM0079/1-331       --------------------------------------------------------------------------TTACACAGGAAACAGCTCGGGATCCGCCCGGGCTAGAGCGG------CCGCCACCGCGTGGAGCTCCAGCTT-TTGTTCCCTTTAGTGAGGGTTAATTGCGCGCT-TGCGTAATCATGGTCATAGCTGTTTCCTATGTGAAATTGTTATCCGC----
 44 NM0078/1-331       ---------------------------------------------------------------------------------------AAGACAAAATATGCACGATGTCACATGCAGGACCGCCGATTGTATTGATACCATTACGTTATGCGTGGACGTCGGCTGT---AGTCCTAAGCGCACCCCG--ACCGAGTTCTGT-GTACGAAAC--CTACCAGCTCCTTCGACC
 45 NM0030/1-331       ---------------------------------------------------------------------------------------CAAACGACCAATACGCCCGTTATCTTAAG--ACTTATCGAGTCATTC----GCTGGTTA-AACTATCACATGCA-ACC---AATGTTTTGTTGGCGTGGTCTCAATGCCCTCTCAACCAATATAATGGGCTCCGACCCACTGTT
 46 NM0084/1-331     ------------------------------------------------------------------------------------------GCTCGCACACCCGACATAGGGCATCGTAACGCCTGTAATAGCTGACACCCG--CTA-CGTA-CGCTATTCGCGC---CGCTGCCCTGATTACCACTGCTCCCACGCCTCTCGGT------AGGCCTACGCCTTTGCTTGTC
 47 NM0051/1-331     --------------------------------------------------------------------------------------------CGACGACTACAGCCGCTATTCAGGTTACACCCAAACCCGAGTCATTCGTGCCTGCTACCAGCTC-CATTGC-TACACGCGGCTAA-----CCCCTGATGACGGAGT--ATTTGCTGCGCAGCTCCCAT-TGCCCCACC
 48 NM0087/1-331     -----------------------------------------------------------------------------AGGGAGGGTTCTGACCTCAGAACTCAGGAGG--TGGATCAGAGCCCCAGACTGCTAGACACATGCCCTGAAAGAGGAAAGCTTGCCTGCAG-AGAGTGCTCTGACCA-------CTGGAACTCAGGAAAGAGC---TAGT-------------------
 49 NM0028/1-331     --------------------------------------------------------------------------------------------------------TTGGTCAC-TGTGTTAAGTAGGGTGTGGTGGCATTCACCTGTAGTCCTAACATTCGTGAGGTAGAGGTAGGAGGGTC---AGAAATTCAAAGTTATCTTCAGCTGCTCTATAGTTCAAGGCCGGGGC
 50 NM0083/1-331     --------------------------------------------------------------------------GAGGACAAGAGCACCTAACTACACT---AAAGCCGGACCGTTGGCGCTCACCCTGTGGTG--ATCAATCTCCCACCACGCTTTCCACCTGACAGCG---CAGAGTATCCCAGTCAATATAGTTCCGCATCAAGACAGAAACGTCAATGACCC----
 51 NM0085/1-331     -------------------------------------------------------------------------------AGCAGCAGCAGCAACAGTAGTAGAAGCAGCAGCACTAACGACAGC-ACAGCAGTAGCAGTAATAGAAGC-AGCAGCAGCAGTAGCAGTAGCAGCAGCAGCAGCAATTTCAACAAC-AGCAGCAGCAGCT-----------------
 52 NM0064/1-331     -------------------------------------------------------------------------ACTCAAAGAACAAAGATCCTGCTAACCACGG-CATTAGGGAACGG-GCGGTACCGATGCCGTTCTGGT--CGACAGCGCATAGCCC-CGGTCCAACTCCGTGCGGCCTAGAACGTTACGTACCCTAGATGC-AGCGGAACTCTTGCGTGTCG-----
 53 NM0027/1-331     -------------------------------------------------------------------TACACACACTAACACACACACATGCACACATACACACAGACACATGCACATATACACACACATACACACGCATACACACACATACACACACATATACACACACATGCA-CACTTACACACACATGCACACACAC------------------------------
 54 NM0056/1-331     ----------------------------------------------------------------------------CTTCCTCATGCATGAGCTTGCATGAGCTTGCATATGCTCACATACCACACATGTGAGTCTACACACAATG-AGC------ACACACACACACACACACATCACTAACCGTCTCGGTCTGGCCATCATAGTCTGGC-----------------------
 55 NM0067/1-331     ------------------------------------------------------------CAC-TGTGACAACAATGTGGAACATTGTGACATCACAGTGGCGAACAGCGACAAAACAGTAAAGGAGTCTGACAGCACAGTGGAAAACAGTGACAGAACTGTG-GAGCACTGTGATTGCACCATGGAGCA-TGTTACACCAC----------------------------
 56 NM0026/1-331     -----------------------------------------------------------------TACACACACCACATCATGCATACACACACATCAATGCA-ATGCATACACACATACATACACATACTAACACATACACTCACACACACGCAGAAAT---TATGCATGCATCATCGACATTGGCACGCA------------------------------------
 57 NM0073/1-331     ------------------------------------------------------------CAC-TGTGACAACAATGTGGAACATTGTGACATCACAGTGGCGAACAGTGACGGAACAGTAAAGGAGTCTGACAGTACAGTGGAAAACAGTGACGGAACTGTG-GAGCACTGTGATTACACCATGGGACA-TGTTGCACCAC----------------------------
 58 NM0061/1-331     -------------------------------------------------------------------------------------------TGTGAACAACCAATCAACGGTGGCAGTGCAGCATGG----TCTA-TCAGGTTGTA-CAGGCCAGAGCGAGACTAAAAT--CAATTCC------ACACAAACCCTCTTACCAA----CGTTAGGACCATGATCT--CTCGC
 59 NM0040/1-331     ------------------------------------------------------------------------------------------CAACGCTCACCTGGTCCGGACCCTCG--ACGCCTCTATCCACTTCCACCTAGCCAA-TGGA-CGCTCGACGAGCTTACAGCTCCGCGCGCACTCCTAATCTGTAACCTTAAACT------GCGCATTGGCCCCGATTCCAT
 60 NM0029/1-331     --------------------------------------------------------------------------------------------CCCTAAGCTGTCAGGAGCTTCTGATACCACCGGCCTTAGTGCCAGCTGTGCCAT-CGC--CGGTGCTTC--GACGCTGCTTGTCCGCGGCATAACTTACTATCATG--CAGCACACCGTTAATCGCTTTCCTTTTGCTT
 61 NM0032/1-331     --------------------------------------------------------------------------------------------CCCTAAGCTGTCAGGAGCTTCTGATACCACCGGCCTTAGTGCCAGCTGTGCCAT-CGC--CGGTGCTTC--GACGCTGCTTGTCCGCGGCATAACTTACTATCATG--CAGCACACCGTTAATCGCTTTCCTTTTGCTT
 62 NM0071/1-331     ------------------------------------------------------------CAC-TGTGACAACAAAGTGGAACATTGTGACATCACAGTGGCGAATAGCGACAAAACAGTAACGGAGTCTGACAGCACGGTGGAAAACAGTGACAGAACTGTG-GAGCACCGTGGTTGCACCATGGAGCA-TGTTACGCCAC----------------------------
 63 NM0058/1-331     ------------------------------------------------------------CAC-TGTGACAACAATGTGGAACATTGTGACATCACAGTGGCGAACAGTGACAAAGCAGCAAAGGAGTCTGACAGCACAGTGGAAAACAGTGACAGCAGACTGTGAGCACAGTGATTGCACCATGGAGCA-TACTACACCAC----------------------------
 64 NM0035/1-331     ----------------------------------------------------------AGCAC-TGTGACAACACAGTGGAGCAGCTTAACACCACAGTGTAGCACTATGACATCAGAGTTGAGCACTGTGTCACCACTCAGAGAAC-TATGACACTACAGTA-GAGCACTGTAACATCACAGTCGAGCACTGTAACACCACATATGCGCAT-------------------
 65 NM0074/1-331     ------------------------------------------------------------CAA-TGTGACATCACA-TGTAGCATGGTGAAATCCCAGTGGAATACTGTGACACCACATTGGAGCACAGTGACGCCACAGTGGAGCA-TGTGGCACCACAGTG-GAGCACTGTGAAACCACAGTGGAGCACTGGG------------------------------------
 66 NM0065/1-331     ------------------------------------------------------------CAA-TGTGACATTACA-TGTAGCATGGTGAAATCCCAGTGGAATACTGTGACACCACATTGGAGCACAATGACACCACTGTGGAGCA-TGTGACACCACAGTG-GAGCACTGTGAAACCTCAGTGGAGCACTGGT------------------------------------
 67 NM0075/1-331     ---------------------------------------------------------------CCAGATAGCATTCCGGCTC-CCTAACGACTGTGACGCTGGTCTGTGCAGCAACG-----CCAATGACTTCACACCAATTGCTT-TCCTGCTCTACCCA-GATGTACAGATGTG-GTCTTCAAAAGTAACCTTGTTCGTGCGCAATACTGCCCG---------------
 68 NM0066/1-331     ---------------------------------------------------------------GCAGTGCTCACAT-ACAGCGCACACATACAGTGCTCCATATA-GTGCACACATACA-GTGCACACATACG-GTGCTCACGTACA-GTGCTCACATACAG-TGTACACATACAGTACACACATACAC-TGCACA-----------------------------------
 69 NM0059/1-331     ---------------------------------------------------------GAGCACCTGTGACACCACAAGGGGGCCTTGTGACTGCACAGAGGGGCACTGTGTCACAACAGTGGAATGCTGTGACAGTACAGTGGAGCAGTGTGACAAAACAGTG-GAGTACTGTGACACAATAGTAGGGCAATATGAC----------------------------------
 70 NS0002/1-331     -----------------------------------------------------------------------------------------------------------------------AAACTGAAAATCACGGAAAATG--AGA-AATACACACTTTAGGACG-----TGAAATATGGCG---AGGAAAACTGAAAAAGGTGGAAAATTTAGAAATGTCCACTGTAGGA
 71 NM0060/1-331     --------------------------------------------------------ATGCGCAGACGCACACACATGAGCATGCGCAGACGCACATACATGAGCGTACGCAGACGCACACACATGAGCATGCGCGCGCG-------ACACACACACACACACACACACACACACG--AGTGGCAAGGCGGGGG-------------------------------------
 72 NM0068/1-331     --------------------------------------------------------ATGCGCAGACGCACACACATGAGCATGCGCAGGCGCACATACATGAGCATACGCAGACGCACACACATGAGCATGCGCGCGCACACACACACACACACACACACACACACACACACACACG--AGTGGCAAGGCGGGG--------------------------------------
 73 NM0063/1-331     --------------------------------------------------------ATGCGCTGACGCACACACATGAGCACGCGCAGACGCACATACATGAGCGTACGCAGACACACACACATGAGCATGCGCGCG----------ACA-ACACACACGCTCACACACACG--AGTGGCAAGGCGGGG--------------------------------------
 74 NM0036/1-331     ------------------------------------------------------AGGTCTATAAGCGTCTATAAGCGTCTATG--AAC-GTCTATAAAC-GTCTATAAACGCCTATAAACGCCTATAAACGCCTATACAAGCCTA----TAAACGCCTATACACGTCTATG-CACGACTATACACGTCT------------------------------------------
 75 NM0044/1-331     ------------------------------------------------------AGGTCTATAAGCGTCTATAAACGTCTATA--AAC-GTTTATAAAC-GTCTATAAACGCCTATAAACACCTATAAACGCCTATACAAGCCTA----TAAACGCCTATACACGTCTATA-CACGCCTATACGCGTCT------------------------------------------
 76 NM0018/1-331     ------------------------------------------------------AGGTCTCTAAGCGTCTAAAAACGCCTATA--AAC-GTTTATAAAC-GTCTATAAACGCCTACAAACGCCTATAAACGCCTATACAAGCCTA----TAAACGCCTGTACACGTCTACA-CACGTCTATACACGTCT------------------------------------------
 77 NM0062/1-331     -----------------------------------------------------------GAGAGTAACATAGGCACAGGTGTGGAGAGTAACACAGGCACAGGT--GTGG-AGAGT----ACACACAGGCACAGGCGTGGA-GAGTACACACAGGCACAGG-TGTGGAGAGCACAC-ACAGGCACAGGTGTGGAGAG---------------------------------
 78 NM0045/1-331     ------------------------------------------------------AGGTCTATAAGCGTCTATAAACGTCTATA--AAC-GTCTATAAGC-GTCTATAAACGCCTATAAGCGCCTATAAACGCCTATACGAGCCTA----TAAACGCCTATACACGGCTATA-CACGTCTATACACG---------------------------------------------
 79 NM0082/1-331     -----------------------------------------------------------GAGAGTAACACAGGCACAGGTGTGGAGAGTAACACAGGCACAGGT--GTGGGAGAGTG---ACACACAGGCACAGGTGAGGA-GAGTACACACAGGCACAGG-TGTGGAGAGCACAC-ACAGGTGCGGAGAG----------------------------------------
 80 NM0077/1-331     -----------------------------------------------------------GGGCTGTAGAATCTGATGGAGGTGTAGGATGGATGGACAGTATGACAAAAG-----GG--TACTAGCCTGGGACAGCAGGATTGGTG-GAAAGGTTACAGGC-AGGCCCAGCAGGCTCGGACGCTGTATAGAG--------------------------------------
 81 NR0013/1-331     --------------------------------------------GGAATCCCAACAATTACATCAAAA---TCCACATTCTCTTCAAAATCAATTGTCCTGTACTTCCTTGTTC-ATGTGTGTTCAAAAACGTTATATTTATAGGA-TAATTATACTCTATTTCTCAACAAGTAATTGGTTGTTTGGCCGAGCGGTCTAAGGCGCCTGATTCAAGAAATATCTTGACCGCA
 82 NM0081/1-331     ----------------------------------------------------------TAGACCAGGTGAGCAGGAGGCGGACAGCAGGGAA------CAG--TCTGGAGGGCAGGAAAGAGCTCTGAGGAG------CCATAGCGGGTAAAGCTG----------AGGATGGGTTT--AAGC---AAAAGCCAGA--CCAAGGACAGGAGGATGTGCACACTGC-----
 83 NP0017/1-331     ---------------------------------TTCTAGAATCAAACGTACCACAAACAGATTCAGGAATACTCGGAATTCAGTATAAACTAAAGCAACTTTTTAAAATTAG-TAGGGAGAACCAGGTGTGGTGGTACACACCTTTAATTCCAGCACATGGGAGGCAGGGGCAGGCAGATCTTTGTGAGTTCAAGGGCAGCCTGGTCTACATGGCAAGTTCCAGGCTAGCC
 84 NG0041/1-331     --------------------------------------------------------TCTTACTACAATTTTTTTGTCTAAAGAGTAATACTAGAGATAAACATAAAAAATGTAGAGGTCGAGTTTAGATGCAAGTTCAAGGAGCGA-AAGGTGG--ATG-GGTAGGTTATATAGGGATATAGCACAGAGATATA--TAGCAAAGAGATACTTTTGAGCAATGTTTGTGGAA
 85 NP0020/1-331     -------------------------------------------------------ATCTGAAAGTTTCCCCATGTCCAACAAGACTAGAACAAACAAGTCCTGCGTAGTCGCCTGTCGGTTTCTGGGTGTGGTGGTATAGCCCTGTAATCCCAGCATTTGGGAAGCTGAGGTGGGAGGATC---GGGAGTTCAAGGTCAGCTTGGGCTACTTAGAAAGACCTTGTCTCAAA
 86 NR0023/1-331     ---------------------------------------------------GTCTTTCTTAGAAAACGTTGTTAGAGACGCTGTCACTTACACTGAACACGCTAGAAGAAAAACCGTCACTGCTATGGA-CGTTGTCTACGCCCTC-AAGAGACAAGGCAGAACTCTCTATGGTTTCGGTGGTTGAACAAAATATTTATCTTAAAAAATTAAAAAGTAAAAAGCTGCATGC
 87 NG0018/1-331     -----------------------------------------------GGTTCCAGCTGACACCGAA-ATGGAAGAGGTAGATTAGGTAGAACATCATGGCCTTGAATAGGTTATAAACAAAACATAATATAACGTATAGGTA--TT-CGAAT----GAATAAATAAGTATGTAAATAGGGCATCTGCATGGAAATA-ACTGGGTAAAACATTACAATGTATTTTTTTAGAC
 88 NP0010/1-331     ------------------------------------------------------TGGTCCCTGTTTTCGAAGAGATCGCACATGCCAAATTATCAAATTGGTCACCTTACTTGGCAAGGCATATACCCA-TTTGGGATAAGGGTAA-ACATCTTTGAATTGTCGAAATGAAACGTATATAAGCGCTGATGTTTTGCTAAGTCGAGGTTAGTATGGCTTCAT-CTCTCATGA
 89 NP0005/1-331     --------------------------------------------------TATTACCTTCTGCTCTCTCTGATTTGGAAAAAGCTGAAAAAAAAGGTTTAAACCAGTTCCCTGAAATTATTCCCCTACTTGACTAATAAGTATATA-AAGACGGTAGGTATTGATTGTAATTCTGTAAATCTATTTCTTAAACTTCTTAAATTCTACTTTTATAG---TTAGTCTTTTTTT
 90 NR0009/1-331     ----------------------------AGATATTAACTACCCATTTGGATTTAATATCCGATCCAAGACTCTATACCTGATGGTTGGAAGTTCTAAAACAATTTATTTTGTTGGGG-AAGACAAAGAGTCGAATAAAACGCTAAT-TAATTTTATTGTACATCTTGGCTTCAGTGCCATTATT--AATTTTCAAGCTTTAAAGTTTTAGTTCAAGTCTGGAATCGTAGCA
 91 NR0004/1-331     -----------------------------------------TTATACATTTAAATGCTAGAAAATTTA--AGTAAAACATTTATAAATAAAAGTAAAATAGTTTTAGGAATATGAGTAAATAGTTTTTTTTAT-GTAAAAACATT-TTATCA--ATTTCATTTATTCATTTTAGTTAAATTTTTCATTCACAAAAAACTTTTTTTTGGTAAAATAAAGACTTTATAAAGA
 92 NR0018/1-331     -------------------------------------------------------GCTTTAAAGCAGCATTCAGAAATAATTTTCGATCGTTAAAAATACAG-ATACTTTCTTAAATAACATAATATTAATATAATAAATAATATT-TTCTTA--AATTATTAATGTGTTTGCTTGTTATTAATGTGTTTTCATGCATATTTTAAAATAATTTGTTCTGAAATCAATCGAT
 93 NR0021/1-331     -------------------------------CAGAGTCCCTTGCCCCGCACGACACTTTCACTGCTTTAAAGGATGTTCGCACAAAGAACAAATAAAAATAT-ATATTTTTGTATTAACCTCTTCTTTAATTTAATAAATGTTGAG-CTGACA--ATTAAAATAAGATTATTTTTATTCCTTCTCTGGCAGAATTAGTATTTTTACATATGAT-TTACCGTTTATATGTAC
 94 NP0024/1-331     -----------------------------------------------------------AAAGTGTTCACTGGGGAACTGCACAATATGACTGCTTTTAACCGTAGTGATTTCAAATATTGAGCCATGCTGTTGCAGTCTTAAAAACTGGAGACCTAAGGGCAGCTTTCTTCTAGTCACCCAATCCAGCACTTTTTTAAAAAATCAGTAAAACTCTTCGACCACCAAGGAA
 95 NM0025/1-331     ----------------------------------------------GGTGTGTGTGA-GTGTTTATGAGTATGTGTACACATG-TTCGTATAC-ATGTGTGT-ATACAAGCAT-GTA--TGCAT-ATGTGCGCG-TGTATGCGTGT-TCATGAGTATGTGT-GTACATGTTCGTATACA-TGCGTGTATGCAT-GCATGC---------------------------
 96 NG0039/1-331     ----------------------------------------------TTTTGAAAAGCAAGCATAAAAGATCTAAACATAAAATCTGTAAAATAACAAGATGTAAAGATAATGCTAAATCATTTGGC--TTTTTGATTGATTGTACA-GGAAAATATACATCGCAGGGGGTTGACTTTTACCATTTCACCGCAA---TGGAATCAAACTTGTTGAAGAGAATGTTCACAG--
 97 NP0009/1-331     -------------------------------------ACAAAGAAATATATATTAAATTAGCACG----TTTTCGCATAGA----ACGCAACTGCACAATGCCAAAAAAGGTAAAAGTGATTAAAAGAGTTAATTGAATAGGCAATCTCTAAACGAATCGATACAACCTTGGCACTCACACGTGGGACT-AGCACAGACTAAATTTATGATTCTGGTCCCTGTTTTCGAAG
 98 NM0076/1-331     ----------------------------------------------------TAGCTTGTGTCACCTGGCATG-GCATGTGTCACTTGGGCATAGCATGTGTCACCTGGGCATAGCA--TGTGTCACCTGGGCATAGCACGTAT-C-ATCTGGGCATAGCTTGTGT-CACCTGGGCATAGCACGTGTCATCTG-GCA---------------------------------
 99 NR0005/1-331     ---------------------------------ATTTTAAATCCTAAAAATAAATAATTCATATATAAGAGCGCATTTAGAAGTATAAATTTTGCTGTCGGCCATACTAAGGTGAAAA-CACCGGATCCCATTCGAACTCCGAAGT-TAAGCGCCTTAAGGCTGGGTTAGT-ACTAAGG-TGGGGGACCGCTTGGGAAGTCCCAGTGTCGACAGCCTTTTTATTTTTTTTG
100 NP0001/1-331     ---------------------------------------------AAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCAT-ATACATATCCATATAC-ATATCCATATCTAA-TCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAAACCTTCTCTTTGGAACTTTCAGTA-ATACC
101 NM0001/1-331     ------------------------------------------------CAAGACCGAGTTACTAAACAGGACTATTACTGCCACGCCAATT--GTAGCGCGCAGCACGTCTCTGCTCACCACTATCCTCTTGTTGACGCTATTGCT-ACTATCGCATCCCGCTTAGCTATACCTACTGA-TGCTCAATTACCCGCC-----------------------------------
102 NP0021/1-331     ------------------------------------------------TCTTTGTATGATAATGTCCATACAATATATTAATATTGAAAACAGTCATGCCATAGAACAA-TTTTGTTCAGTAAATG-TACATCACATAAGTAACTT-CAAATTTTAGAAGGGATAATTTTAAAGTCAGGCAGGCAGCATTAAAGAAGAATGAGGAAATCAATTTAAAATTCTTTTTTAAAG
103 NR0003/1-331     ------------------------------------------------TTACATATTTTAGCTATTTGACTACTTTAATGCTAGTAAATTAAAATGAATTTAATTCATTTTCACTTTAAAACACTTATTTTAAT-A-AAA-TATATG-ATTTTA--AAATGATAAAATATTTTTTAAGAGGTAAATTTAAGAAATTAGTTAAATTTTAAAGAAAAAGCATCTAAAAATGGAC
104 NG0042/1-331     --------------------------TTTTCGAGTTAGCGTGTTTGAATACTGCAAGATACA-AGATAAATAGAGTAGTTGAAACTA-GATATCAATTGCACACAAGATCGGCGCTAAGC-ATGCCACAAT-TTGGTATATTATGT-AAAAC---ACCACCTAAGGTGCTTGTTCGTCAGTTTGTGGAAAGGTTTGAAAG-ACCTTCAGGTGAGAAAATAGCATTATGTGC
105 NR0002/1-331     --------------------------------AGATAAGAATTAGATGTTTATATTCTGCTAATTTCACTGGTGAAAATGTAGCAAATAGAAATTATTTTAATCTAATAA-ACTAGCAAATAGTATTTAAAAC-A-AAAATATTTG-TTTTTT--ATGTTGTAAAATGTTTTAAATTAGATAAAATTTACAAATTTACAAATTTTCAAGCAAAATAGGTTCTAAAAAATGA
```

```
106 NP0014/1-331    -----------------------------------------------------GGGGCGCGGCGATTTTTAAAATGTATGCATGTAAATAAAACCTGTGCTGGTTAGGCATCATCATTTTTGGTCTCATCGTTCTATGATAAATATA-AATA---TAAATATGTACACGAGTACATAGCATTTAGTTTTCATAGCTAAGCAAGTCTAATTTTAATGAATCTAGCCAAATTTT
107 NM0052/1-331    ----------------------------------------------------------GTGTGTGCACGTGCGCGTGTGCATGTG-TGTGTGCGT-GTGCGTGT-GTGTGTGCAT-ATG--TGTGT-GTGCGTGTG-TGTGCGCGCGT-GTGTGTGTGTGTGT-GTGTGTGTGTGTGTGTGTGTTGAGACAAGGTCT-------------------------------------
108 NR0010/1-331    ---------------------------------------TTGATAAGAACTTCAATCTTTGACTAGCTAG---CTTAGTCATTTTTGAGATTTAATTAAT--ATTTTATGTTTATTCATATATAAACTATTCA---AAATATTATAGA-ATTTAA--ACATTTTAACATCTTAATCATTCATAAATAACTAAAAATCAAAGTATTACATCAATAAATAACTTTTACTCAATGT
109 NR0001/1-331    --------------------------------GGCCGGGCTTGTTTTCCTGCCTGGGGGAAAAGACCC--TGGCATGGGGAGGAGCTGGGCCC-CCCCCAGAAGGCAGCACAAGGGGAGGAAAAGTCAGCCTTGTGCTCGCCTACGGCCATACCACCCTGAAAGTG--CCCGATATCGTCTGATCTCGGAAGCCAAGCAGGGTCGGGCCTGGTTAGTACTTGGATGGGAGACC
110 NG0031/1-331    --------------------------------------------------CGGCCCCAGCGAGGAAGAAATGATTATGCAACAGATGATGATCAAGCTCAGCATGGGCATCAGTGGACAGTGCTTCAAGGAGTGTGTGACTAG--CTTCAGCTCTGGCCAGATGGTTCCCCAAGAGGCAACTTGCATTCAGTCATGCGCCAAGCGCCAACAGTCTGCATTCATGGCCATGAAT
111 NP0004/1-331    ------------------------------------------------CAGTTCATAGGTCCATTCTCTTAGCGCAACTACAGAG--AACAGGGCACAAACAGGCAAAAAACGGGCACAACCTCAATGGAGTGATGCAACCTGCCTGG-AGTAAATGATGACACAAGGCAATT-GACCCACGCATGTATCTATCTCATTTTCTTACACCTTCTATTACCTTCTGCTCTCTCTC
112 NP0006/1-331    ----------------------------------------------------CTTTGTTTGACAAAAATGTTGCCTAAGGGCTC-TATAGTAAACCATTTGGAAGAAAGATTTGACGACTTTTTTTTTTTGGATTTCGATCCTATAAT-CCTTCC--TCCTGAAAAGAAACATATAAATAGATATGTATTATTCTTCAAAACATTCTCTTGTTCTTGTGCTTTTTTTTTACCA
113 NP0018/1-331    -----------------------------------------GCTAGCCAGGGCAGCATAGTGAAGCT--CTGCTTACAAAAGCAAAACAAAAAACCGGGAAAACCTTCAGTTTTTGAATAAGACACATATGTGAGTTCCTAAGATGCA-TGAA-CGCCACAGAGAC---CACTCAAGAGAGCACCAAGGCTCCTACGCAAAGGCAAGT-TAGCTCCAGAGCCCTAGGAGAGCT
114 NR0015/1-331    -------------------------------CTTTTCTTACCTTTTACATTTCAGCAATATATATATATATTTCAAGGATATACCATTCTAATGTCTGCCCCTAAGAAGATCGTCGTTTTGCCAGGTGACCACGTTGGTCAAGAAATC-ACAGCCGAAGCCATTAAGGTTCTTAAAGCTATTTCTGATGTTCGTTCCAATGTCAAGTTCGATTTCGAAAATCATTTAATTGG-
115 NM0057/1-331    ------------------------------------------------TGTGTATGTGGTGTGTATGTGTGTGTATGTG-GTGTG-TGTGTGTAT-GTGTATGT-GTGTGTGCGT-GAA--TGTGT-GTGTGTGTA-TGTA-GTGTGT-GTGTGTGTATGTG--GTGTGTGTGTGTATGTGGTGTGTGTGTGTG--------------------------------------
116 NG0010/1-331    ----------------------------TATCACATAATGAATTATACATTATATAAAGTAATGTGATTTCTTCGAAGAAT--ATACTAAAAAATGAGCAGGCAAGATAAACGAAGGCAA---AGATGACAGAGCAGAAAGCCCTAGT-AAAGC--GTATTACAAATGAAACCAAGATTCAGATTGCGATCTCTTTAAAGGGTGGTCCCCTAGCGATAGAGCACTCGATCTTC
117 NG0015/1-331    -----------------------------------------AACGGGAAAAAGTTAGTTGTGGTGATAGGTGGCAAGTGGTATTCCGTAAGAACAACAAGAAAAGCATTTCATATTATGGCTGAACTGAGCGAACAAGTGCAAAATTT-AAGCATCAACGACA-ACAACGAGAATGGTTATGTTCCTCCTCACTTAAGAGGAAAACCAAGAAGTGCCAGAAATAACATGAGCA
118 NG0011/1-331    -----------------------------------------------------------CCCCTAGCGATAGAGCACTCGATCTTCCCAGAAAAAGAGGCAGAAGCA-GTAGCAGAACAGG-CCACACAATCGCAAGTGATTA-ACGTCCACACAGGTATAGGGTTTCTGGACCATATGATACATGCTCTGGCCAAGCATTCCGGCTGGTCGCTAATCGTTGAGTGCATTGGT
119 NP0003/1-331    ----------------------------------------------------TATATAAATGCAAAAACTGCATAACCACTTTAACTAATACTTTCAACATTTTCGGTTTGTATTACTTCTTATTCAAATGTAATAAAA-GTATCAAC-AAAAAATTGTTAATATACCT------CTATACTTTAACGTCAAGG-AGAAAAAACTATAATGACTAAATCTCATTCAGAAGAAC
120 NR0022/1-331    ----------------------AAAAATTTATAAATAATTTTAAAACAATAAATAGAAAAACAAATAAGATTATAAAAACTTACAAAAATGGCCGGTGGTAAAGGTGGTAAAGGTATGGGTAAAGTCGGAGCCAAGAGACACTCCAGA--AAGTCTAACAAGGCTTCCA---TTGAAGGTATTACT--AAGC--CCGCTATCAGAAGATTAGCTAGAAGAGGTGGTGTTAA--
121 NG0001/1-331    -----------------------------------------------ATATCTTCCCATTTTTGGGTGGTGCCGGACCATACTACTCTTTCCCT--GGCGACTATGGTATTTCTCGTGATTTGCCTGAAGGTTGTGAAATGAAGCAAC-TGCAAATGGTTGGTAGACATGGTGAAAGATACCCTACTGTCAGTCTGGCTAAGACTATCAAGAGTACATGGTAT--AAGTTGAC
122 NG0030/1-331    ------------------------------------------GCTGAGCAAACAAACAAGCGCAGCGAACAAGCTAAACAATCTGCAATAAAGTGCAAGT-TAAAGTGAATCAATTAAAAGTAACC---AACAACCAAGTAATTAAAC-TAAAAACTGCAACTACTGAAATCAACCAAGAAGTCATTATTGAAG--ACAAGAAGAGAACTCTGA--ATACTTTCAACAAGTCC
123 NG0016/1-331    ------------------------------------------------------AGAGGAAAACCAAGAAGTGCCAGAAATAACATGAGCAACTACAATAACAACAACGGCGGCTACAACGGTGGCCGTGGCGGTGGCAGCTTCTTTAGC-AACAACCGTCGTGGTGGTTACGGCAACGGTGGTTTCTTCGGTGGAAACAAC-GGTGGCAGCAGATCTAACGGCCGTTCTGGT
124 NS0011/1-331    ---------------------------------------------------------AAACTGCTTCCTCTGTATAAATCAAAGCAAAATGTAAATAGCGTTGACAAGTGATTACAGAAGTTAGGTGAGGTTAATTACCAATTTTTTT-TTTTAA--AATTGGTGAAATAAGATTACGTTTAAAGGAGCATTAACAGGTTTACTCATAACAATCATTTTCAAATTTCCCTATC
125 NR0011/1-331    ---------------------------------------------CTACGCTTAGATTTTAACTTTATCCC-ACTTTAATTTCAA-GCGTAAAAATAA---AAATCCCACACAAAAATTAAGTGGAAATTGATGCAAAAATTTCACTA-AAATTT--AATTCAATAAATATGTAAAAATGGTTGATCTCTATAATTTATGAGATTTGCATTATTTAAGGCTTATAAGAAATTT
126 NM0055/1-331    --------------------------------------------------TGCGTGTGGTGCGGATGTGTGCATGTGTGTGTGTGGTACAGGTGT-GAGTATGT-GTGTGTACAC-GCA--TGTGT-TTGTGTGTGGTGCAGATATGT-GCATGTGTGTGTGTGGTACAGGTGTGTGAGTA-TGTGTGTGTG----------------------------------------
127 NR0014/1-331    -----------------------------------------------------AGAAAAAGGAAAGGTGAGAGGCCG-GAACCGGCTTTTCATATAGAATAGAGAA-GCGTTCATGACTAAATGCTTGCATCACAATA-CTTGAAGTT-GACAATATTATTTA-AGGACCTATTGTTTTTTCCAATA-GGTGGTTAGCAATCGTCTTACTTTCTAACTTTTCT-TACCTTTTA
128 NG0036/1-331    ---------------------------------------------ATTCCGATGCTGACTTGCTGGGTATTATATGTGTGCCCAATAGAAAGAGAACAATTGACCCGGTTATTGCAAGGAAAATTTCAAGTCTTGTAAAAGCATATAA-AAATAGTTCAGGCA-CTCCGAAATACTTGGTTGGCGTG-TTTCGTAATCAACCTAAGGAGGATGTTTTGGCTCTGGTCAATGAT
129 NS0001/1-331    -----------------------------------------------GGAAGCCCATAGAGGGCTATGGTGAACAACGAAATATC-TTCCGTTCAAAACTGGAAAGAAGCTTTCTGAGAAACTGCTCTGTGTTCTGTTAATTCATCTC-A---CAGAGTTACATCTT-----TACCTTCAAGAAGCCTTTCGCTAAGGCTGTTCTTGTGGAATTGGCAAAGTTATATTTGGAA
130 NR0016/1-331    -------------------------------------TCAAGTTCGATTTCGAAAATCATTTAATTGGTGGTGCTGCTATCGATGCTACAGGTGTCCCACTTCCAGATGAGGCGCTGGAAGCCTCCAAGAAGGTTGATGCCGTTTTGT-TAGGTGCTGTGGGTGGTCCTAAATGGGGTACCGGTAGTGTTAGACC-TGAACAAGGTTTACTAAAAATCCGTAAAGAACTTCAA
131 NS0007/1-331    ----------------------------GGGCTTCTCATGGTGGCTCTGAGAAGCCAGGGAA---ACTGGAGGTGGGAGGGGCCTC-TCGGGACTCCACTGGGCTTGGTGCA-TTGGAAGAGGGCCTCATCTCCAG--TTGAGGCAGGAACCGCAGG-GTACCTCT--GATTTCAGACTCCGATCGCAGGGTCCCTGCAGACTGGGGACAGG--AGAGTCAGGC-CTCGTCTT
132 NP0007/1-331    ----------------------------------------------------AAGCTTTCCTTTTCCTTTTGGCTGGTTTTGCA-GCCAAAATATCTGCATCAATGACAAACGAAACTAGCGATAGACC---TTTGGTCCAC---TTC-ACACCCAACAAGGGCTGGA-TGAATGACCCAAATGGGTTGTGGTACGATGAAAAAGATGCCAAATGGCATCTGTACTTTCAATA
133 NR0012/1-331    -------------------------------------------TAAGAAATTTTAAATTTAACGCGGAAG---CTTCATTTTTAGATAAAATTTATTAATC-ATCATTAATTTCTTGAAAAACATTTTATTTATTGATCTTTTATAAC-AAAAAA--CCCTTCTAAAAGTTTATTTTTGAATGAAAAACTTATAAA-AATTTATGAAAACTACAAA-AAATAAAATTTTTAAT
134 NG0013/1-331    -----------------------------------------------CCAGAGCGGTGGTAGATCTTTCGAACAGGCCGTACGCAGTTGTCGAACTTGGTTTGCAAAGGGAGAAAGTAGGAGATCTCTCTTGCGAG--AT-GATCCCGCATTTTCTTGAAAGCTTTGCAGAGGCTAGCAGAATTACCCTCCACGTTGATTGTCTGCGAGGCAAGAATGATCATCACCGTAGTG
135 NP0002/1-331    -------------------------------------------------------------TACTAGCTTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATGAACGAATCAAATTAACAACCAT-AGGATGATAATGCGATTAGTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATC-TATTAACAGAT
136 NG0037/1-331    --------------------------------------ACCTAAGGAGGATGTTTTGGCTCTGGTCAATGATTACGGCATTGATATCGTCCAACTGCATGGAG--ATGAGTCGTGGCAAGAATACCAAGAGTTCCTCGGTTTGCCAGTTATTAAAAGACTCG-TATTTCCAAAAGACTGCAACATACTACTCAGTGCAGCTTCA---CAGAAACCTCATTCGTTTATTCCCTT
137 NP0025/1-331    -------------------------------------------CTCTGCGCACGGGCTTTTTCTGAGAGA---CCCATGTTTCCTTTTTACTTTTATAAACAGT--TTACATGCT--ATGTTTCTAGAAGGAGGGGAAACCTA----A-TCCCCCTAATCCAATGGCGGGGAGGAAATAGGGTGG--GGTGGGGTGGGGTGGGGTGGGGGA--GGGAAATATCTCG-CTACT
138 NM0007/1-331    -----------------------------------------TATCGTGTCCGGTG-CTGGCTCTCAACGCGTGGCAGTAAGCTCGCCGTTAGTC-ACTCTATCTGCATTTTTC-TGTATAACGGTTCACGTTTATC---CCTGCATCA-CGCATTCTT-------TAGCTCGTACCGGCCTCGCT---TGCATTCTACTCCCG-----------------------------
139 NG0017/1-331    -------------------------------------CTTGACTAAGTTATTATATGAAACTGCTTTGTTGACTTCCGGCTTCAGTTTGGACGAACCAACTTCCTTTGCATCAAGAATTAACAGATTGATCTCTTTGGGTTTGA-ACA-TTGATGAGGATGAAGAAACAGAG-ACTGCTCCAGAAGCATCCACCGCAGCTCCGGTTGAAGAGGTTCCAGCT--GACACCGAAA
140 NG0035/1-331    ------------------------------------------GCATTGGTGACTATTGAGCACGTGAGTATACGTGATTAAGCACACAAAGGCAGCTTGGAGTATGTCTGTTATTAATTTCACAGGTAGTTCTGGTCCATTGGTGAAAGTTTGCGGCTTGCAGAGCACAGAGGCCGCAGAATGTGCTCTAGATTCCGATGCTGACTTGCTGGGTATTATATGTGTGCCCAATA
141 NP0016/1-331    --------------------------------------------------------AGTTATCATGACCTCTTAGTTGGCTATGGTCTATGGGTAAT--TATTA-ACCATTATTTAATTCAAAGAAAAATA-TATAATATTAAGCATT-AGTTTA--AAATAA-AATGT-TATATTATTTGATAAAGTGACCAAGGGTATATGTTTTGTTGTTTAATTTTGATTTTTTT---T
142 NG0021/1-331    ------------------------------------GAGCGTGAGGTGACCATCGAGCAGACTGGCGAGCCGGCAAAGAAGTCCGCCGAGGAGCCAAAAGACAAA-ACCGCCAGTCAGTAGAAATAAGTTGAGATTATACTAAAACCG-ATAAA-ATGCTAGT--GAACTCCTATGTTTAGATATTCCAAAACCTAT---CAAATTTAAGTTCTTGTTAAATTAACAAGTTAA
143 NS0010/1-331    --------------------------------------------GTATCTTTTGTCCAATATGAAGAAGGTCAACATGAGGATGG--GAATGATAATT-GATAG-CATATAATATTATTCTTTTGTCAATACTAGTGTTTAGGATATT-TCTACT--AATACCTAATACCTCAATGGTCCAATACTAAATAAGGTACTATTCATTGTATTGATTGATTCTGCATTTATCT--T
144 NG0007/1-331    --------------------------------------------AAGTTTGTCAAGGAGAACCCATATTTCACAAACAGAATGT-CTGTCAACAACCAACCATTCAAGTCTGAGAACGACCTAAGTTACTA-TAAAGTGTACGGCCTA-CTGGATCAAAACATCTTGGAATTGTACTTCAACGATGGAGATGTGGTTTCTACAAA----TACCTACTTCATGACCACCGGTAA
145 NG0012/1-331    ---------------------------------------------------TTGAGTGCATTGGTGACTTACACATAGACGACCATCACACCACTGAAGACTGCGGGATTGCTCTCGGT--CAAGCTTTTAAAGAGGCCCTACTGGCG-CGTGGAGTAAAAAG-GTTTGGATCAGGATTTG-CGCCTTTGGATGA-GGCACTTTCCAGAGCGGTGGTAGATCTTTCGAACAGC
146 NR0008/1-331    ----------------------------------------GTTTTTGTTTCGTCGTGCTTGGCATAATGGAAAATC----ATCCTTCTTTGTGATAAAATAGAAAACAAAATGGCTAGTACAATCTTGAGACACCGA-TGCTGGGTT-TTGGGTACAAA--G-ACTGCCAATACGTCCATCCCCTA-TCGAATTACCAA--ATATTGTTTTCTAATTTGGGTATTAAGTTCC
147 NS0013/1-331    -------------------------------------------------TATATTATTATTCAATAGAAGTAATAAAGAAAAAGTTGGTAAAGCAACTTAACAGTAAAAAGGTAATGATTGAAAAAG---TTTTTGAACATCT--------AAGCTATATGTTGATGGGTTTACAATTTTACCATTAGTACTCATGCCTAGTACTTTTCTGTTCGTCCTTAATGTCCGCGATT
148 NR0006/1-331    --------------------------------------------------TGAGAACTAAACACTTGAAT--TCAGATATAAGTATTATAATTCAGGAATCTGAGATTTACGGAATTATAAATCCAGATGGAC-TAGAGCATGCCCAA-GCGTAAGTAAACGGTTTTTCAAAGACTGGAGA---GATGTTTTGGGTGAGTTCGATTTTAGGTGTTGAGTATATAAAG-AGTGC
149 NS0012/1-331    -----------------------------------------TTTAGAGCAAGCGCCTTTGTGAGCCCTCCCGGTTACGACGCCTTGGCAATGTAGCAGATAACTCTGCACTTCTAGAATCATTCCACTACGACATTTGGCTCATCACC-AGCT-CGCGAGAAATGT---AAATAAGCCAACAACCAAGAATGCGTAACATTAAAGAATACAGTTGCTTTCATTTCGGCGTGAT
150 NG0025/1-331    -------------------------------------------AAAGGATTTATTAAATTAATATGAGTATTCTAGTCTTTAGCTCATTACAAATACAATGCAATAAATTAGGAACAATTAAGTAGG-------GGAATGATGATGGA-TGATGTTGGATGATGATGGCTCCTTTACTTGTCCTTGCCGTTGGGTG--CTCCGTTCTCCTTGCCCTTCACCTCGCTTTCATTT
151 NP0012/1-331    ----------------------------------------------------AGTGCAGTTGTCAGTTGCAGTTCAGCAGACGGGCTAACGAGTACTTGCATCTCTTCAAATTTACTTAATTGATC---AAGT---AAGTAGCAAAAG---GGCACCCAATTAAAGGAAATTCTTGTTTAATTGAATTTATTAT--GCAAGTGCGGAAATAAAATGACAGTATTAATTAGTAA
152 NG0033/1-331    -----------------------------------------------GAGAGCAACGATTCACTTTTATTATAACAAAG--ATTAAAAACATACTCTTTTCTTA-A-ATACTCCAGCTAGTCACACCATTTACTTCATATCAGCCCTT-TCTTCA--GTGCATTTTATTTT--TTCTATGAAATCTTGTTTACTCTTCTTCATACAAAATATTTTATCTCACTTTTAATGCTT
153 NG0003/1-331    ----------------------------------------------GCTCCATGGTGGTTGATTACAACAACACGAGTGGGTTTTTCAATGATACTATTGATCCAAGACAAAGATGCGTTGCGATTTGGAC-TTATA-ACACTCCTGA-AAGTGAAGAGCAATACATTAGC--TATTCTCTTGATGGTGGTTACACTTTTACTGAATACCAAAAGAACCCTGTTTTAGCTGCC
154 NR0019/1-331    ----------------------------------------------------------GTCTATGTAGTCAAATAA-ATAAATCAAATAAAAAATAAAAAGAAA-CAATTTTTGTATTA---TCCTTTATTATGTAATATATATTACA-TTCCGC--AACAAAATTAGCCAATTTCCGTGCTGAATTTTACATAGGTTTTATTTTTTACAGACGT-TTGAAGGACAGTGTCA
155 NR0024/1-331    ------------------------------------------------ATCTGAGAGAGTATGCTTTTTTC-TATCGAGTGTTA--GTGTAGCAATTTTCTAAAGTGCATTGAGAGATTGAGCAGAAATGTTTAGAACTTATTCACAT-CAAATT--AACTTAAGAAAAATAATAACTTACTTAATCAATTCACAATTAGCCATTATGAATAACTAAA--CTAAAACATAAAG
156 NS0008/1-331    ------------------------CCGCTTGCCTCTCGAGATGTCCCCGGGGAGAGAGGCCGC-TTGTCGAGCTGTATTTGGAA---CCTGGGGTTTTTCCGAACGATGCACGGAAAACTGCCCCTTCGTGTTG--ACTTCATTCACAGGCTGGAGTTCGGAGA--GG---TGTCCGGGCATC---GGGTTCTTATCAAGAGGGGACCGG--GAAATCGGGGTCCTACGGA
157 NG0005/1-331    ---------------------------------------TGCCAACGAAGGTTTCTTAGGCTACCAATACGAATGTCCAGGTTTGATTGAAGTCCCAACTGAGCAAGATCCTTCC---AAATCTTATTGGGTCATGTTTATTTCTATC-AACCCAGGTGCACCTGCTGGCGGTTCCTTCAACCAATATTTTGTT--GGATCCTTCAATGGTACTCATTTTGAAGCGTTTGACA
158 NG0020/1-331    --------------------------------------------TGGACGAGAGCGTGGTCCTGGTGGAGGCAAAATCGGAGCAGCAGGA-GGCCGAACAAGGTGGCTATAGTTCCAGG--CACTT-CCTCGGCCGATACGTTCTGCC-GGATGGATACGAGG-CGGACAAGGTGTCCTCGTCGCTGAGCGACGA-CGGCGTTCTGACCATCAGTGTGCCCAATCCTCCAGGC
159 NG0027/1-331    ----------------------------------------CACAAAGTGGCGCATGATGTGACCATGGTCGTCCTGGCGTTCCTCATGCTTGCCCTCG-ACCAAGATGGAGTCGTCCACCACCTTCACGT-TGAGCTCACTGGGCTTG-AACTGGGCGACGTCCATGCACACCTGGAATCCATCCTTGCCCACATGGGCGGTTGCCGGCCAGTGAATGTCGTTGCGGTTGGCC
160 NG0028/1-331    --------------------------------------GAATGTCGTTGCGGTTGGCCATCTCGCGCCGCAAAGCCAAAACCTGGCCGGCGGGCGACGATGGGCATATCGGCGATGCGCAAGGGCATCCGTTGATGGA-ACGGCGCTG-CTGAGTGCCAAGGG-GCAGCACGTAGCGGGAATGCGGA-TGCAATCCCAGTCCAAGCTCGTAGATGGGGCTGCGGGGCTCCTGC
161 NS0014/1-331    -------------------------------------------------TTAGAGCAATCATTGAAAGTACTAGATACATTTTAGCCAGAGAGGACTCGT-TGACGTAGAATTAAAATTC----AAATGAA-TTTCCGCCCCATTCAT-ATACCCCAAATAACAAACATATTAAAACTTCATAATTATTCAAAATGTGGAGTAGTAATAGAAGAGCAGTACCT-TCAAAATTC
162 NG0009/1-331    ------------------------------------TTCATTCAACGTTTCCCATTGTTTTTTTCTACTATTGCTTTGCTGTGGGAAAAACTTATCGAAAGATGACGACTTTTTCTTA-ATTCTCGTTTTAAGAGCTTGGTGAGCGCT-AGG--AGTCACTGCCAGGTATCGTTTGAACACGGCATTAGTCAGGGAAGT---CATAACACAGTCCTTTCCCGCAATTTTCTTT
163 NP0015/1-331    ---------------------------------------------------GGTTGAGATTAACGCCTATGACCAGGGCAACCGCACCACGCCGTCCTACGTGGCTTTCACAGACTCGGAACGCCTCAATGGTGAACCGGCCAAGAACCAGGTGGCCATGAACCCAGAAACACAGTGTTTGACGCCAAGCGACTCATCGGCCGAAAATACGACGAT-CCCAAAATCGCAGAG
164 NP0008/1-331    ----------------------------------------------AAAACTTTATGATTTCAAAGAATAACCTCCAAACCATTGAAAATGTATTTTTATTTT-----TATTTT--CTCCCGACCCCAGTTACCTGGAATTTGTTCTT-TATGTA--CTTTATATAAGTATAATTCTCTTAAAAATTTTTACTACTTTGCAATAGACATCATTTTTTCACGTAATAAACCCAC
165 NS0003/1-331    -------------------------------------------------ATTACATGCGAATCCTATTGGGAACCTACTGAATTCACCATGATACTTAGATTCCGTTCCTCAAA-ATGTTGCTCCATATTG-AAAAG-CAAACTCATA-CAAGCATGTCCCAT-TGGGAAC--TCACTGAATTCGCCTAGAAATTTTGATTCCATTCG-TGAAAATTTTCTATATC-CCGAA
```

```
166 NS0005/1-331       --------------------------------------------AAAGAACGTCCGCTCTG-CTCTCGAATCGCGACGGGTATC-TCTTGGAGCTCACTGGGTGGACTCA-AGGGAGTCAAGCCTCCTGAGGCG--TTTGGAGAGAGGTCGCGAGATTGGTCTC--TAGGCCACGCAGGAGACGAAGGCCCTCATCTCTCGATGACGGGG--GAATCTCGGGGTTGTTCTC
167 NS0004/1-331       ---------------------------------------TGGCCCAGGCAAGTCCAATCTTCCATTCGAGTTGCGAAGGAAAGC-TGGGGATTGCTCTCGAGTGACTGCA-GGGCCAATAGACCTCATCTAGGC--TTGTGTCCAGAAGCCAGTG-TTCCTCTC--CAGGGGCGACAGGGATCTCGGGGTTGCATTCCAGACGCACCCGG--GGAGACAGGCATTCATCTC
168 NG0004/1-331       ------------------------------------------------AAAGAACCCTGTTTTAGCTGCCAACTCCACTCAATTCAGAGATCCAAAGG-TGTTCTGGTATGAACCTTCTCAAAAATGGA-TTATGACGGCTGCCAA-AT-CACAAGACTACAA---ATTGAAATTTACTCCTCTGATGACTTGAAGTCCTGGAAGCTAGAATCTGCATTTGCCAACGAAC
169 NR0007/1-331       ----------------------------------------CCGAACTCCGAAGTTAAGCGGTTTAAGGCCTGTTAAGTACTGA--GGTGGGGGACCACTCGGGAACTTCAGGTGCTGATAGCTTTTTGCTCCTGAA-GCTATCTTT-TTGCACTCTTTCTT-TTTATCTATCCACCCTTCAGTAC-TTCTCACACTATCCAGGTTGGCGAGTTTTTGTTTCGTCGTGCTTC
170 NG0002/1-331       --------------------------------------------------ATGAAAAAGATGCCAAATGGCATCTGTACTTTCAATACAACCCAAATGACACCGTATGGGGTACGCCATTGTTTTGGGGCCATGCTACTTCCGATG-ATTT-GACTAATTGGGA---AGATCAACCCATTGCTATCGCTCCCAAGCGTAACGATTC----AGGTGCTTTCTCTGGCTCCAT
171 NG0032/1-331       ------------------------------------AGGCCAAGCAAGGAGCTGGCATGTTCTGAGCAACTTCTTCAAGTCCTCCTTCTCATCACTGAGCGAGTAAGGATTTTAAGTGTTGTTAAGTTGACTTTGCAGCACTCTTG-ATAAATGTACTTGACGGAATCACTCTCTTAAGAAGATATAAAACTT-----TGAACTAGAGCATCAGTGTAACTAA-ATGTTAA
172 NR0020/1-331       ---------------------------------------------CTCTTCAATAATAACACATTCTTCAGTTA-ACACATGGAAAAAATATAAAATAGCTA-GTTTTATTTTATTA---TTTTCTGTTATTTTATAAATATTGCC-GTCATC--AATGTAATTTCATTAATTTCAAACTGTCTTACACAGAATTCAGATTTTTTTTTTAGATGTTTAAGGTTCAGAGTCC
173 NS0009/1-331       ----------------------GAGGCCACGTCTGGA-ATGTCTTCGTG-AGACCGGCCTC-ATCCTGAGGTGCGACCGGAAGGATCGGGAACCCCTTCCAGACAAAGCA-GGGGAGTCGACCCTCCTGTCCAG--ATCAGGAGGGGAGAAAGGGCTCAGAGGA--GGGGGTGCCGGAAAACCTCAGTGTTCCTCTCGAG-GGAGACCGG--GATTTCGGGGAACTTTGTC
174 NG0014/1-331       --------------------------------------AGTGCGTTCAAGGCTCTTGCGGTTGCCATAAGAGAAGCCACCTCGCCCAATGGTACCAACGATGTTCCCTCCACCAAAGGTGTTCTTATGTAGTGACACCGATTATTT-AAAGCTGCAGCATACGATATATATACATGTGTATATATGTATACCTATGAATGTCAGTAAGTATGTATACGAACAGTATGATAC
175 NS0006/1-331       -------------------------------------------------GTGTGCGGTTT--CTCACGAGGTACGACGGCGAGG-TCAGTGAGCCTCTCGTGGGGCGCCA-GGGAAGTCGGGTCTCCATGCGAG--TGGCGAGGGGGAGCGCGTCATTGCTCCC--GAGCCATGGTAGGGGAATGTGGCC-TCGAGACGTGTTGAAGAAG--GTCTCTCGAGGTCTTTCTC
176 NG0040/1-331       ---------------------------------------------ATGTGGATTGCGCATACT-TTGTGAACAGAAAGTGATAGCGTTGATGATT---CTTCATTGGTCAGA-AAATTATGAACGGTTT-CTT---CTATTTTGTC-TCTATATACTACGTATAGGAAA--TGTTTACATTTTCGTATTGTTTTCGATTCACTCTA-TGAATAGTTCTTACTACAATTTTT
177 NG0029/1-331       ---------------------------------------------GTACGTAATTTTAAATTAAAAACACTAACAATCATCTGCATGCAATTGTCTGTATTAATCTAATAAATAAATAGCTTTTTTAAGTTAGTATGTAAATACAT-TTTGAAGAATATCTTGTCAAAGT--TCCATAGGCCTTTCTGGCGG--ACAACATCCG--CTAACA--AACCCTTCGATTATCTC
178 NP0023/1-331       --------------------------------------CTCAGACCCTGAGGCGCCGGCCATGGCCCCACTGAGACACAGGAAGGGCCGCGCCAGAGCACTGAAGACGCTTGGGGAAGGGAACCCACCTG--GGACCCAGCCCCTGGTGGCTGCGGCTGCATCCCAGG-TGGGCCCCCTCCCCGAGGCTCTTCAAGGCTCAAAGAGAAGC---CAGTGTAGAAAAGCAAAC
179 NG0022/1-331       ---------------------------------------------TCGATGGTGCCCTTCTATGAGCCCTACTACTGCCAGCGCCAGA-GGAATCCCTACTTGGCCCTGGTT---GG--ACCGATGGAGCAGCAGCTGCGCCAGCT-GGAGAAACAGGTGG-GCGCCTCGTCGGGATCGTCGG-GAGCCGTGT-CGAAAATCGGAA-AGGATGGCTTCCAGGTCTGCATGC
180 NG0006/1-331       -----------------------------------------------ACTCGACTGGTACCCTAGAGTTTGAGTTGGTTTACGCTGTTAACACCACACAAACCATATCCAAATCCGTCTTTGCCGACTT-ATCACTTT--GGTTCA-AGGGTTTAGAAGATCCTGAAGA-ATATTTGAGAATGGGTTTTGAAGTCAGTGCTTC----TTCCTTCTTTTTGGACCGTGGTAA
181 NP0019/1-331       --------------------------------------------ACCAAGCTGAGAGTCAGCTTGTGTGCCCAGGAGGGAGGCGTTGGGTCA------GAGCCTCTGGAGGACCCCTGAAGTCTCTTCTCAGTGTTCTCTATCACAGGGAGAGCTGTCAGCCCCTGGAATGTGGTTCT--ATGTCTAGAAAACTATC--CCATAAATAACAGGAAGCCCAAGGTTTACCAA
182 NG0008/1-331       ---------------------------------------------CCTACTTCATGACCACCGGTAACGCTCTAGGATCTGTGAACATGACCACTGGTGTCGATA-ATTTGTTCTACATTGACA---AGTTC-CAAGTAAGGGAAG-T-AAAATAGAGGTTATAAAACTTATTGTCTTTTTTATTTTTTT-----CAAAAGCCATTCTAAAG-GGCTTTAGCAACGAGTGA
183 NP0013/1-331       -----------------------------------------------GCGCACTAGCTCTGCTTTTGCGCGTACGACAACAACTACATTTAAAATTTCTCGA-AACTCATGGCATTTATTGGGAAAGGTTAGTTA---GTTTTATT-TTTTG-----TTTTTAGAGCAGCATTCAATTTAGACTTTTATAAAAGAAATTTCTAATT-TGATCCCTCGTTTATCAAACGATA
184 NP0011/1-331       --------------------------------------------ATTTGTTTCTCAGTGCACTTTCTGGTGTTCCATTTTCTATT-GGGCTCTTTACCCCGCATTTGTTTGCAGATCACTTGCTTGCGCATTTTTA--TTGC-ATT-TTACATATTACACATTATTTGAACGCCGCTGCTGCTGCATCCGTCG--ACGTCGACTGCACTCGCCCCCACGA-GAGAACAGTA
185 NP0022/1-331       -----------------------------------TAACCGATGGGAACACGTCTCCACCAAGACAGCGCTCAGGACTGGTTCTCCTCGTGGCTCCCAATTCAGTCCAGGAGAAGCAGAGATTTTGTCCCATGGTGGGTCATCTGAAGAAGGCACCCCTGGTCAGGG-CAGGCTTCTCAGACC-------CTGAGGCGCCGGCCATGGC---CCCACTGAGACACAGGAA
186 NP0026/1-331       --------------------------------------------CGGATCACCGGCTTTTGGCTGCTCTCACCAAATCAGCTGCAAGAAGATTAGAGCTCAAAAGAATTACA-GAAAGAGAGCC--------TTTTTCTTTTCTTCCTTGTGGG-GTTCCTTTCATTT-CGTGCTCTCCTTTCTCTGCCAGCCAGTCCGTCCGTCCTTGCG--TCCACTGCACCTGCACAC
187 NG0034/1-331       ----------------------------------------TACCTGGGCGGGACGCGCCAGGCCGACTCCCGGCGAGAGGATGGGGCCAGACTTGCGGTCTGCGCTGGCAGG-AAGGGTGGGCCCGACTGGATTCCCCTTTTCTGCTGCGCGGGAGGCCCAGTTGCTG-ATTTCTGCCCGGATTCTGCTGCCCGGTGAG---GTCTTTGC---CCTGCGGCGCCCTCGCCC
188 NG0038/1-331       -----------------------------------------GCAGCTTCACAGAAACCTCATTCGTTTATTCCCTTGTT--TGATTCAGAAGCAGGTGGGACAGGTGAACTTTTGGA---TTGGAACTCGATTTCTGAC--TGGGT-TGGAAGGCAAGAGAGCCCCGAAA--GCTTACATTTTATGTTAGCTGGTGGACTGACGCCAGAAAATGTTGGTGATGCGCTTAGA
189 NR0017/1-331       ------------------------------------------------GTTTACTAAAAATCCGTAAAGAACTTCAATTGT-ACGCCAACT--TAAG-----ACCATGTAACTTTGCATC-CGACTCTCTTTTAGAC-TTAT--CT-CCAATCAAGCCACAATTTGCTAAAGGTACTGACTTCGTTGTTGTCAGAGAATTAGTGGGAGGTATTTACTTTG-GTAAGAGAAA
190 NG0019/1-331       ------------------CGCCCGTTTGGAGTGTGGCGCTACCGAGGAACTGGCAGCATATTGCCCGCTGGCAGGAGCAGGAGTTGGCTCCGCCGGCCACCGTCAACAAGGATGGCTACAAACTCACCCTGGA---CGTCAAGGACTACAGCGAGCTGAAGGT---CAAGGTGCTGGACGAGAGCGTGGTCCTGGTGGAGGCAAAATCGG--AGCAGCAGGAGGCC-----
191 NG0024/1-331       ---------------------------------------------------------------------------------------------------------GATCCAGCAGGTGGGACCCGCCCATCTCAATGTGAA--GGAGAATCCCAAGGAGGCGGTGGAGCAGGACAATGGCAACGATAAGTAGAGGACTCGTTCCGGGAGATGCCCTGCATTATTTAACCAT
192 NG0026/1-331       -------------------------------GCTTTCATTTGCCTTAACGT-TGAGGTGAGCGGGTCCCACTTGCTGAATTTGAATGATGCGCTCC-TTGGACTTG---TCCTCGACGGCCTGCGGCT-TGGGAATACTGACGGTG-AGC---ACGCCATCCGACGACAGCTGCGAGACCACTTGCTCCGCCTTG-TAGCCATCGG-----GAACCTTGTAGCGGCGCACA
193 NG0023/1-331       TGAGGCTGATAAGGTGGCCTCCACCTTGTCCTCCGATGGTGTCCTGACCATCAAGGTGCCCAAGCCACCGGCAATCGAGGATAAGGGCAACGAGCGCATCGTTCAGATCCAGCAGGTGGGACCCGCCCATCTCAATGTGAA--GGAGAATCCCAAGGAGGCGGTGGAGCAGGACAATGGCAACGATAAGTAGAGGACTCGTT-----------------------------
    consensus/100%     .......................................................................................................................................................................................................................................
    consensus/90%                                         ....................................................................................................................................................................................................
    consensus/80%                                              ...............................................................................................................................................................................................
    consensus/70%                                                   .....................................................................r.................y..r......... .....................................................................................
MView 1.47.3, Copyright © Nigel P. Brown, 1997-2002.
```

```
         .         :         .         .         .         .         3         .         .         .] 331
A-CGTG----------------------------------------------------------------------------------------------
A-CG------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
A-CG------------------------------------------------------------------------------------------------
A-CGTGG---------------------------------------------------------------------------------------------
A-CGTG----------------------------------------------------------------------------------------------
A-CGTGG---------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
A-CGTGGA--------------------------------------------------------------------------------------------
A-CGTG----------------------------------------------------------------------------------------------
C-GTGGA---------------------------------------------------------------------------------------------
A-CGTGG---------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
A-CGTGG---------------------------------------------------------------------------------------------
A-CGTGG---------------------------------------------------------------------------------------------
A-CGTGG---------------------------------------------------------------------------------------------
ACCTGA----------------------------------------------------------------------------------------------
A-CGTGGAATATGGC-------------------------------------------------------------------------------------
CC--------------------------------------------------------------------------------------------------
A-CGTGGT--------------------------------------------------------------------------------------------
ACCTGG----------------------------------------------------------------------------------------------
A-CGTGGAATATGGC-------------------------------------------------------------------------------------
CCTGG-----------------------------------------------------------------------------------------------
ACATGGA---------------------------------------------------------------------------------------------
A-CGAGGAATATGG--------------------------------------------------------------------------------------
A-CGTGGAATATGGC-------------------------------------------------------------------------------------
ACGTG-----------------------------------------------------------------------------------------------
ACCTGG----------------------------------------------------------------------------------------------
ATT-------------------------------------------------------------------------------------------------
A-GGTGGAATATGGCA------------------------------------------------------------------------------------
ATCGTGGAATATAGCAGGC---------------------------------------------------------------------------------
ACCTGGAATATGGCGAG-----------------------------------------------------------------------------------
ACCTGGAATATGGCGAG-----------------------------------------------------------------------------------
ACCTG-AATATGG---------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
ACCTGGAATATGGCGAG-----------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
GCGATGT---------------------------------------------------------------------------------------------
TCACCGCTTTCGCC--------------------------------------------------------------------------------------
```

```
CGACACTGTCCCG----------------------------------------------------------------------------------------
CCTCCGC----------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
CTG--------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
CATCTCCCCCCAACCCCG-----------------------------------------------------------------------------------
TATCACCCGCCCTCT--------------------------------------------------------------------------------------
TCCGCACAGCTCAC---------------------------------------------------------------------------------------
TCCGCACAGCTCAC---------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
A-CGTGGAATATGGCAAGAAAACTGAAAATCATGGAAAATGAGAAACATCCACTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAG
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
AGTTAACTGTGGGAATACT----------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
CAGG-------------------------------------------------------------------------------------------------
AGCGGTATTCGCAATATTTTAGTAGCTCGTTA---------------------------------------------------------------------
AAGAAGTGGAGGGGGGGTGGTGGTGGTG-------------------------------------------------------------------------
CTTACTCAAAGGTAATAGTGTAA------------------------------------------------------------------------------
GTTTTCGCTATTCCGACGCGTCTAGT---------------------------------------------------------------------------
AGAATAAGAACAACAACAAATAGAGCA--------------------------------------------------------------------------
TTAGTTTTAAAACACCAAGAACTT-----------------------------------------------------------------------------
AC---------------------------------------------------------------------------------------------------
ATAACTTAAAGAAAAAG------------------------------------------------------------------------------------
TTTCCTCGCCACATATGCATTACCGTCTA------------------------------------------------------------------------
CCGTTT-----------------------------------------------------------------------------------------------
AAAAAAAAAAGGATGGAGGTTAAAAGACG------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
---GCGCATACGCTACAATGACCCGA---------------------------------------------------------------------------
GAGATCGCACATGCCA-------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
GTCAAGT----------------------------------------------------------------------------------------------
GCTTAACTGC--TCATTGCTA--------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
GACTTACTTACTG-GATTTTG--------------------------------------------------------------------------------
CAAAAATGAAGTATTTCCTTTTT------------------------------------------------------------------------------
CTGCT------------------------------------------------------------------------------------------------
AGAAAATA---------------------------------------------------------------------------------------------
```

```
TAAGTTTTACCATGACATGATCAGA------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------------
TCAAAGAAATTATTGGGG-------------------------------------------------------------------------------------------------------------------
CGCCT--------------------------------------------------------------------------------------------------------------------------------
TGATATCCAAGGTCAACTCC-----------------------------------------------------------------------------------------------------------------
GATTTGGAAAAAGCTGAAAA-----------------------------------------------------------------------------------------------------------------
ATATATCTTACTTTTTTTTTTCTC-------------------------------------------------------------------------------------------------------------
TGCTCATGGGACAGGGC--------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------------
CCCA---------------------------------------------------------------------------------------------------------------------------------
AACTACAATAA--------------------------------------------------------------------------------------------------------------------------
TGACTTACACATAGACGACCATCACACCAC-------------------------------------------------------------------------------------------------------
GTGATTGTACCTGAGTTCAATTCTAGCGC--------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------------
GCAATTACACTCG---TCAATTC--------------------------------------------------------------------------------------------------------------
GTTACCGAGGAAGAACTCA------------------------------------------------------------------------------------------------------------------
TGGTAGATGGATCGATGGCAAACA-------------------------------------------------------------------------------------------------------------
GCATGTTTAGAGCAAGCGCCTTTGTGAG---------------------------------------------------------------------------------------------------------
TTAAATTTAACGCGGAAGCTT----------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------------
ACATTTCAGCAATATATATATATATTTC---------------------------------------------------------------------------------------------------------
TTACGGCATTGATATC---------------------------------------------------------------------------------------------------------------------
AGTCCATAGAGGGCTATGGTGAAAA------------------------------------------------------------------------------------------------------------
ATTGTAC------------------------------------------------------------------------------------------------------------------------------
TGGGTTGAG----------------------------------------------------------------------------------------------------------------------------
ACAACCCAAATGACACCGTATGGGGTACG--------------------------------------------------------------------------------------------------------
TTAAAATAATTTTGATAAGA-----------------------------------------------------------------------------------------------------------------
GAGAGTGCGTTCAAGGCT-------------------------------------------------------------------------------------------------------------------
TATA-TAAATGCAAAAACTGCATAACCACTTT-----------------------------------------------------------------------------------------------------
TGTTTGATTCAG-------------------------------------------------------------------------------------------------------------------------
TTTTTAATCCGGACAAGCTCATTTGCGT---------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------------
ATGGAAGAGG---------------------------------------------------------------------------------------------------------------------------
AGAAAGAGAA---------------------------------------------------------------------------------------------------------------------------
TTTCTGCAACAGCTATG-AGCATTGTGCAAACATATT------------------------------------------------------------------------------------------------
ATTTTAAAACAA-------------------------------------------------------------------------------------------------------------------------
TTCTCTTGAACCGTAAATATC----------------------------------------------------------------------------------------------------------------
ACGCTCTAG---GATCTGTGAA---------------------------------------------------------------------------------------------------------------
GCCGTACGCAGTTGTCGAACTTGGT------------------------------------------------------------------------------------------------------------
GAATCTCGCCAATATTTAAG-----------------------------------------------------------------------------------------------------------------
TTAGAGCAATCATTGAAAGTACTAGATA---------------------------------------------------------------------------------------------------------
GCCTTACTATAATTATCGCTATCGGC-----------------------------------------------------------------------------------------------------------
TGGT-ACGGCACCCA----------------------------------------------------------------------------------------------------------------------
T--GCCTTAACGTTGAGGTGAGC--------------------------------------------------------------------------------------------------------------
AATATTTTGTAAAATCATATATAATCAAATT------------------------------------------------------------------------------------------------------
TAAGTTT---GCATTTCTCTTTAATCT----------------------------------------------------------------------------------------------------------
CAACTCCACTCAATTCAGA------------------------------------------------------------------------------------------------------------------
AATTGTCC-CGTACGACCTCTTCAATAATAACACAT-------------------------------------------------------------------------------------------------
GCTAATTTATTACTTATACATAAA-------------------------------------------------------------------------------------------------------------
AATGTGGA-----------------------------------------------------------------------------------------------------------------------------
AATCAATCTAGAG------------------------------------------------------------------------------------------------------------------------
CGTGCAGGAGACACTCAAG------------------------------------------------------------------------------------------------------------------
CATCTCGCGCC--------------------------------------------------------------------------------------------------------------------------
GAGTTCATCC---------------------------------------------------------------------------------------------------------------------------
GATTTCTTCAGTTTCCCACCCGGGA------------------------------------------------------------------------------------------------------------
TTTCTATTACT--------------------------------------------------------------------------------------------------------------------------
GGACATGAAGCACTGGCCTT-----------------------------------------------------------------------------------------------------------------
CAATCGTAATGTAGTTGCCTTACA-------------------------------------------------------------------------------------------------------------
ACAGTCCACTTATTACTACTGCGGCC-----------------------------------------------------------------------------------------------------------
```

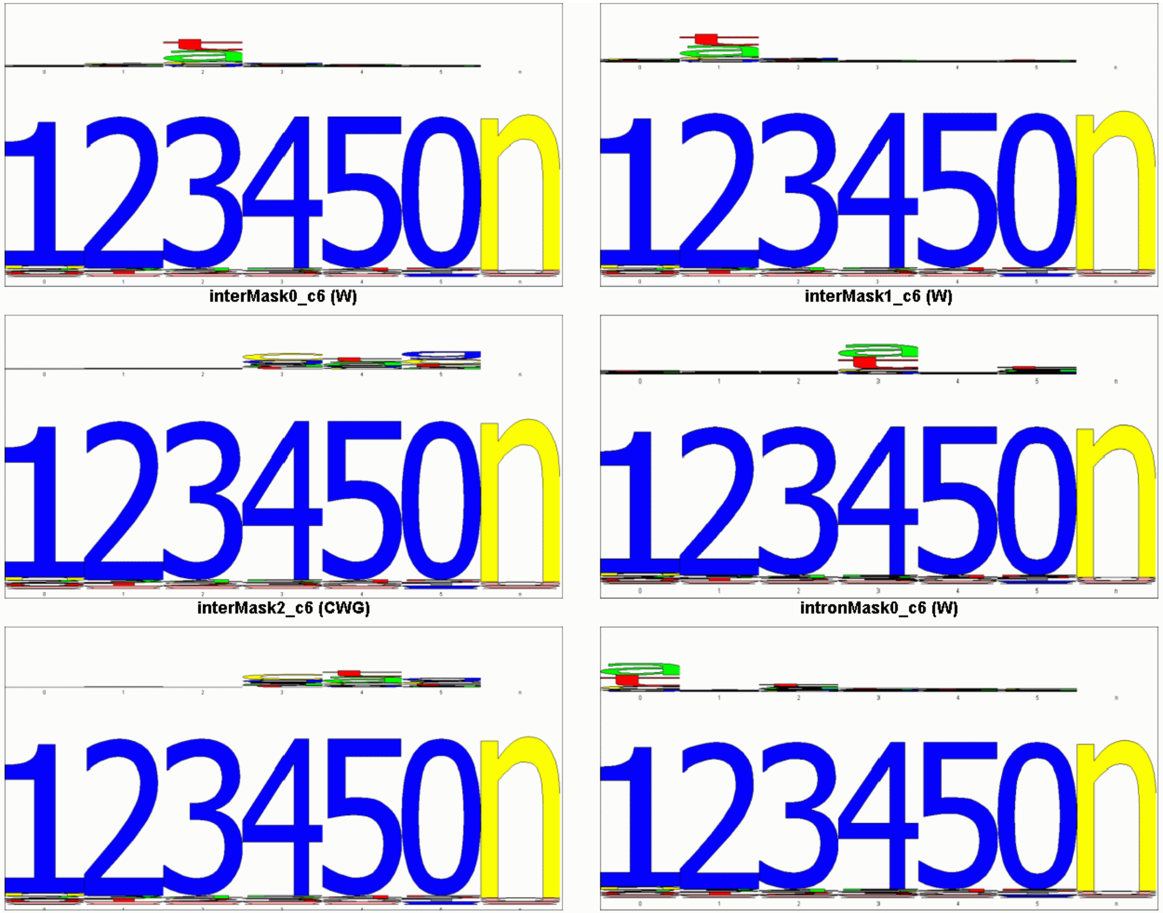
```
CGAGCGGCGGCCCCAGTGTGCGG------------------------------------------------------------------------------
CGAGTGGAAGCAAAGAAC-----------------------------------------------------------------------------------
GGTTTCTTAGGCTACCAATACGAAT----------------------------------------------------------------------------
GGCATAATGGAAAATC-------------------------------------------------------------------------------------
TGGTGGTTGATTACAACAACACGAGTGGG------------------------------------------------------------------------
ATTTGCACCGCTT----------------------------------------------------------------------------------------
CCTTGCCC-CGCACGACACTTTCA-----------------------------------------------------------------------------
GGG--------------------------------------------------------------------------------------------------
CTGAAGATG--------------------------------------------------------------------------------------------
CGGGTTGAGGCAGGAAACCCTGGGTTCCCT-----------------------------------------------------------------------
TTTGTCTAAAGAGTAATACTAGAGATAA-------------------------------------------------------------------------
CTAAC--ATAATTAACTTAAGCAGCC---------------------------------------------------------------------------
CAGGTCAGGCCCGG---------------------------------------------------------------------------------------
GATGTGTCGCACTTCAAGCCCAGCGA---------------------------------------------------------------------------
ACTCTAAGGTCAAGTTTGTCAAGGAG---------------------------------------------------------------------------
ATCTCTGCTGTACAGGATGTTCTA-----------------------------------------------------------------------------
ACGAAT--GTAAAACTTTATGATTTCAAAG-----------------------------------------------------------------------
ACAAAGCTATATTCATAATTTTTTCTCT-------------------------------------------------------------------------
ATTTAAGGAGCTGCGAAGGTCC-------------------------------------------------------------------------------
AGGGCCGCGCCAGAGC-------------------------------------------------------------------------------------
CAGGTCA---CCCCGACCCGCACTGTTCTA-----------------------------------------------------------------------
CAGGGCAAAGTCCCAGCC-----------------------------------------------------------------------------------
ATTAAATGGCGTTATTGGTGT--------------------------------------------------------------------------------
AGGAAGACGATGGTGATGGTGTCGCTTGGGAT---------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
TTATCAAAGTCATACATCTGTTTTATAAGCTGTAGTTATCCAAGGACACTTCACTCATACACAATAGCCATTAAGGG------------------------
AAAGTGGCGCATGATGT------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------
.....................................................................................................
............................
........................
....................
```

B. Cyclical Hidden Markov models trained from various types of sequences

The models illustrated in this appendix follow a 3-field naming scheme:

[training source][unique training ID]_c[no. of wheels in model architecture]

For the *unique training ID* field, digits represent a specific training run from the respective training source.



interMask0_c6 (W)

interMask1_c6 (W)

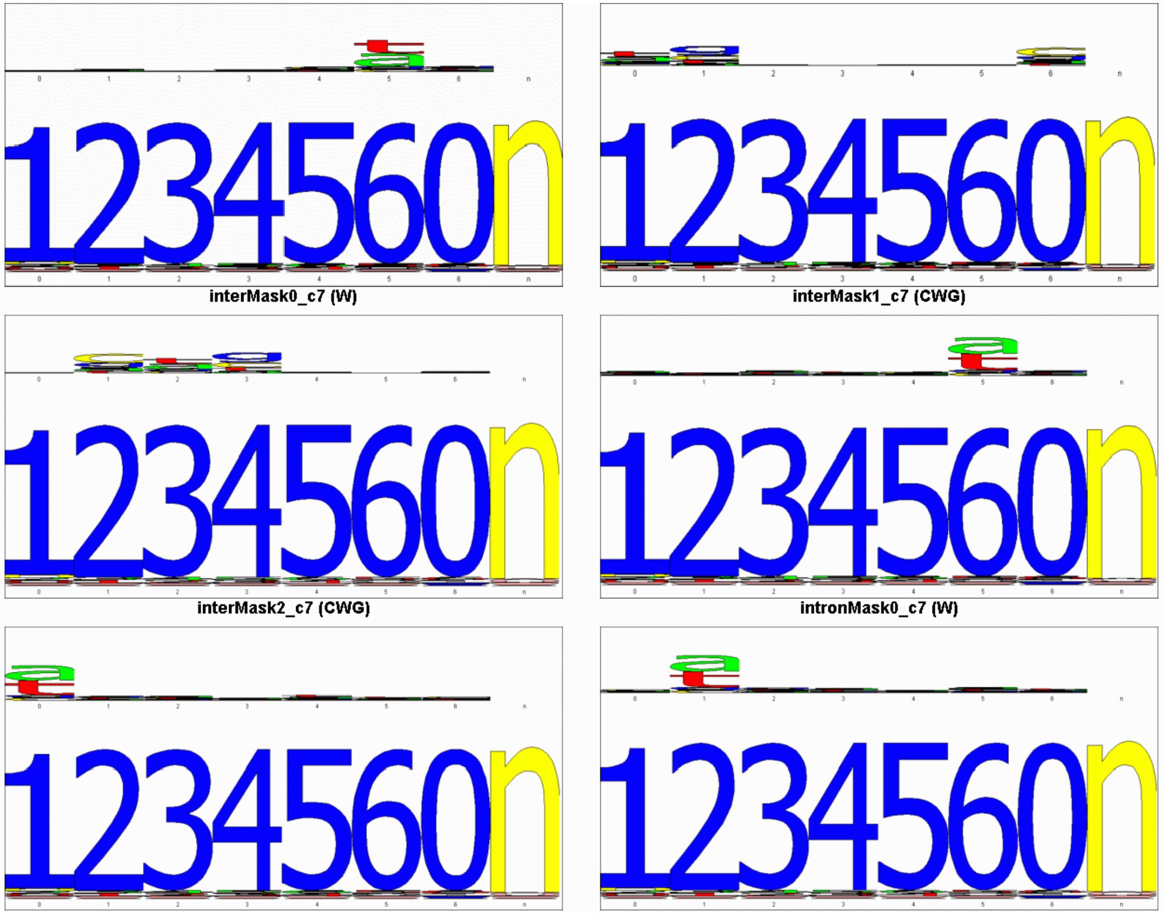
interMask2_c6 (CWG)

intronMask0_c6 (W)

intronMask1_c6 (CWG)

intronMask2_c6 (W)

6 state wheel models



interMask0_c7 (W)

interMask1_c7 (CWG)

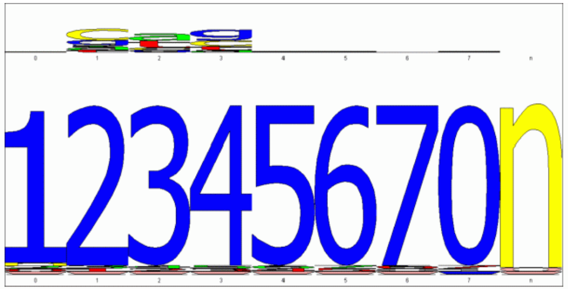
interMask2_c7 (CWG)

intronMask0_c7 (W)

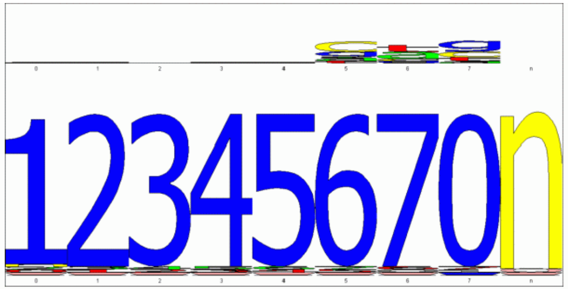
intronMask1_c7 (W)

intronMask2_c7 (W)

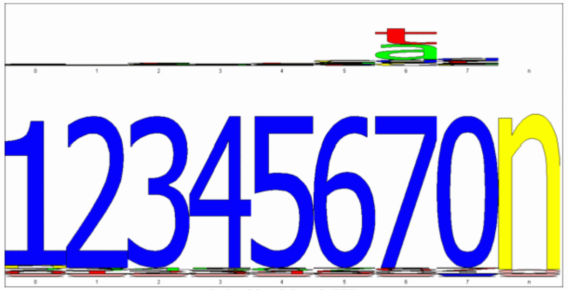
7 state wheel models



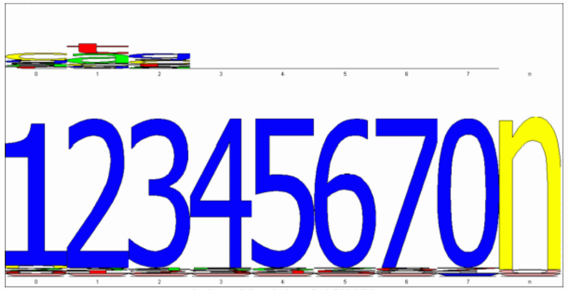
interMask0_c8 (CWG)



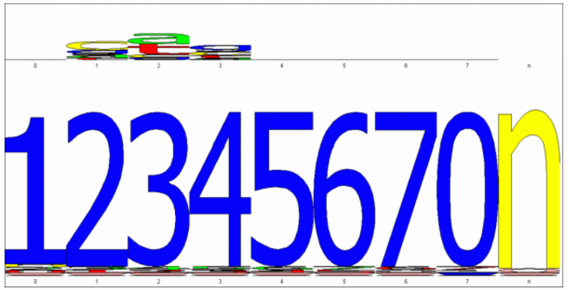
interMask1_c8 (CWG)



interMask2_c8 (W)

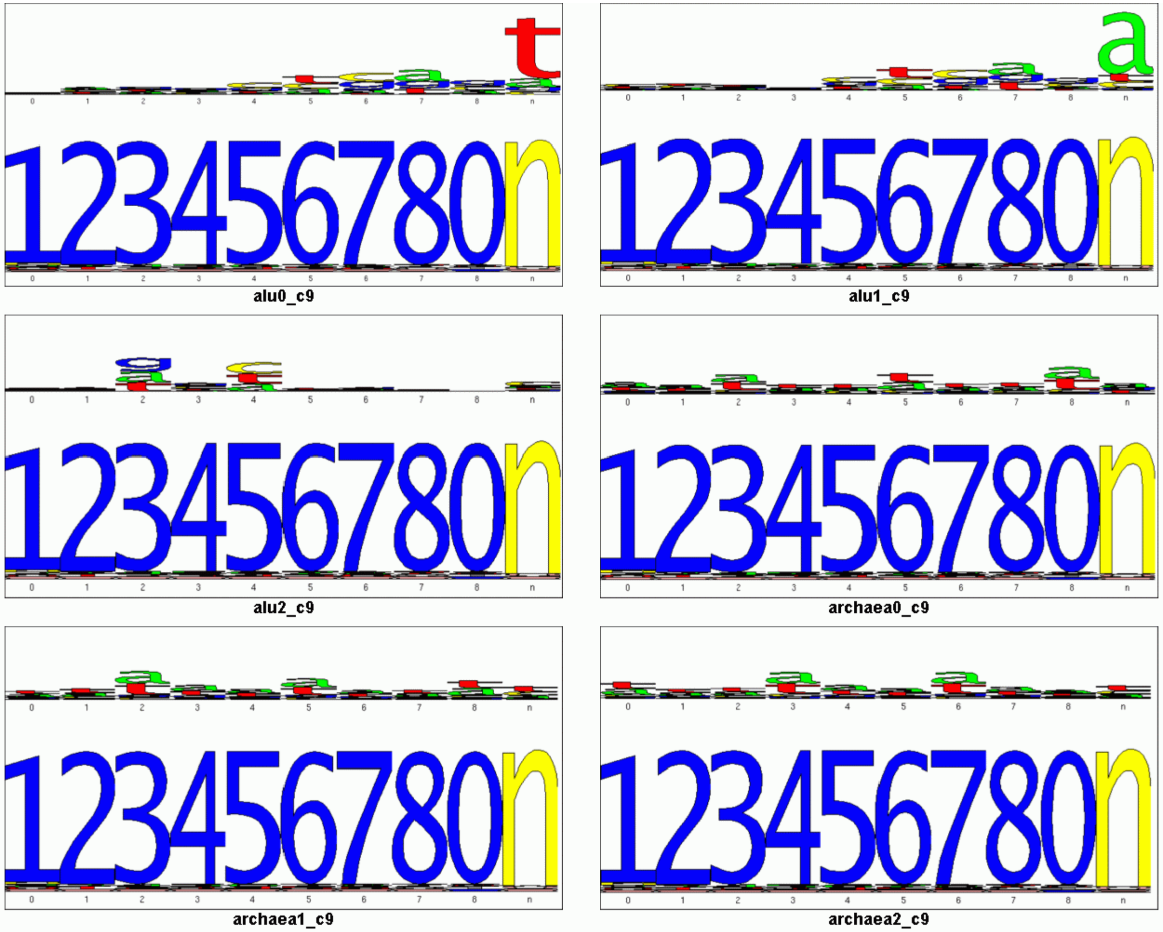


intronMask1_c8 (CWG)



intronMask2_c8 (CWG)

8 state wheel models



alu0_c9

alu1_c9

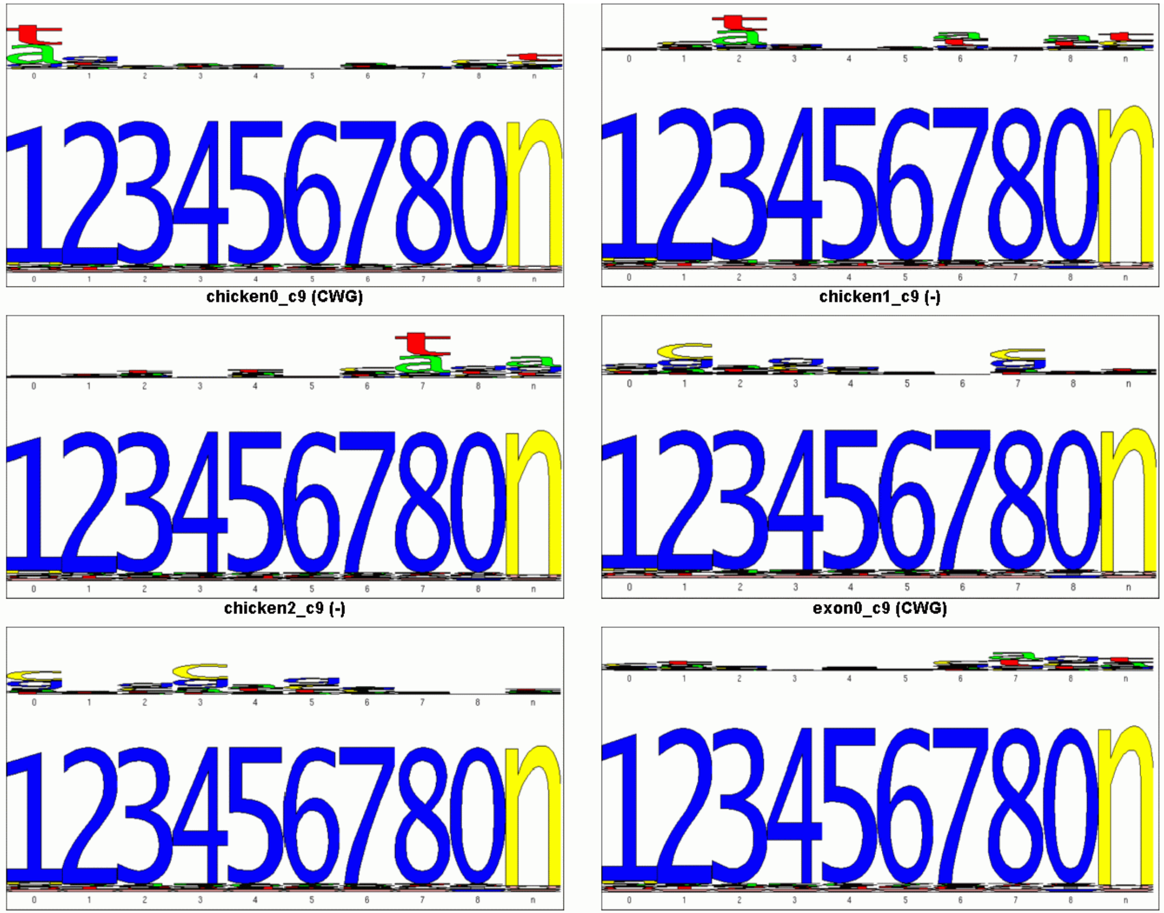
alu2_c9

archaea0_c9

archaea1_c9

archaea2_c9

9 state wheel models



chicken0_c9 (CWG)

chicken1_c9 (-)

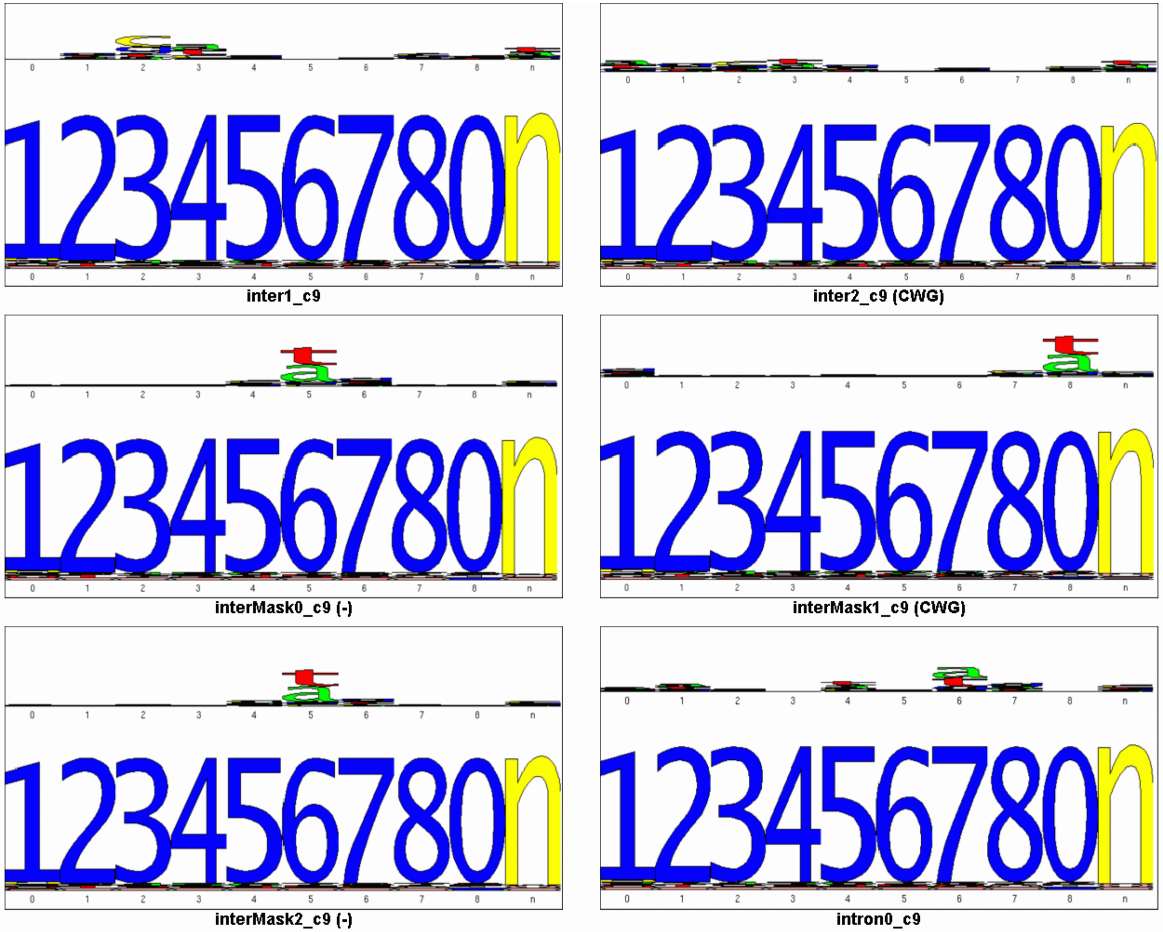
chicken2_c9 (-)

exon0_c9 (CWG)

exon2_c9 (CWG)

inter0_c9 (CWG)

9 state wheel models



inter1_c9

inter2_c9 (CWG)

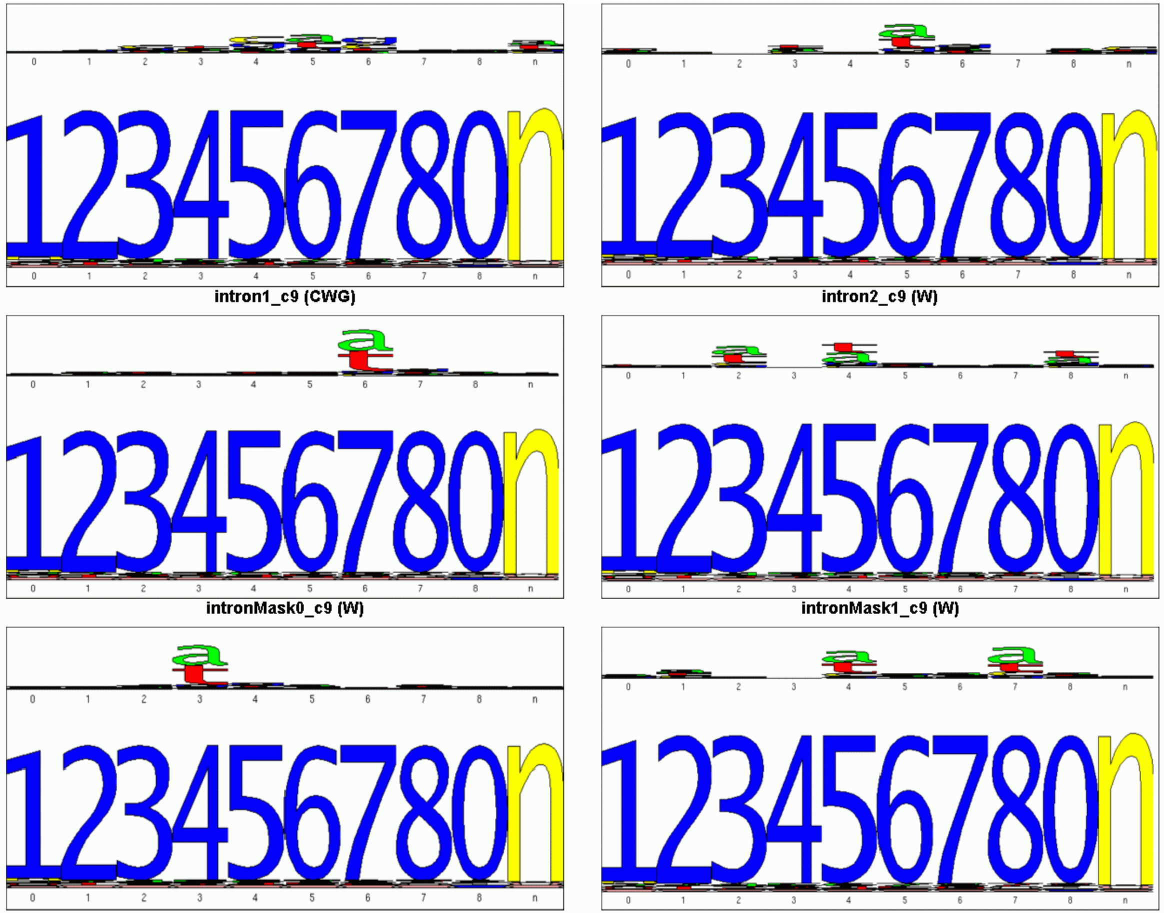
interMask0_c9 (-)

interMask1_c9 (CWG)

interMask2_c9 (-)

intron0_c9

9 state wheel models



intron1_c9 (CWG)

intron2_c9 (W)

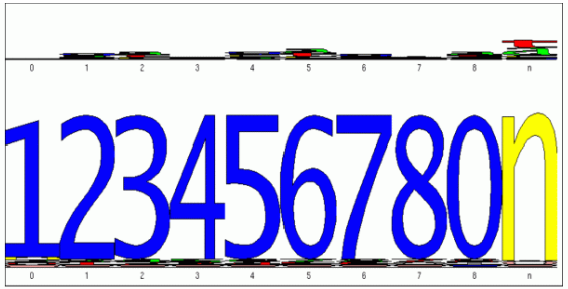
intronMask0_c9 (W)

intronMask1_c9 (W)

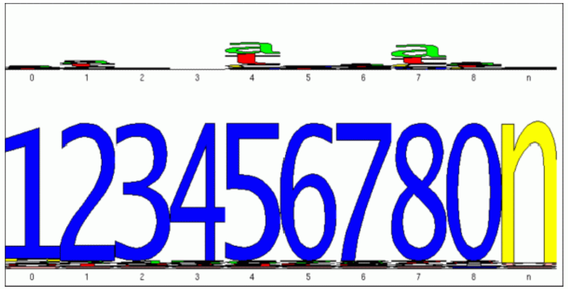
intronMask2_c9 (W)

levitsky0_c9 (W)

9 state wheel models

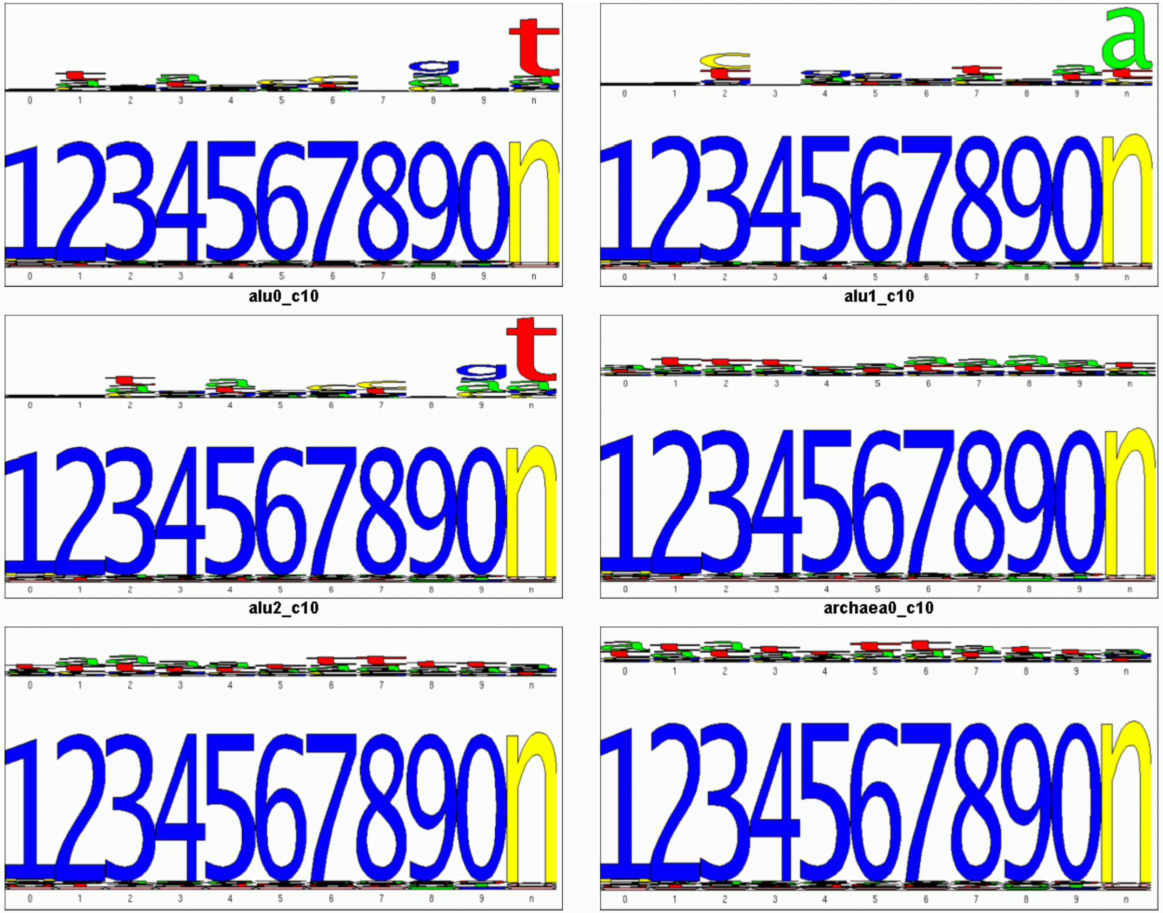


levitsky1_c9



levitsky2_c9

9 state wheel models



alu0_c10

alu1_c10

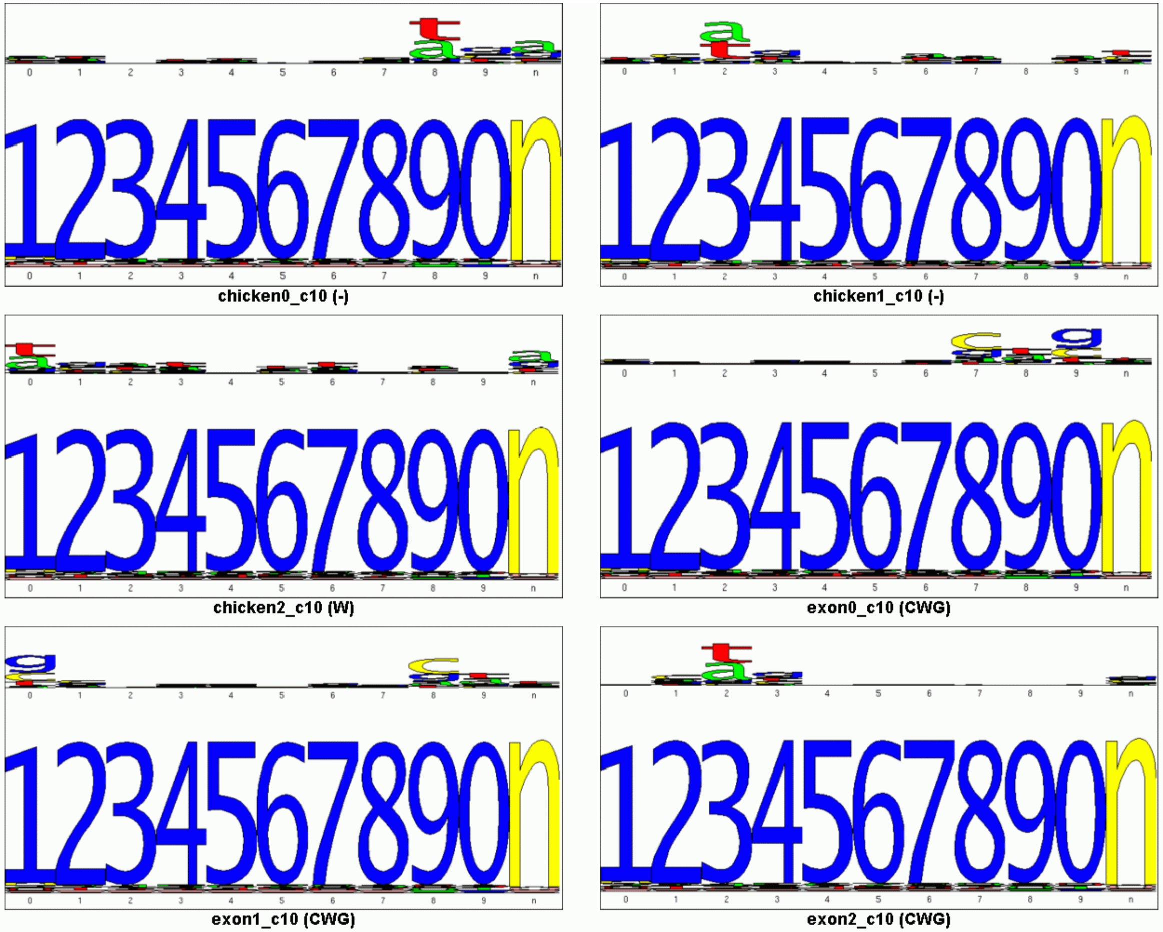
alu2_c10

archaea0_c10

archaea1_c10

archaea2_c10

10 state wheel models



chicken0_c10 (-)

chicken1_c10 (-)

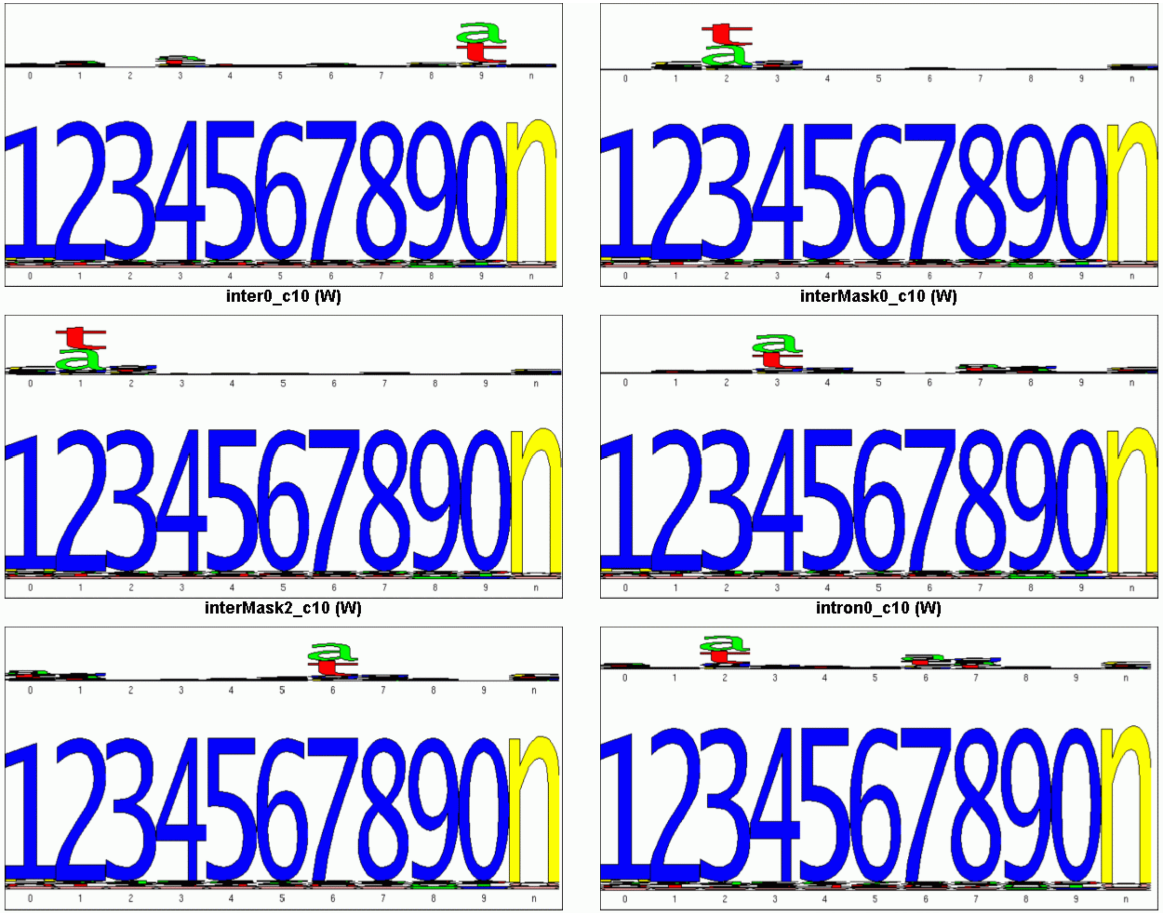
chicken2_c10 (W)

exon0_c10 (CWG)

exon1_c10 (CWG)

exon2_c10 (CWG)

10 state wheel models



inter0_c10 (W)

interMask0_c10 (W)

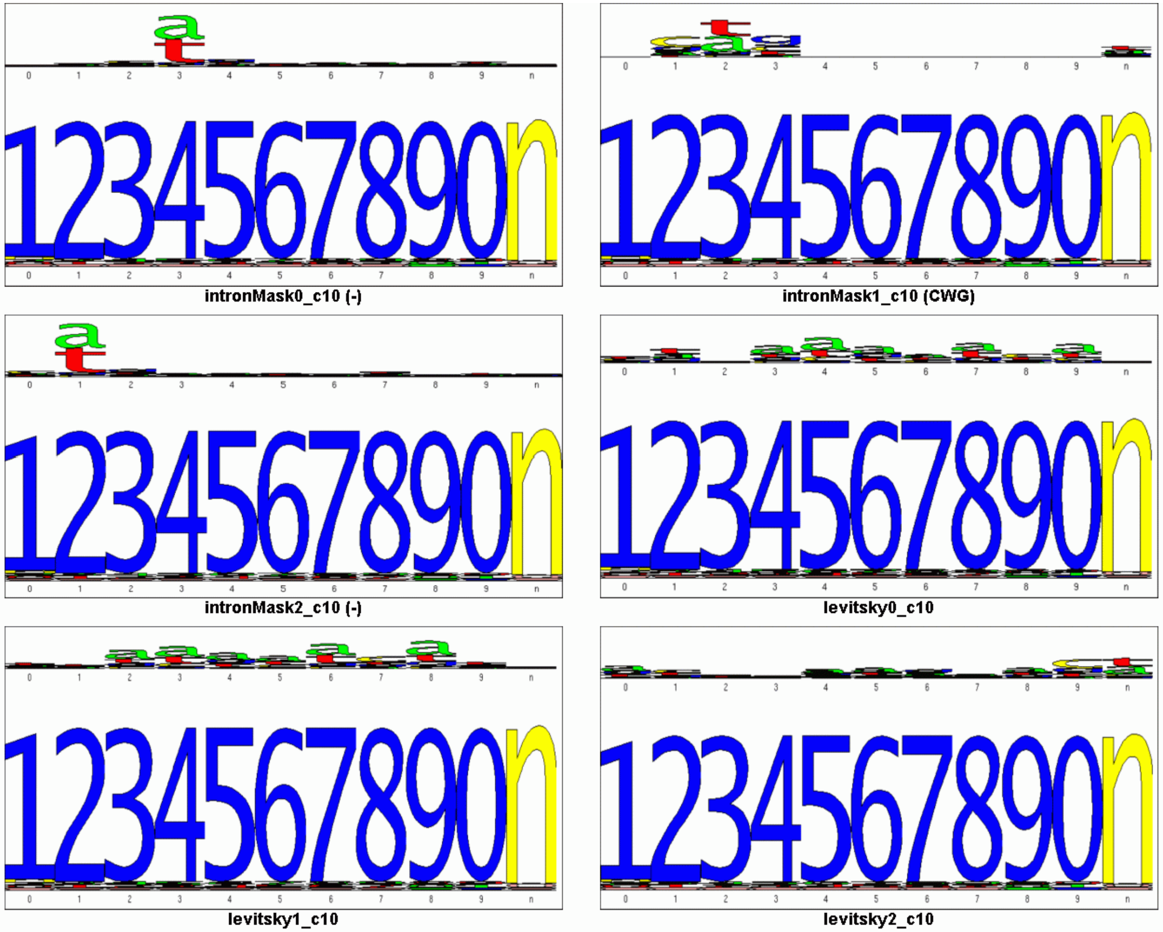
interMask2_c10 (W)

intron0_c10 (W)

intron1_c10 (W)

intron2_c10 (W)

10 state wheel models



intronMask0_c10 (-)

intronMask1_c10 (CWG)

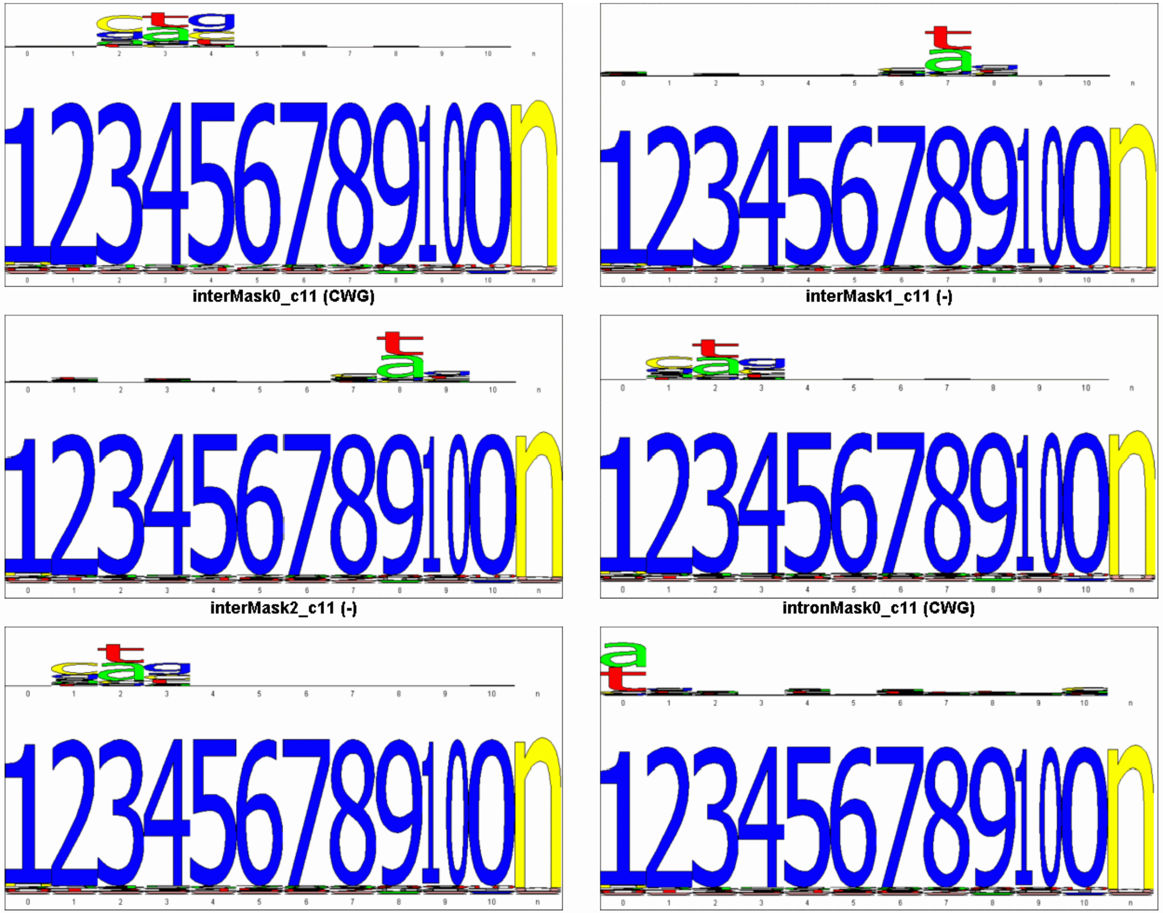
intronMask2_c10 (-)

levitsky0_c10

levitsky1_c10

levitsky2_c10

10 state wheel models



interMask0_c11 (CWG)

interMask1_c11 (-)

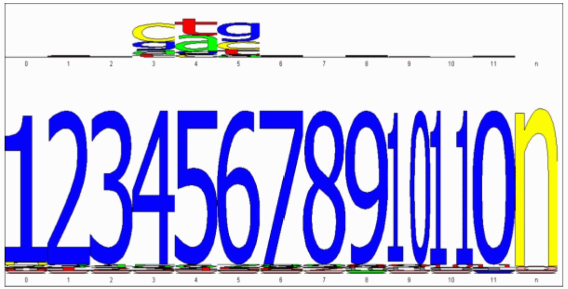
interMask2_c11 (-)

intronMask0_c11 (CWG)

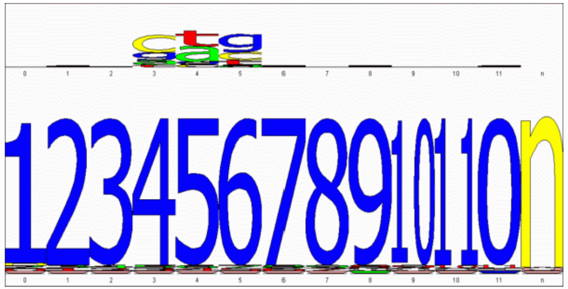
intronMask1_c11(CWG)

intronMask2_c11 (W)

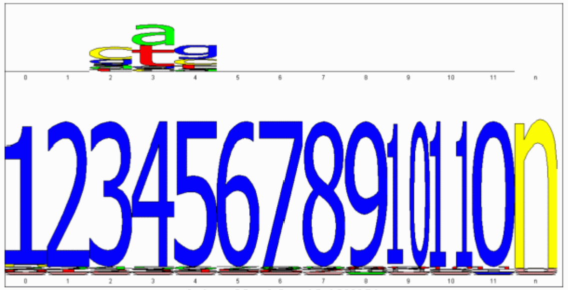
11 state wheel models



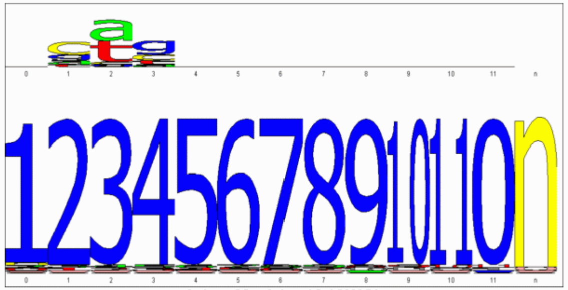
interMask1_c12 (CWG)



interMask2_c12 (CWG)



intronMask0_c12 (CWG)



intronMask1_c12 (CWG)

12 state wheel models