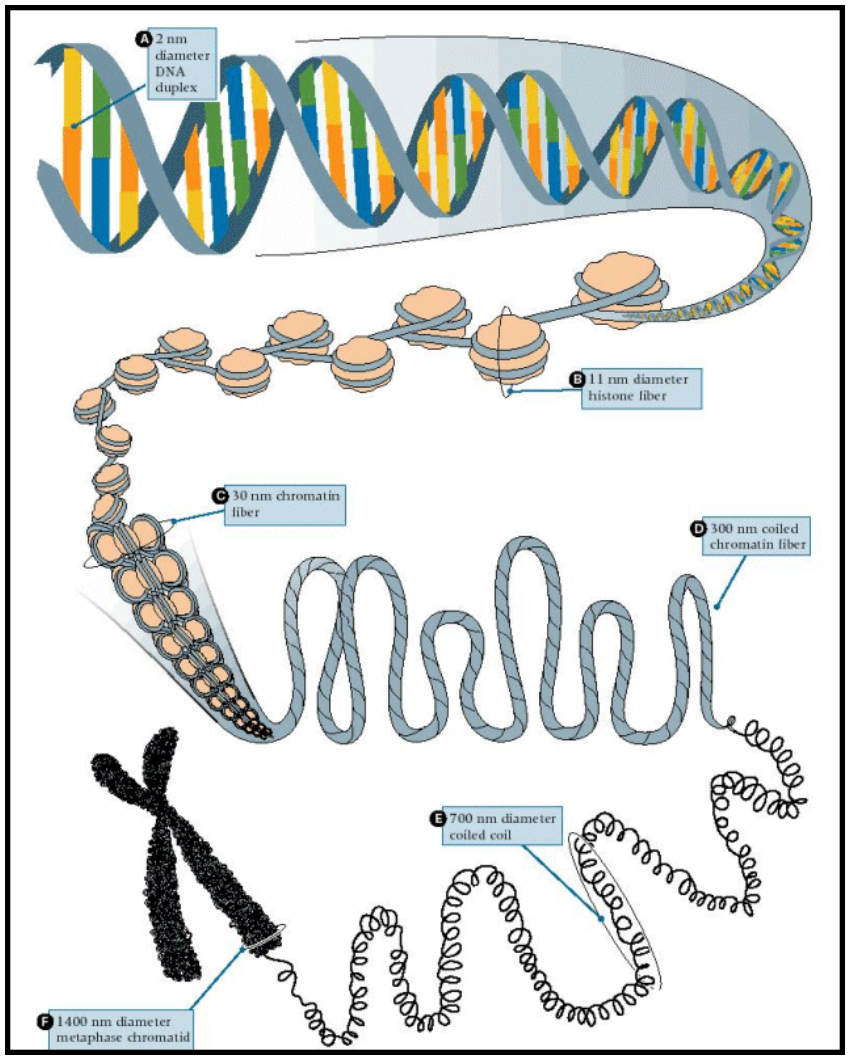
1 A General Introduction to Nucleosomes and Nucleosome Positioning

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).

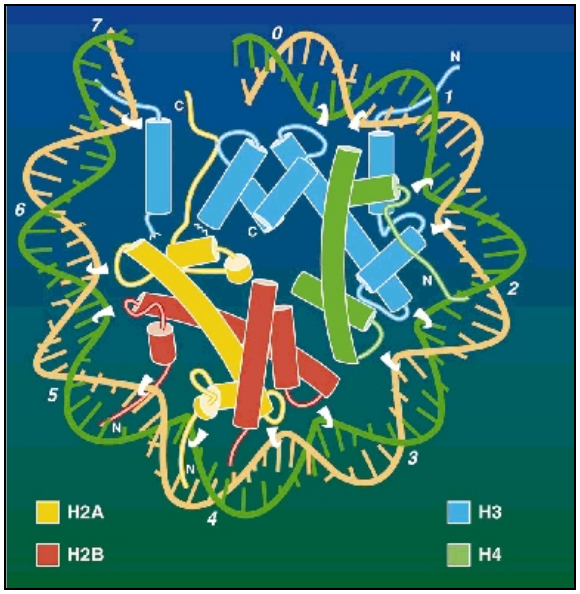
Figure 1.1: A hierarchical view of chromatin structure. Reproduced figure (Hartl & Jones, 1998).



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 $CA_{4-5}T$ 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 $CA_{4-5}T$ 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.

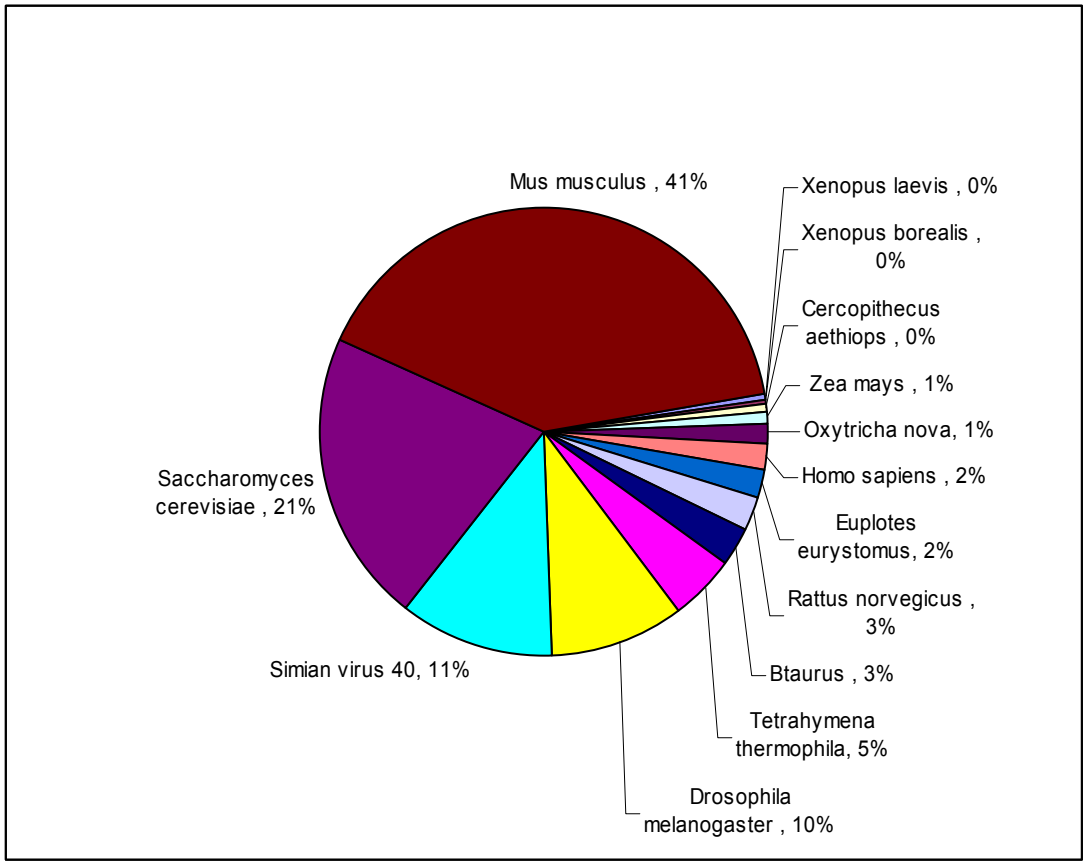
For some of the experiments performed in this thesis (Chapter 3 and Chapter 5), additional chicken genomic sequences were required which could be used as a background test set to these chicken core DNA sequences. Two chicken genomic clones were available for this purpose: AC092403 (144,369 bp) and AC120196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

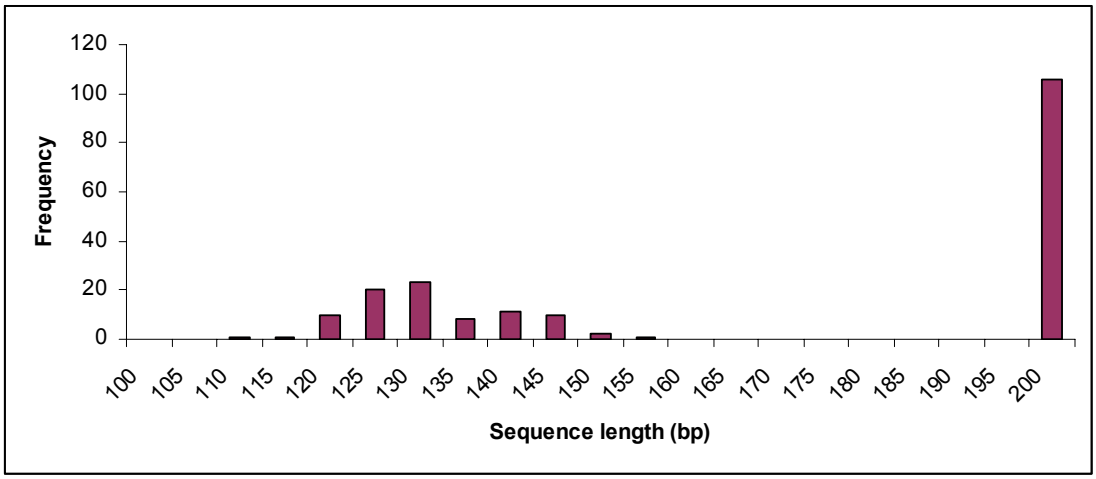


Table 1.1: Accuracy of different nucleosome mapping methods (Ioshikhes & Trifonov, 1993).

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.

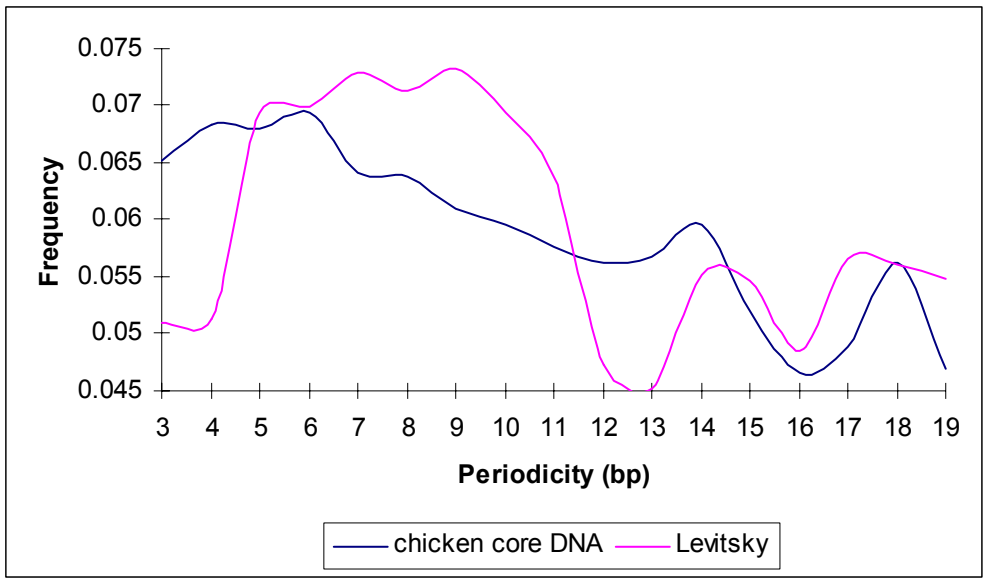
1.9.2 [AA/TT] rotational positioning pattern obtained using multiple sequence alignment

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.