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).