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

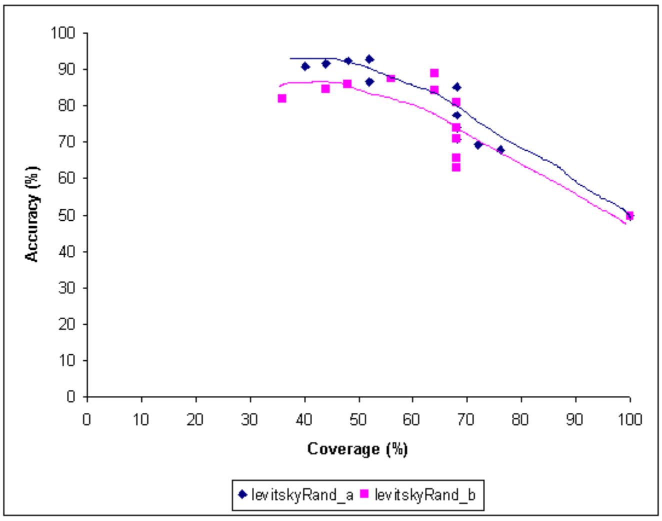
(a) Model levitskyRand_a

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).



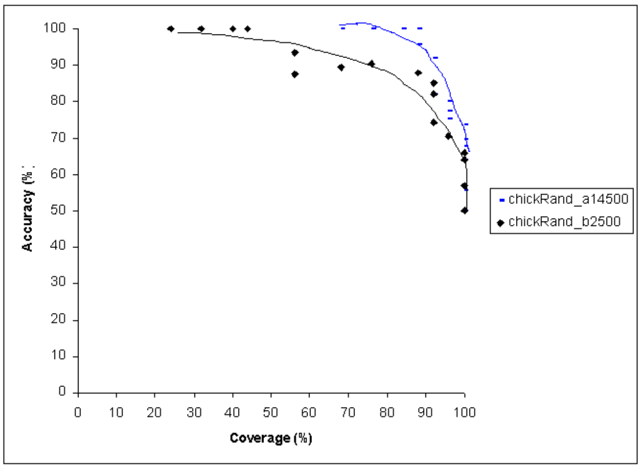
However, it was observed that the models had learnt multiple CpG motifs in the negatively-weighted matrices; these are highlighted yellow in Table 5.2. An important fact known about long runs of CpG motifs is that they occur very rarely in eukaryotic genomes (Cooper & Gerber-Huber, 1985; Sved & Bird, 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.

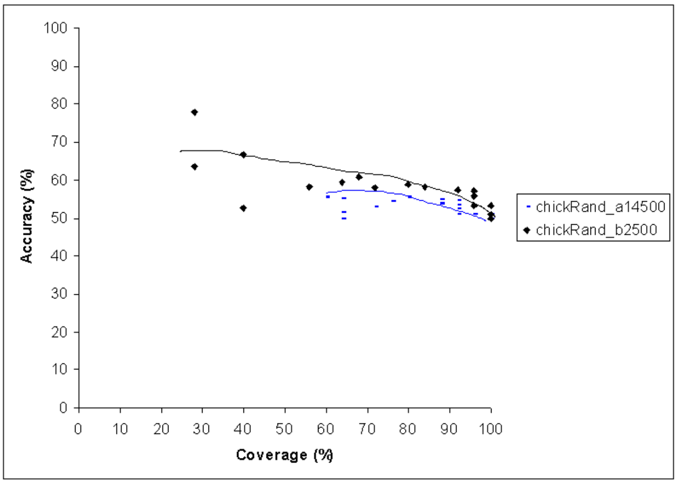
5.3.2 Anchored training results using randomized chicken nucleosome sequences as negative data

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)

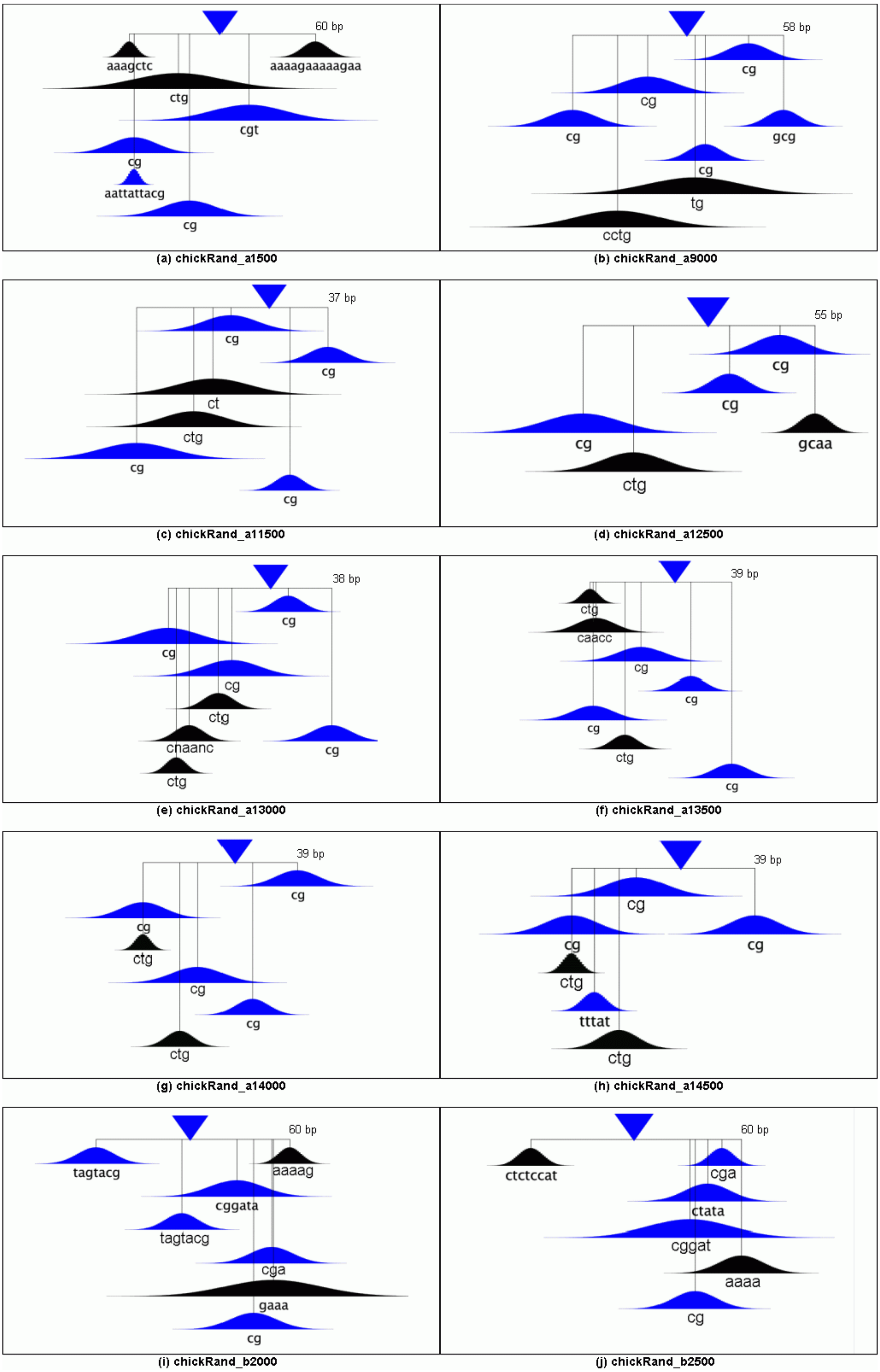


(b)

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

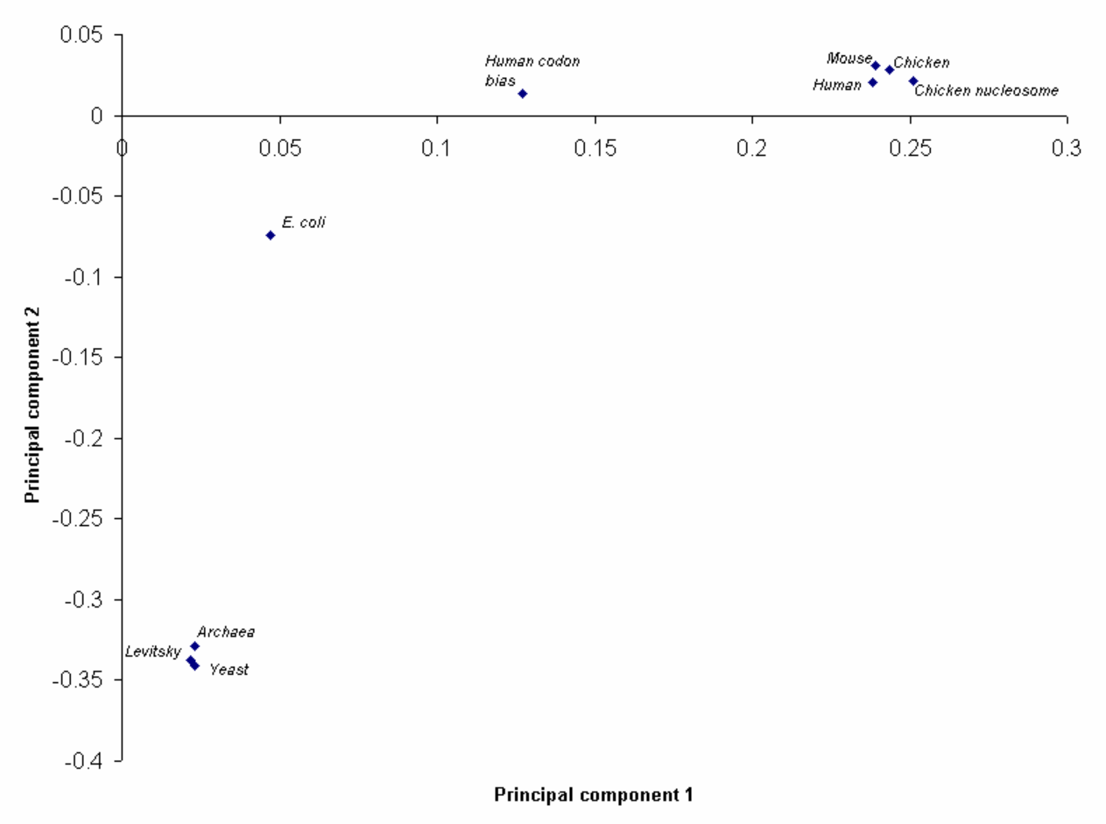
(i) chickRand_b2000

(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).

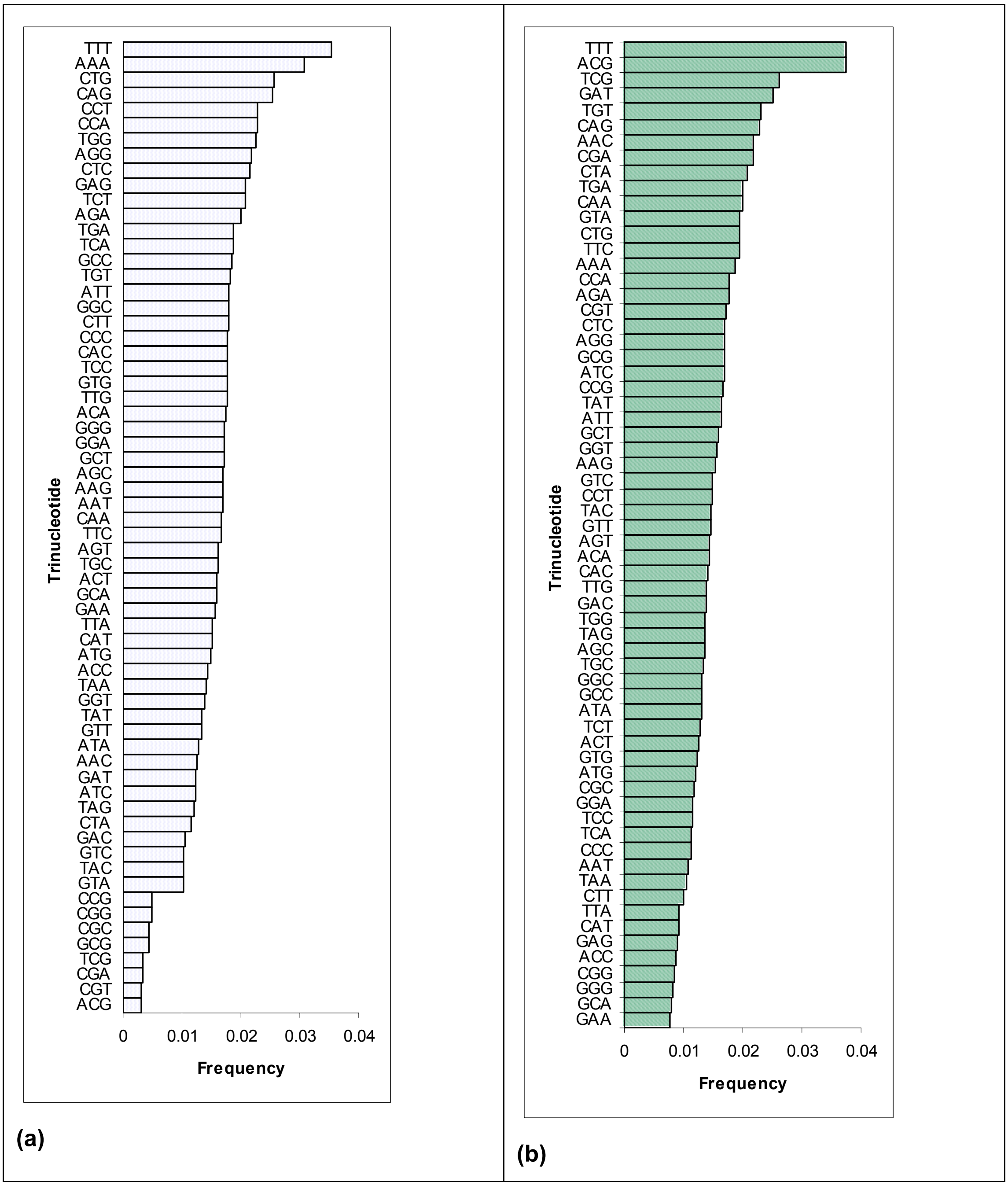
Figure 5.4: Principal components analysis of the background trinucleotide distributions of different genomes and the 2 nucleosome datasets.



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.



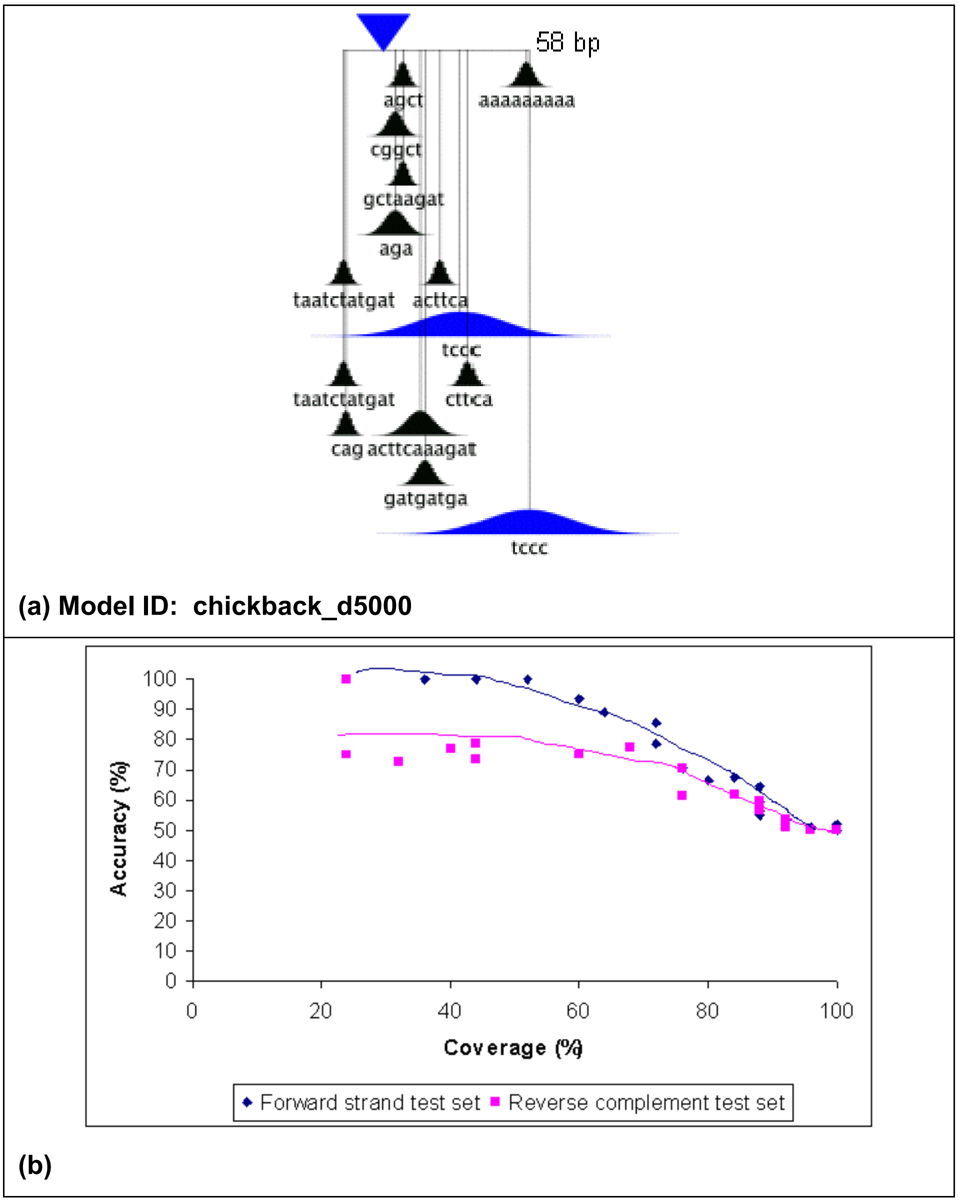
The background trinucleotide distribution for the Levitsky data was quite far from the distribution of the higher organisms along the principal components axes (Figure 5.4, Figure 5.5(b)); the correlation co-efficient between the human and Levitsky distributions was 0.02. The means of the distributions 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.

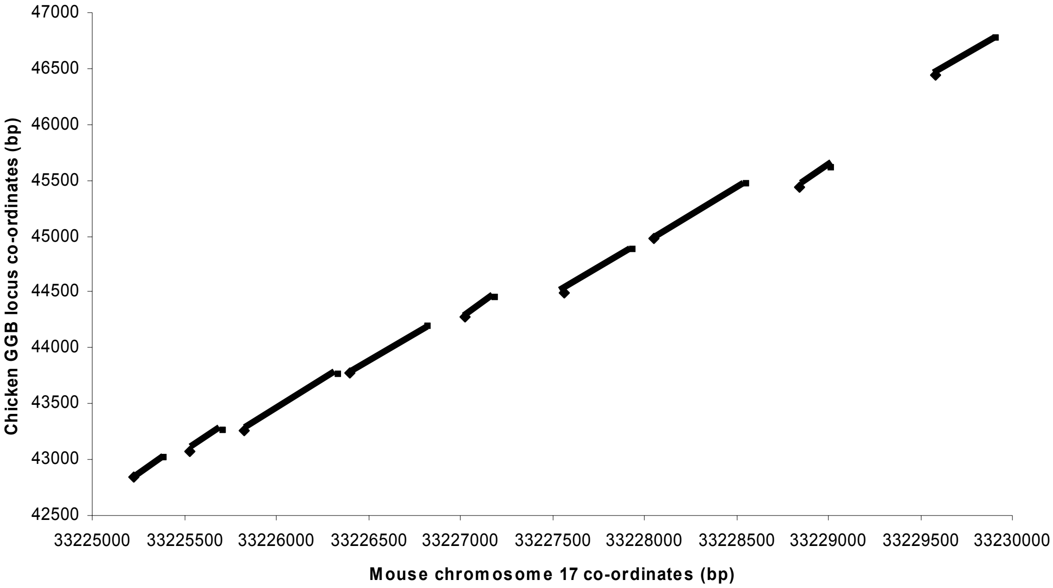
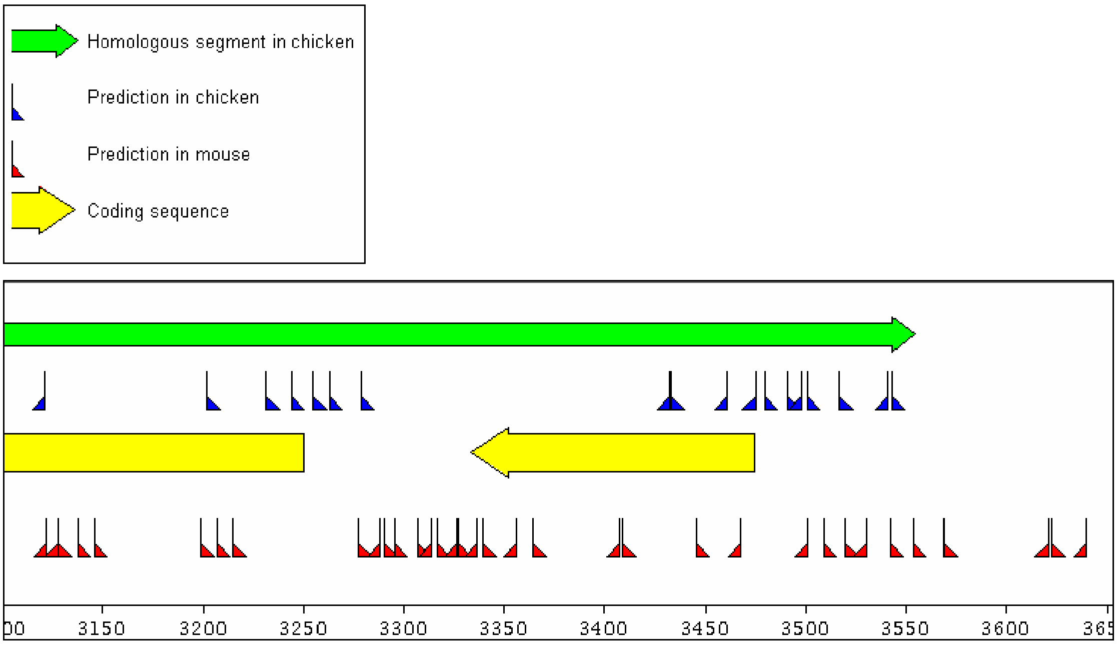


Figure 5.8: Prediction using model *chickBack_d5000* (Figure 5.6(a)) on the chicken GGB locus and homologous regions in mouse. The sequence co-ordinate axis represents the mouse sequence.

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.