# COMPUTATIONAL ANALYSIS OF NUCLEOSOME POSITIONING DATASETS

Ross Macdonald Fraser

PhD

University of Edinburgh
2006

# Abstract

Chromatin is a complex of DNA and histone proteins that constitutes the elemental material of eukaryotic chromosomes. The basic repeating sub-unit of chromatin, the nucleosome core particle, is comprised of approximately 146 base pairs (bp) of DNA wrapped around an octamer of core histones. Core particles are joined together by variable lengths of linker DNA to form chains of nucleosomes that are folded into higher-order structures. The specific distribution of nucleosomes along the DNA fibre is known to influence this folding process. Furthermore, on a local level, the positioning of nucleosomes can control access to DNA sequence motifs, and thus plays a fundamental role in regulating gene expression. Despite considerable experimental effort, neither the folding process nor the mechanisms for gene regulation are currently well understood.

Monomer extension (ME) is an established *in vitro* experimental technique which maps the positions adopted by reconstituted core histone octamers on a defined DNA sequence. It provides quantitative positioning information, at high resolution, over long continuous stretches of DNA sequence. This technique has been employed to map several genes: globin genes (8 kbp), the beta-lactoglobulin gene (10 kbp) and various imprinting genes (4 kbp).

This study explores and analyses this unique dataset, utilising computational and stochastic techniques, to gain insight into the potential influence of nucleosomal positioning on the structure and function of chromatin. The first section of this thesis expands upon prior analyses, explores general features of the dataset using common bioinformatics tools, and attempts to relate the quantitative positioning information from ME to data from other commonly used competitive reconstitution protocols. Finally, evidence of a correlation between the *in vitro* ME dataset and *in vivo* nucleosome positions for the beta-lactoglobulin gene region is presented.

The second section presents the development of a novel method for the analysis of ME maps using Monte Carlo simulation methods. The goal was to use the ME datasets to simulate a higher order chromatin fibre, taking advantage of the long-range and quantitative nature of the ME datasets.

The Monte Carlo simulations have allowed new insights to be gleaned from the datasets. Analysis of the beta-lactoglobulin positioning map indicates the potential for discrete disruption of nucleosomal organisation, at specific physiological nucleosome densities, over regions found to have unusual chromatin structure *in vivo*. This suggests a correspondence between the quantitative histone octamer positioning information *in vitro* and the positioning of nucleosomes *in vivo*.

Further, the simulations demonstrate that histone density-dependent changes in nucleosomal organisation, in both the beta-lactoglobulin and globin positioning maps, often occur in regions involved in gene regulation. This implies that irregular chromatin structures may form over certain biologically significant regions.

Taken together, these studies lend weight to the hypothesis that nucleosome positioning information encoded within DNA plays a fundamental role in directing chromatin structure *in vivo*.

# Declaration

I do hereby declare that this thesis was composed by myself and that the work described within is my own, except where explicitly stated otherwise.

Ross Fraser

# Acknowledgements

My thanks must go to my co-supervisors, Dr Martin Simmen and Dr James Allan, for their invaluable help and support throughout my PhD, who time and again went the extra mile during the many vicissitudes throughout my studies.

Thanks are also due to the various people who contributed to the superb datasets I've been analysing for the past few years, and in particular to Dr Colin Davey and Dr Marieta Gencheva, both of whom took this "wet biology" novice of a physics graduate under their wing to show him the ropes of monomer extension.

My thanks also go to the following people:

Dr Tim Brown, whose help, advice and support were invaluable, and without which I'd never have gotten close to finishing. Thanks to my fellow Allan PhDs, Jennifer and Gillian and to the George Square crew, Lorna, Mark and David.

I'd especially like to thank family and friends who have provided encouragement and support during my time at Edinburgh, especially my parents, brother, Brenda, Paul, Matthew, and Roxaneh.

Finally, my thanks go to the MRC for funding the first three years of my PhD studies.

# Abbreviations

| | |
|---|---|
| BLG | β-lactoglobulin gene |
| bp | Base pairs |
| bp/nuc | Base pair per nucleosome (a measure of nucleosome density) |
| globin | Combined $\beta^A$- and ε-globin gene regions |
| M | Number of MCS simulated. |
| MC | Monte Carlo |
| MCS | Monte Carlo Step(s) or Sweep(s) |
| ME | Monomer Extension |
| MN | Micrococcal Nuclease |
| N | Simulated Number of Nucleosomes |
| nbps | Number of base pairs |
| PSD | Power Spectral Density |
| SS | Single-Stranded |

# Contents

# List of Figures

CHAPTER 1

# An Introduction to Nucleosome Positioning

## 1.1 Chromatin and the Nucleosome

Genetic information, which is encoded into the DNA sequence, is packaged and organised within eukaryotic cells into structures known as chromosomes (Figure 1.1). The packaging of the DNA is facilitated by the formation of a nucleoprotein complex called chromatin. It is within this context that the DNA sequence information is accessed by the various biological machinery required to allow cells to function.

The fundamental repeating unit of chromatin, the nucleosome, is a complex of DNA and histone proteins. The nucleosome plays a fundamental role in the packaging of DNA in eukaryotic cells. The nucleosome "core particle", the sub-repeating unit of chromatin, consists of approximately 1.65 left-handed turns of DNA, or ~146 bp, wound around a cylindrical protein complex called the histone octamer. The histone octamer itself is compromised of 8 highly conserved proteins: a central tetramer of H3 & H4, flanked by two H2A/H2B dimers (Figure 1.2). Early, low resolution attempts to solve the crystal structure of the core particle (Klug *et al.*, 1980) led to more accurate studies at 7 Å resolution (Richmond *et al.*, 1984; Uberbacher *et al.*, 1988). Within the last 10 years, the crystal structure of the nucleosome core has been resolved to considerably higher resolution: 2.8 Å in 1997 (Luger *et al.*, 1997) and more recently to 1.9 Å (Davey *et al.*, 2002). The structure of the core particle, and its constituent elements, is therefore relatively well understood.

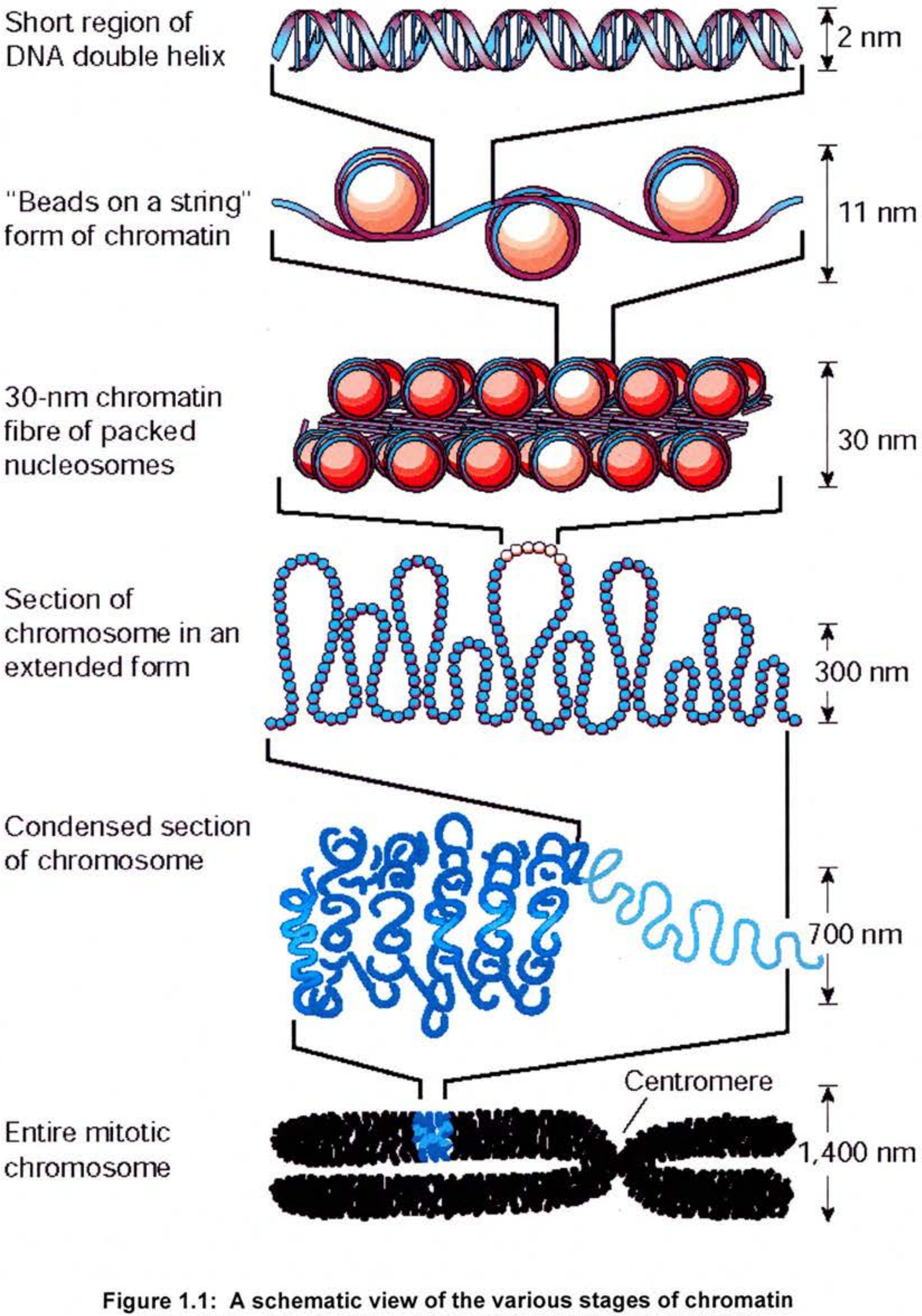

Short region of DNA double helix
2 nm
"Beads on a string" form of chromatin
11 nm
30-nm chromatin fibre of packed nucleosomes
30 nm
Section of chromosome in an extended form
300 nm
Condensed section of chromosome
700 nm
Centromere
Entire mitotic chromosome
1,400 nm

**Figure 1.1: A schematic view of the various stages of chromatin structure. Adapted from (Felsenfeld and Groudine, 2003)**

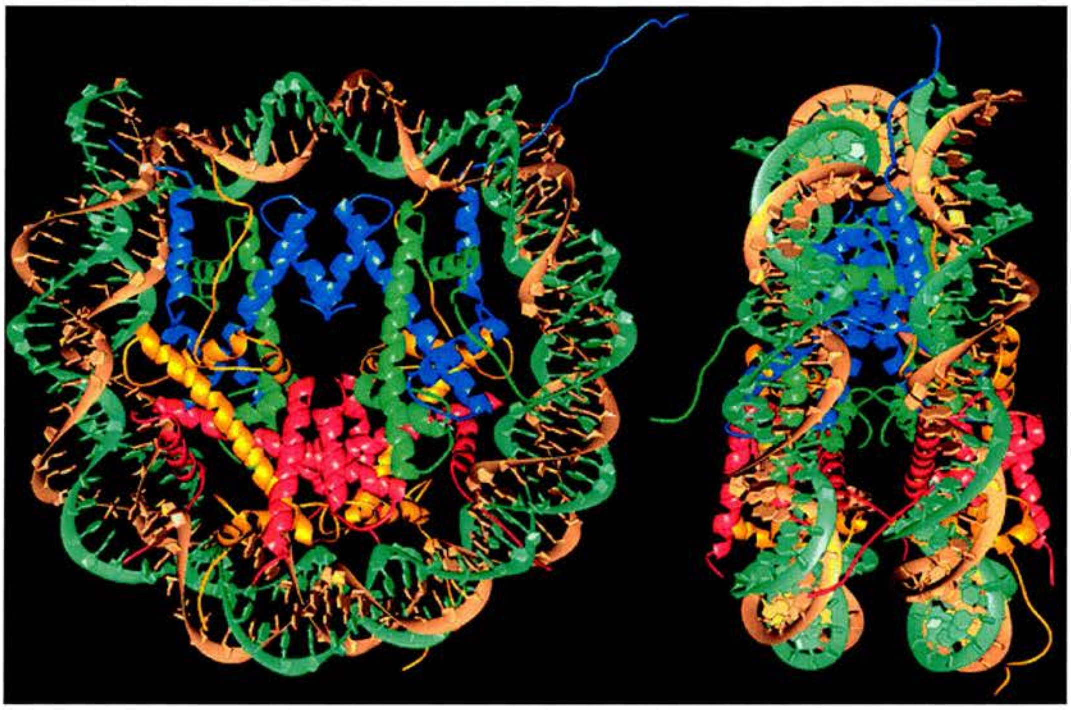


**Figure 1.2: The crystal structure of the nucleosome core particle at 2.8Å resolution. Ribbon traces for the 146 bp DNA sugar-phosphate backbones are represented in brown and turquoise. The eight histone proteins are represented thus: H3 in blue, H4 in green, H2A in yellow, and H2B in red. The left view is down the DNA superhelix axis, and the right view is perpendicular to the superhelix axis. The pseudo-dyad axis is aligned vertically with the DNA centre at the top. Adapted from (Luger *et al.*, 1997)**

This first level of compaction, commonly referred to as the "beads on a string" form of chromatin or the 11 nm fibre with the DNA wrapped around the core histone octamer, condenses the DNA by a factor of ~6. This level of compaction, as it is readily visible by microscopy techniques, is well understood.

To form the chromatosome, another histone, H1, known as the "linker histone", is required. The addition of the linker histone sequesters a further 20bp of sequence into the DNA-protein structure, and stabilises the nucleoprotein structure. It follows that the chromatosome is comprised of the core histone octamer, one molecule of H1, and approximately 168 bp of DNA. Histone H1 is located at the site where the DNA enters and leaves the histone octamer (Satchwell and Travers, 1989).

The DNA bent around the histone octamer is the "nucleosomal DNA", whereas the DNA joining adjacent nucleosomes is referred to as the "linker DNA". The length of the linker DNA determines the nucleosome spacing, the distance between adjacent nucleosomes. Unlike nucleosomal DNA, which is a constant length, linker DNA is highly variable *in vivo* and *in vitro*, ranging from between ~20 to over 100 bp in physiological conditions (Van Holde, 1989). The nucleosome itself, *sensu stricto*, comprises the chromatosome with a variable length of linker DNA, although nucleosome and histone octamer will be used interchangeably throughout the thesis.

The next level of packaging, third from the top in Figure 1.1, is the 30 nm fibre. This fibre is formed when, in appropriate conditions, the array of nucleosomes is coiled into a fibre of ~30 nm in diameter (Van Holde, 1989; Wolffe, 1998). Once in this state, the DNA is compacted by a factor of 40. Most of the DNA within eukaryotic cells is packaged into a fibre of this form (Langmore and Schutt, 1980).

The fundamental constituents of the 30 nm fibre were first observed by electron microscopy, using either low salt conditions (Oudet *et al*., 1975) or the removal of the linker histones (Thoma and Koller, 1977) to unpack the fibre to the "beads-on-a-string" or 11 nm fibre. The fibre can refold itself back into the more compact 30 nm form with the addition of mono- or divalent cations (Thoma *et al*., 1979) and thus it was determined that the chromatin fibre is reliant on electrostatic interactions to maintain its form (Clark and Kimura, 1990). *In vivo*, however, the fibre is compacted by the linker histones, which reduce the electrostatic free energy of the fibre by displacing bound cations and reducing the residual charge (Blank and Becker, 1995).

Whereas the external structure of the 30 nm fibre is well characterised, the exact internal structure is still largely unknown (van Holde and Zlatanova, 1996) and is the subject of a substantial ongoing research effort, for example (Schalch *et al*., 2005). A wide range of experimental techniques have been employed to ascertain the structure of the 30 nm fibre, ranging from electron microscopy to neutron diffraction, sedimentation analysis to electric and flow linear dichroism. Based on this evidence, there have been many different models proposed, although none fits all the available data.

The proposed models fall into two general classes, each with variants: 3 helical models and one non-helical. The helical models include the solenoid (Finch and Klug, 1976; Thoma *et al.*, 1979; Graziano *et al.*, 1994; Daban and Bermudez, 1998), the twisted/helical ribbon (Worcel *et al.*, 1981; Woodcock *et al.*, 1984) and the zigzag/crossed-linker model (Williams *et al.*, 1986; Schalch *et al.*, 2005). Lastly, there is the non-helical "superbead" model (Hozier *et al.*, 1977; Renz *et al.*, 1977; Zentgraf and Franke, 1984). Two of these models, the solenoid (Figure 1.3 a)) and the crossed-linker (Figure 1.3 b)), are the best supported by the available evidence and the consensus view is that the 30 nm fibre structure is likely to be similar to one of these two.

The precise spatial distribution of nucleosomes within the 30 nm fibres will have impact on nucleosome interactions (see the top down views in Figure 1.3). The linker DNA in the solenoid model is wrapped around the inside of the solenoid and therefore must be bent to connect each nucleosome with its neighbour (Figure 1.3 a)). However, in the cross-linker model, DNA is thought to criss-cross between neighbouring nucleosomes in a direction roughly perpendicular to the fibre axis with H1 located on the inside of the fibre. For the solenoid family, nucleosomes are stacked with no close face-to-face contacts, whereas there is face-to-face stacking in the cross-linker model. It is clear, therefore, that the spatial arrangement will have an influence on the nature of the internucleosomal interactions within the fibre.

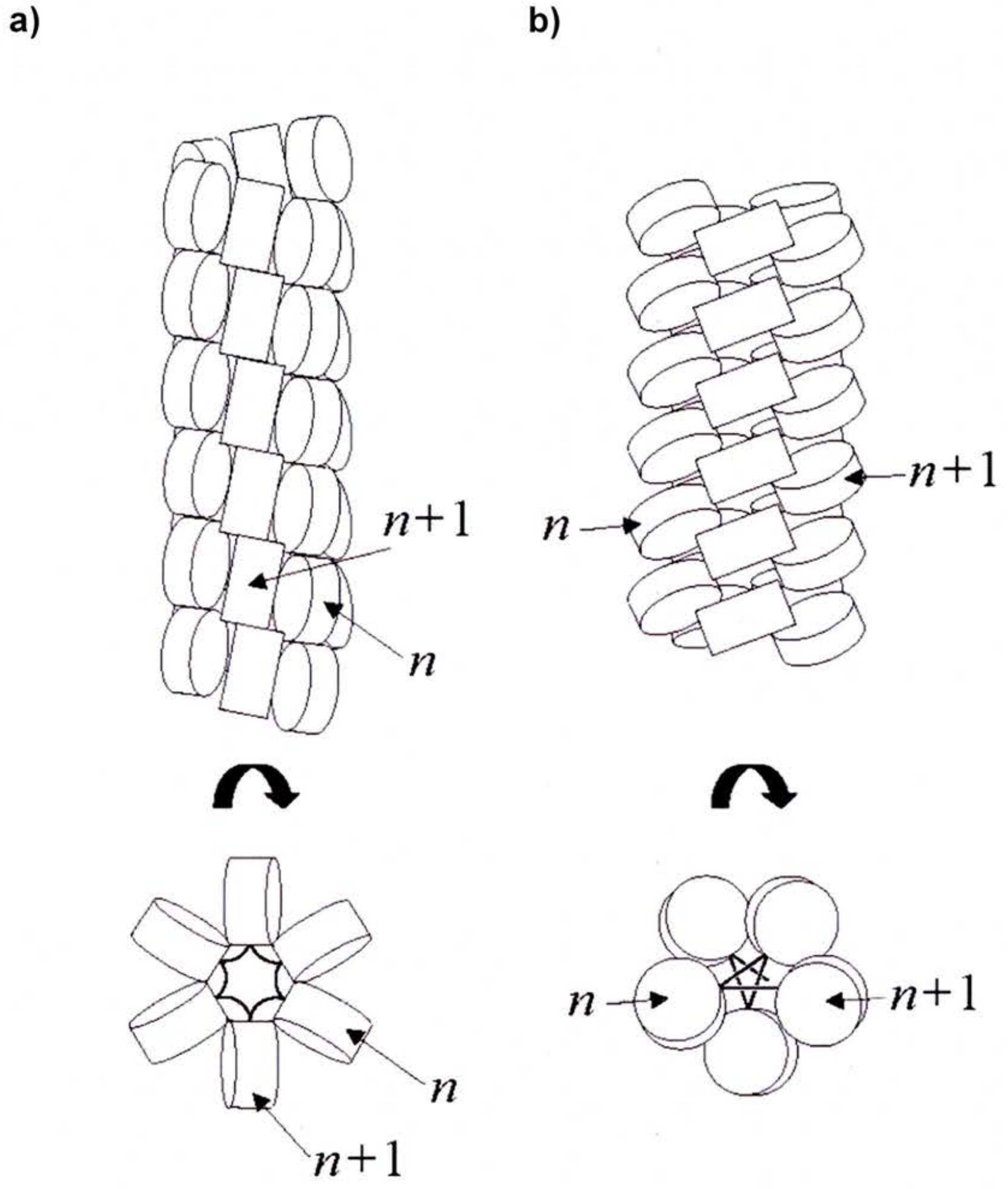


**Figure 1.3: Idealised models for the 30 nm fibre. a) The "solenoid" model, while b) presents the "crossed linker" model. Both models are shown from the side (upper illustration) and from the top (lower illustration). The top view highlights the different linker geometries and the relationship between the neighbouring nucleosomes *n* and *n*+1. Real fibres will tend to be less regular, depending on the nucleosomal arrangement on the underlying DNA sequence. Adapted from (Schiessel, 2003)**

Recent work using micromanipulation techniques, where a single chromatin fibre is stretched and its behaviour under stress observed (Cui and Bustamante, 2000; Bennink *et al.*, 2001; Brower-Toland *et al.*, 2002), appears to favour the cross-linker model, as does the recently resolved x-ray structure of a crystallised tetranucleosome (Schalch *et al.*, 2005).

In both the solenoid and cross-linker models, the linker histone H1 is believed to be located mainly towards the middle of the 30 nm fibre, where it may play a role in stabilising the fibre (Staynov, 2000). Studies of the crystal structure of the nucleosome core particle by Luger *et al.* (1997) have suggested that neighbouring nucleosomes can interact with one another: the H4 basic tail is not locally bound to a nucleosome and can interact with acidic regions of neighbouring H2A and H2B dimers. Consequently, it was suggested by Luger *et al.* that histone H4 also has an influence on higher order structure and stability.

The variation in the length of linker DNA may be important for the diversity of gene regulation seen in nature and it has been observed that average nucleosome spacing does indeed vary depending based on the expression pattern of the gene and the developmental stage (Evans *et al.*, 1990). However, there is some evidence to suggest that the structure chromatin adopts is independent of the average nucleosomal spacing (Allan *et al.*, 1984; Widom *et al.*, 1985; Kornberg and Lorch, 1999).

To form the higher order structures beyond the 30 nm fibre, as shown in the lower half of Figure 1.1, a number of non-histone proteins interact with the histones, and these allow the 30 nm fibre to undergo further levels of compaction. These interactions result in compaction by a factor of between ~1000 to ~10000 depending on the particular type of chromatin: chromatin is organized into condensed regions (heterochromatin) and more open "euchromatin". Heterochromatin is more tightly packaged, and so access to the underlying DNA sequence is consequently additionally inhibited.

## 1.2 Nucleosome Positioning

### 1.2.1 Introduction

Nucleosomes have been shown to locate preferentially on specific positions on DNA (Simpson, 1991; Thoma, 1992). The term "positioning" refers to a predetermined statistical preference for a histone octamer to be located over a particular stretch of ~146 bp of a DNA sequence, rather than having a random distribution throughout the DNA molecule resultant from the nucleosome having equal probability of positioning at any particular location (Widom, 1998). It therefore follows that local DNA structure, which is determined by the DNA sequence itself, may have a primary role in positioning nucleosomes (Wolffe, 1998). "Rotational positioning" is a related phenomenon, which has been observed to occur when a degenerate set of translational positions, differing by integral multiples of the DNA helical repeat (~10.5 bp), are preferred (Widom, 1998). To differentiate it from rotational nucleosome positioning, the term "translational nucleosome positioning" has been adopted when distinguishing the two.

*In vitro* studies have shown that certain DNA sequences demonstrate significantly higher affinity for positioning nucleosomes, and therefore that there is likely to be something inherent to the sequence itself which facilitates this positioning. *In vivo*, however, other factors such as proteins bound to DNA may also influence positioning. Whether the sequence preference for the DNA is the dominant factor in determining the positions nucleosomes adopt in the cell is still a undecided issue (Blank and Becker, 1995).

### 1.2.2 Biological Influence of Positioned Nucleosomes

A substantial body of evidence has accumulated that at least a subset of nucleosomes *in vivo* are positioned, and that they play an important role in gene regulation (Simpson, 1991; Thoma, 1992).

Two kinds of DNA structural patterns have been proposed to direct nucleosome positioning: patterns which favour nucleosome formation and stability, and patterns that strongly inhibit positioning. Nucleosome positioning can help to either selectively expose functionally important DNA sequences by constraining their locations to the linker region or obstruct access to functionally important sequences by protecting them within the core particle. This degree of control over access to sequence motifs can be seen as another level of gene expression regulation.

The presence of nucleosomes on DNA has traditionally been thought of as having a repressive effect on gene expression. This is because the nucleosome restricts access to transcription factor sequence motifs. The curvature of DNA organised in a nucleosome is such that many transcription factors (and other molecular complexes) do not recognise binding sites, so the "rotational setting" of the DNA within a nucleosome can also affect whether factors will bind (Martinez-Campa *et al.*, 2004). As such, the presence of nucleosomes in regulatory regions can have a fundamental role in gene regulation, although the precise effect varies between genes. Although nucleosomes are not static structures and the DNA sequence throughout much of the core particle can become transiently dissociated from histones, DNA does not completely unwrap from the core particle (Li and Widom, 2004). Nucleosomes can also be mobile in appropriate conditions (Meersseman *et al.*, 1992), a process which may be important to allow transcription to proceed.

However, nucleosomes have also been shown to play a constructive role in transcription. The nucleosome can lead to the formation of "supergrooves" of DNA separated by 80bp when brought together by the nucleosome folding process (Edayathumangalam *et al.*, 2004). Similarly, the bending of DNA around histone octamer brings non-contiguous sequences together which can facilitate interactions between two bound transcription factors (Jackson and Benyajati, 1993).

It is therefore not accurate to say that a higher density of nucleosomes leads to transcriptional repression. Indeed, some widely expressed genes have a higher than average density of nucleosomes. It seems likely that there may be some aspects of the distribution of nucleosomes which facilitate rather than inhibit gene expression.

It has been demonstrated that the positioning of nucleosomes can be affected by minor alterations in the DNA sequence, and by DNA methylation (Davey *et al.*, 1997), most probably due to changes in the histone-DNA interaction.

Changes of the nucleosomal repeat length in genes have been linked to changes in gene expression, with the repeat length increasing during the production of red blood cells in chicken (Weintraub, 1978).

### 1.2.3 Translational Nucleosome Positioning

Translational positioning refers to positions nucleosomes prefer to adopt on long molecules of DNA. In this context, "long" refers to a length of DNA significantly longer than the core particle length (~146 bp). The principle of translational positioning is that certain 146 bp stretches of DNA sequence have a higher (or lower) ability to form and position a nucleosome.

A number of different structural features intrinsic to DNA are thought to influence the positioning process. Two features of DNA which have been thought to be particularly crucial for determining translational positioning are bendability (flexibility) and rigidity (Calladine and Drew, 1997). Highly flexible DNA requires a lower energy cost to wrap around a histone octamer than random DNA. Thus DNA, which is significantly more flexible than random DNA sequences, and could therefore be expected to preferentially position nucleosomes. On the other hand, highly rigid DNA, whose structural conformation is more restricted in comparison to random DNA, will be more difficult to bend around the histone octamer. Therefore, stretches of such DNA will be energetically unfavourable for nucleosome positioning. Flexible DNA is distinguished from intrinsically bent DNA in that bent DNA is a permanent feature of the DNA molecule, whereas bends in flexible DNA are transitory. A further consideration is that some DNA may have permanent, non-standard helical twists of a type that could facilitate nucleosome formation. A final consideration is that some DNA, by virtue of their sequences, may happen to form stronger and more frequent bonds with the histone octamer, increasing stability. These features are reviewed in Widom (2001).

It has been suggested that boundaries – which can potentially be created by various proteins, a strongly positioned nucleosome, or a DNA sequence which is particularly unfavourable for nucleosome positioning – could directly influence positioning as the local nucleosomes would have to organise themselves with respect to the excluded sequence.

### 1.2.4 Rotational Nucleosome Positioning

Rotational positioning determines which side of the double helix will face inwards towards the histone octamer, and which will consequently face outwards. As rotational positioning is tied to the DNA helical repeat, rotationally positioned nucleosomes are separated by ~10 bp. This type of positioning is therefore most suited for intrinsically curved DNA (Satchwell *et al.*, 1986), as appropriately bent DNA allows more readily for the formation of nucleosomes, as if the DNA is already suitably curved, it lowers the energy cost of wrapping the DNA around the histone octamer. This energy cost is, in general, substantial for a polyelectrolyte molecule such as DNA, which has a persistence length (a measure of the molecular flexibility) of around 50 nm (≈150 bp) (Hagerman, 1988). Bending DNA significantly shorter than the persistence length is energetically unfavourable, yet DNA is required to be bent almost one and three quarter times to form the nucleosome core particle.

### 1.2.5 Nucleosome Mobility

Although at physiological conditions, nucleosomes cannot be readily exchanged between two DNA molecules, they do exhibit the spontaneous ability to translocate on a DNA fragment in a temperature dependent manner without any interaction from an external source (Beard, 1978; Pennings *et al.*, 1991; Meersseman *et al.*, 1992).

Another, catalysed, form of nucleosome mobility occurs via SWI/SNF, an ATP dependent nucleosome remodelling complex, which can use energy from the hydrolysis of ATP to move nucleosomes on DNA (Becker and Horz, 2002).

There has been a recent theoretical discussion concerning  the possible method(s) of nucleosome translocation (Kulic and Schiessel, 2003b; Kulic and Schiessel, 2003a), reviewed and discussed in detail in Schiessel (2003).

### 1.2.6 *In Vivo* Sequence Dependent Nucleosome Positioning

There is some evidence that suggests that the DNA sequence is also a primary determinant of the positioning nucleosomes adopt *in vivo*, at least in specific, local instances (Buttinelli *et al.*, 1993; Adroer and Oliva, 1998). Unlike the *in vitro* situation, where only the DNA and core histone interact, there are many other factors which may affect where nucleosomes can position. The binding of certain proteins to the DNA molecule, mentioned earlier in respect of boundary-directed positioning, is one proposed factor absent *in vitro*. Indeed, a study of yeast minichromosomes determined that strong rotational positioning was not sufficient to direct nucleosome positioning *in vivo* (Tanaka *et al.*, 1992).

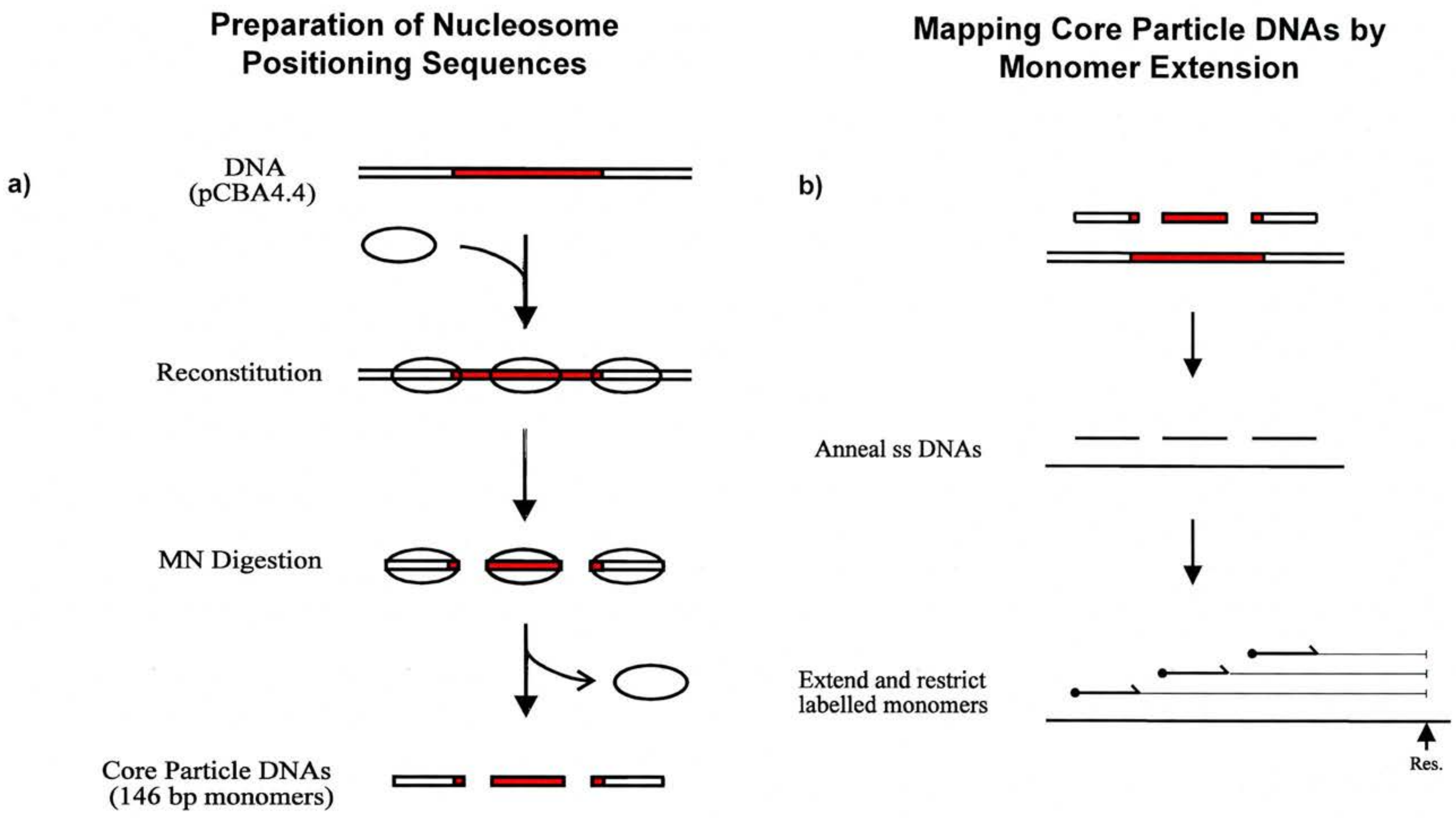
## 1.3 Monomer Extension

### 1.3.1 Introduction

Monomer extension (ME) (Yenidunya *et al.*, 1994; Gould, 1998) is an *in vitro* technique which maps the positions adopted by core histone octamers when they are reconstituted onto a defined DNA sequence. It provides quantitative positioning information, at high resolution, over long continuous stretches of DNA sequence. Throughout this thesis, extensive use will be made of data generated by the ME nucleosome mapping technique. This technique has been used in a number of prior nucleosome positioning studies (Yenidunya *et al.*, 1994; Davey *et al.*, 1995; Davey *et al.*, 1997; Gould, 1998; Shen *et al.*, 2001; Davey and Allan, 2003; Davey *et al.*, 2003; Gencheva *et al.*, 2006).

### 1.3.2 Basic Procedure

A schematic overview of the basic steps of the ME procedure is provided in Figure 1.4. The ability of ME to map nucleosome positioning is a result of the protection of the DNA within the nucleosome core particle to digestion by micrococcal nuclease (MN). When core histones are reconstituted onto a specially constructed plasmid containing the sequence to be mapped, the resulting fibre is subject to MN digestion. This strips away the linker DNA leaving "monomer", core particle DNA (Figure 1.4 a)). The isolated monomer DNAs are then annealed back onto a single stranded version of the original plasmid, and extended to a known restriction site. Extension products are resolved on a denaturing polyacrylamide gel and, from the size of the extension products, the nucleosome boundaries can be mapped to base pair accuracy. In addition, the number of molecules within each band on the polyacrylamide gel gives a relative measure of the number of core particles that were positioned on each possible positioning site. Quantification is achieved by scanning the gel with a PhosphorImager.



Preparation of Nucleosome Positioning Sequences

Mapping Core Particle DNAs by Monomer Extension

Figure 1.4: Schematic outline of Monomer Extension Procedure.

There are key points on which the analyses presented within this thesis are reliant and should therefore be highlighted. Firstly, the conditions at which the reconstituted histones reform onto the DNA sequence are such that the core particles will still be relatively free to translocate to other possible positioning sites within their local region. Secondly, the reconstitution reaction takes place with a limited amount of core histone, such that the nucleosome density will be much lower than seen *in vivo* (approximately 1 octamer to 600-800 bp). This minimises the possibility of nucleosome- nucleosome interactions. It does not, however, eliminate the possibility of such interactions. The DNA can loop back to allow spatially separated nucleosomes the opportunity to interact. Thirdly, there are a very large number of molecules interacting within the reconstitution experiment, a number sufficient that it is possible to view the number of histone octamers which choose to position on a particular 146 bp stretch of DNA as reflecting the probability of finding an octamer at that site in any single molecule.

### 1.3.3 Critical Evaluation of Monomer Extension

Despite its strengths and advantages over competing *in vitro* translational positioning mapping techniques, ME does suffer from a number of problems. These can broadly be broken down into two areas: errors associated in assigning the nucleosome positioning signal to the correct 146 bp of underlying sequence, and errors in determining the strength of the positioning signal itself.

The more significant of these is the uncertainty in determining the underlying sequence responsible for the quantified positioning signal. The error associated with the quantification itself, for the type of analyses generally for which ME data has been used (in this and in other studies), is not of particular concern.

The mapping technique requires that one know the length of the monomer DNAs protected from MN digestion. Whilst digestion by MN does usually leave monomer DNA fragments of ~146 bp in length, there are a number of instances, such as irregular protection of the sequence by the core particle and where MN displays a sequence specificity towards AT rich regions, which cause the lengths of the monomer core particle DNA to fluctuate. Such variations are usually limited to ±3 bp (Yenidunya et al., 1994).

Another concern is the repeatability of the positioning affinities determined by the mapping procedure. However, it would appear that ME does give a quantitatively repeatable assessment of the positioning signal. An example of this can be found in Figure 5 (a) in Davey *et al*., (2003) where methylated and non-methylated DNA was assessed by ME. Outwith the regions where the methylation has an effect, the positioning affinities assessed by ME are quantitatively similar.

The width of band on a denaturing polyacrylamide gel is relatively invariant to migration length; the band width is usually a few pixels in length top and bottom of the gel. However, the calibration between migration length and fragment size is non-linear, as longer DNA molecules will tend to migrate at a considerably slower rate than DNA molecules half their size. A result of this is that one pixel at the top of a gel represents ~1 bp in DNA length, whereas 1 bp normally covers several pixels near the bottom of the gel. When compiling the datasets, one has to assign a positioning affinity for each site so this could potentially lead to a strong positioning signal being inappropriately split into a number of otherwise relatively weak positioning sites, diluting the relative affinity of strong site.

There are also errors associated with the quantification of the amount of DNA material within a band by PhosphorImager. However, such inaccuracies are unlikely to introduce a significant error.

## 1.4 Nucleosome Positioning Datasets

### 1.4.1 Monomer Extension BLG and Globin datasets

Two ME datasets were used in this thesis. The longer of the two, the map of the ovine β-lactoglobulin[1] (BLG) gene region (Gencheva *et al.*, 2006) is plotted in Figures 1.5 a). Figure 1.5 b) is a map which charts the positioning adopted by nucleosomes for the same sequence *in vivo*. This map is discussed in more detail in the next section. Figure 1.5 c) is a schematic representation of the BLG gene structure, with the exons and promoter represented as black rectangles. The other analysed dataset is a map of the region containing the chicken adult ($\beta^A$-) and embryonic (ε-) globin genes[2] (Davey *et al.*, 1995; Edgar, 1999) (Figure 1.6 a)), hereinafter referred to as "globin". Figure 1.6 b) is a schematic representation of structure of the two genes within the globin map, with the exons represented as black rectangles and the enhancer as the red rectangle. For completeness and for comparison, the three other published ME maps of the Human H19, Mouse H19 and Igf2r gene regions are shown in Figure 1.7 a), b) and c) respectively.

### 1.4.2 BLG *in vivo* map

Figure 1.5 b) is an *in vivo* nucleosome positioning map, produced by indirect end-labelling (Wu, 1980), of the ovine β-lactoglobulin gene in liver nuclei (Boa, 1999; Gencheva *et al.*, 2006). The dark blue ovals depict identified positioning sites, the grey ovals were positioning where only one nucleosome boundary was cleaved and white ovals are nucleosome positions inferred from the patterns of the identified positions (to fill in gaps in the *in vivo* map). The light blue ovals depict regions where the otherwise regular nucleosomal array adopts two different positions within the same 146 bp. These alternative arrays are out of phase by approximately 60 bp and consequently give rise to overlapping positions. The red oval represents a region which had an unusual protection to cleavage by copper phenanthroline. Away from the regions of alternative positioning, the map demonstrates that the positions adopted by nucleosomes *in vivo* are regularly spaced, suggesting that a uniform higher-order fibre may form at this location.

[1] A tissue-specific milk gene not found in humans or rodents

[2] Another tissue-specific gene, involved in the production of haemoglobin in vertebrates.

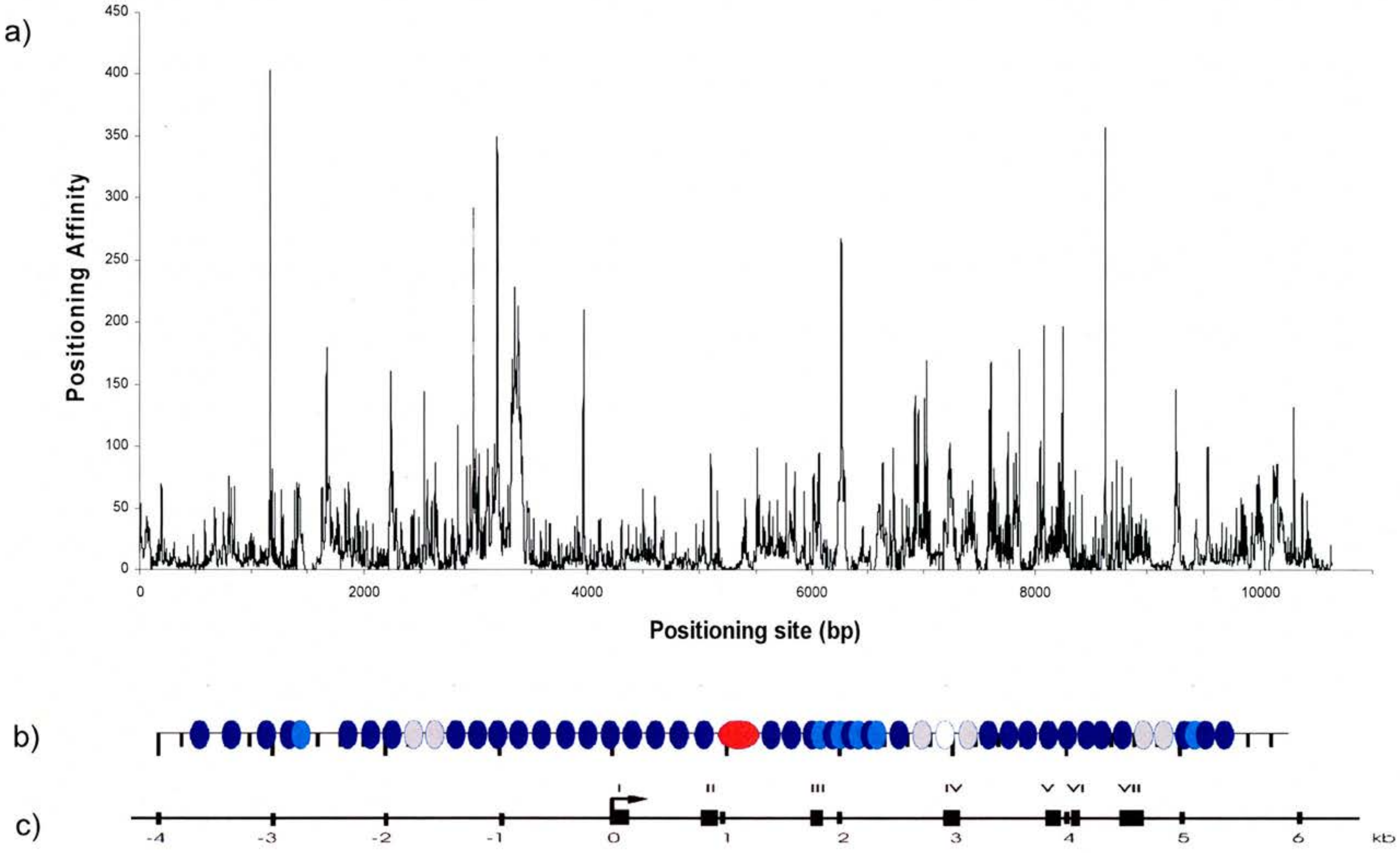


**Figure 1.5: BLG nucleosome positioning dataset. a) Monomer extension map for the nucleosome positioning affinity within the BLG gene region. b) *In vivo* nucleosome positioning map produced by indirect end-labelling (see text). c) Schematic representation of the BLG gene structure. Black rectangles represent the 7 exons.**

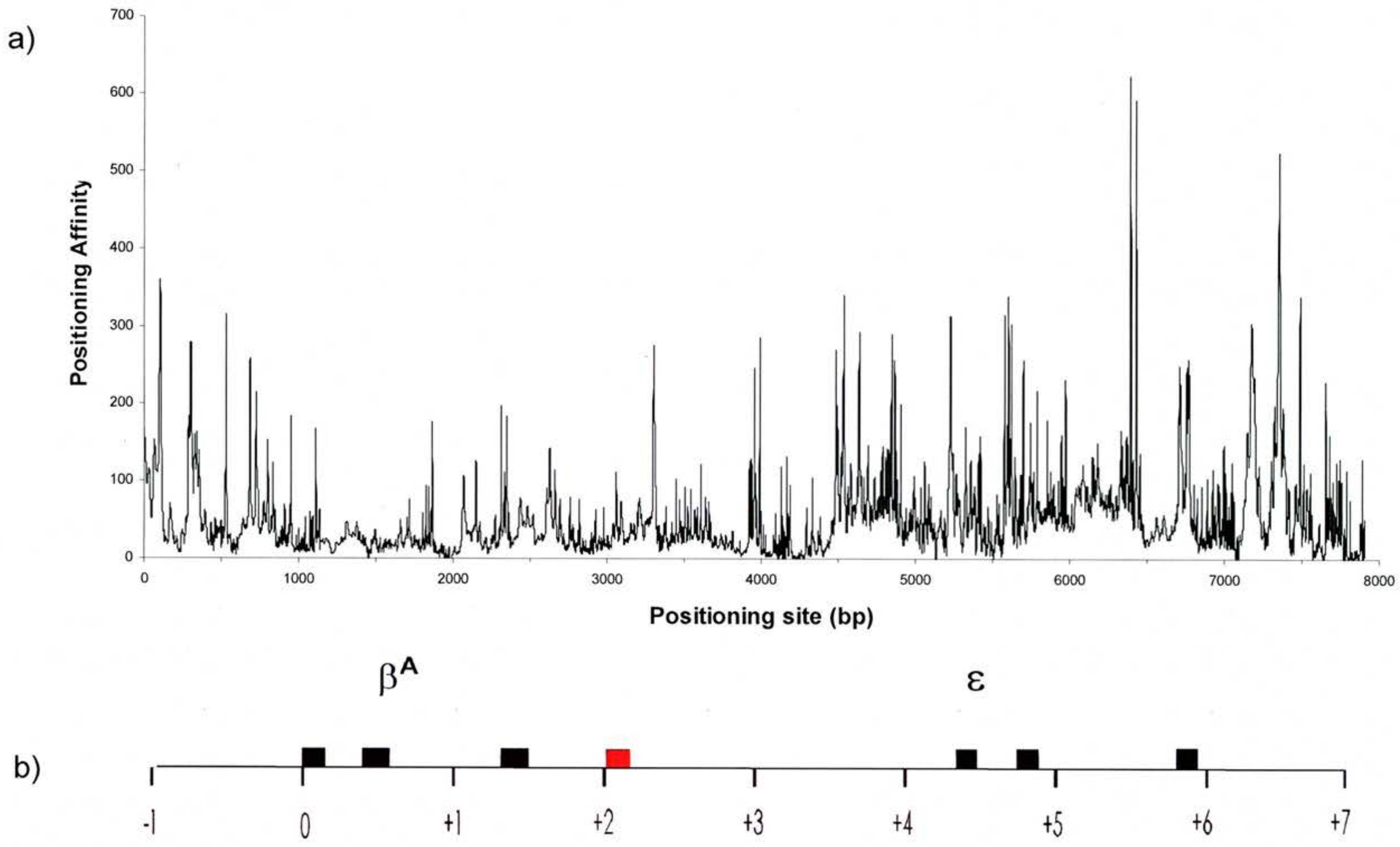


**Figure 1.6: Globin nucleosome positioning dataset. a) Monomer extension map for the nucleosome positioning affinity within the globin gene region. b) Schematic representation of the gene structure for the region mapped, indicating locations of the β$^{A}$- and ε-globin genes, their exons (black) and enhancer (red).**

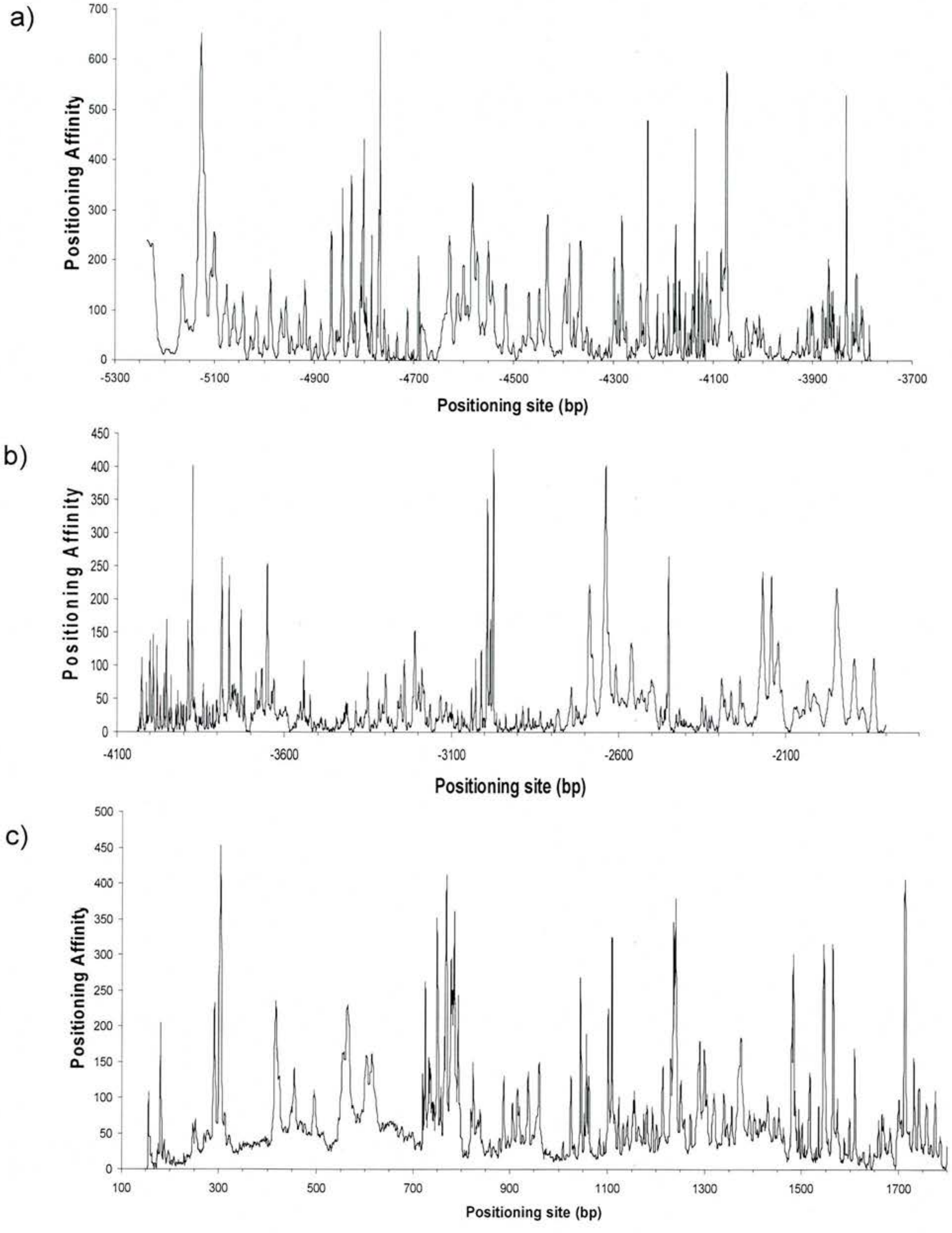


**Figure 1.7: Monomer extension maps for a) Human H19, b) Mouse H19 and c) mouse Igf2r gene regions. Positioning sites plotted relative to transcription start site.**

## 1.5 Prior Analyses of Nucleosome Positioning

In this section, a selection of prior approaches for the analysis of nucleosome positioning will be briefly introduced. The first, BEND, is a program which predicts the curvature and flexibility of DNA *in silico,* and has been used to predict nucleosome positioning (Blomquist *et al.*, 1999; Wada-Kiyama *et al.*, 1999b; Bash *et al.*, 2001; Fiorini *et al.*, 2001). The program can use data about any of the structural parameters of DNA which contribute toward bending the DNA molecule.

It has been demonstrated, by circularising DNA, that AAA and TTT trinucleotides tend to be positioned with the minor groove facing inwards, and that GGG, CCC and GGC were positioned so that their minor groove was on the outside of the DNA (Drew and Travers, 1985). By sequencing 177 identified nucleosome positioning sites, it was observed that AAA/TTT and AAT/ATT trinucleotides were positioned with a ~10.2 bp periodicity, and arranged in such a fashion that their minor groove faced the nucleosome core. Further studies demonstrated GC dinucleotides were also found to have a 10.2 bp periodicity, but which was out of phase with the AT; GC base steps were located so that the minor groove tended to face outwards (Satchwell *et al.*, 1986). These, and other such observations, led to the development of a likelihood matrix for predicting nucleosome positioning (Satchwell *et al.*, 1986; Drew and Calladine, 1987).

Multiple sequence alignments (Ioshikhes *et al.*, 1992; Ioshikhes *et al.*, 1996; Wang and Widom, 2005) have been used to investigate databases of DNA sequences which have been experimentally determined to position nucleosomes. A prominent result from these studies was the apparent 10.2 bp periodic arrangement of AA/TT dinucleotides in the nucleosomal DNA.

Baldi *et al.* (1996), using hidden Markov models, found a non-T, A/T, G (VWG) motif, which is very often found in vertebrate genomes, is found to have a 10 bp periodicity in regions were observed to ordered nucleosomes *in vitro*.

Finally, Levitsky *et al.* (2001), using public accessible nucleosome positioning databases, developed the RECON program for determining the "nucleosome formation potential" of a given sequence of DNA. This program will be explored in more detail in section 2.5.3.

## 1.6 Thesis Aims and Overview

### 1.6.1 Aims

It has been well established in the literature that sequence dependent nucleosome positioning is a feature *in vitro* and, in specific instances on a local level, *in vivo*. There is not yet the similar body of evidence supporting the proposition that the sequence encodes longer range distributions of nucleosomes *in vivo*.

The importance of the ME dataset used within this thesis derives from three factors: its size, currently 24,000+ nucleosome positioning sequences, which is considerably larger than any other published dataset; that ME quantitatively assesses the positioning affinity for mapped DNA sequences; and finally the comparatively high resolution of the determination of the sequence responsible for a given positioning signal (down to one bp). There is no comparable dataset of this size or quality in the literature.

The central aim of this study was therefore to analyse this unique nucleosome positioning dataset, both by developing prior analyses and by designing novel novel computational approaches, in the hope of gaining new insights into the factors that influence nucleosome positioning.

The second main aim was to use this dataset to assess the influence of *in vitro* nucleosome positioning signals, which involve only DNA-histone interactions, on the arrangement of nucleosomes *in vivo*, which involve DNA-histone interaction plus other factors such as proteins. Central to this was the existence of a long range map of the positions adopted *in vivo* for the BLG gene region.

### 1.6.2 Thesis Overview

In chapter 2, a number of different techniques will be used to characterise the nucleosome positioning available datasets introduced in section 1.4. Chapter 3 details the development of a novel simulation technique based on Metropolis Monte Carlo methods. This new approach is then used to study the datasets in Chapter 4.

The final chapter brings together these threads and reviews the key results obtained.

CHAPTER 2

# Characterising Nucleosome Positioning Datasets

## 2.1 Chapter Overview

In the following chapter, four separate analyses are presented: the first three analyses provide some insight into the nature of the ME datasets, whilst also serving as an introduction to some of the core analysis methods used in subsequent chapters. The fourth analysis (2.5) is a study of two commonly used algorithms for the prediction of sequence-dependent nucleosome positioning. The overriding aim of this section is to critically assess the current capacity for the prediction of nucleosome affinity given an arbitrary DNA sequence by comparing the predicted affinities with the experimentally determined quantitative nucleosome positioning data from ME. This final section lays the foundation for the motivation behind, and development of, the Monte Carlo methods-based analysis proposed in the next chapter.

## 2.2 Periodicity Analysis of the BLG and Globin ME datasets

### 2.2.1 Power Spectrum Analysis

The many analyses throughout this thesis make extensive use of signal processing techniques to analyse periodicities found in various datasets. Of particular interest is the power spectral density (PSD) or power spectrum. This is a useful technique for detecting and classifying periodic signals in complex datasets as the PSD gives a quantitative measure of the strength of a particular period within a dataset.

The PSD is the Fourier Transform of the autocorrelation of the dataset in question, as the PSD and the autocorrelation are Fourier Transform pairs, where autocorrelation is the cross-correlation of dataset with itself (Bracewell, 1986). Spectral analysis methods of this type have been used extensively in prior studies of nucleosome positioning, such as Davey *et al.* (1995) and Widom, (1996), and the implementation used hereinafter is much as described in Davey *et al.* (1995).

### 2.2.2 PSD of the entire BLG and Globin gene regions

In Figures 2.1 a) and b) the PSDs for the mapped BLG and globin gene regions are presented. The strongest single periodicity evident in the BLG dataset is at 193 bp, with smaller peaks at approximately half the power surrounding the main peak. There is also a complex array of less significant peaks in the range from 100 – 400 bp. Similar behaviour is evident in the PSD for the complete globin dataset, but with the strongest period now at 210 bp. However, it is clear that there are other strong periodic signals of nearly equal strength in the region between 175- 210 bp, as well as above 250 bp.

It is worth highlighting at this early stage that there is no clear evidence of a 10 bp periodicity in either dataset; a 10 bp periodicity in the nucleosome positioning signals would be regarded as evidence of rotational settings influencing translation positioning (rotational positioning).

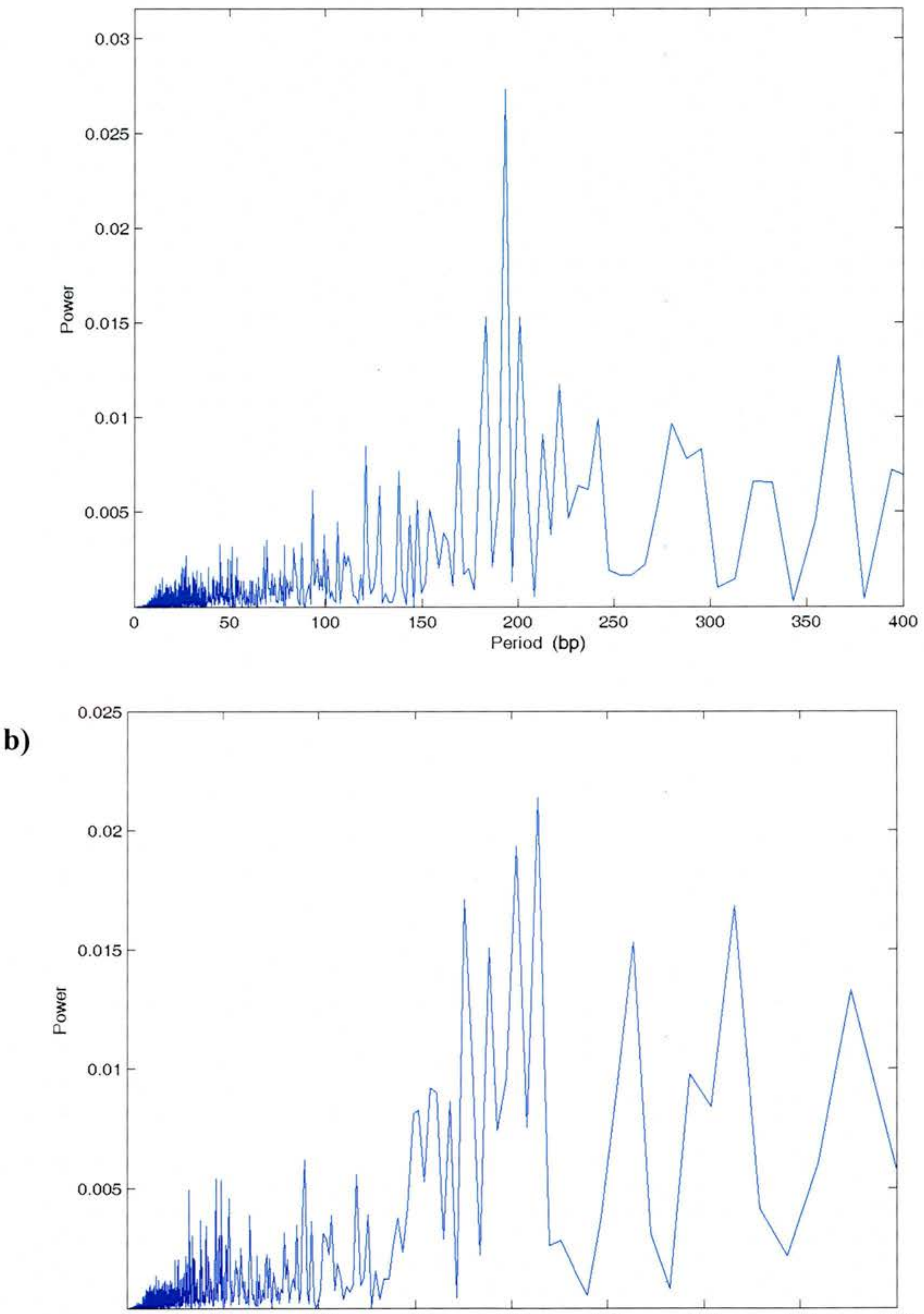


**Figure 2.1: PSD analysis of the entire ME datasets for (a) the BLG and (b) globin gene regions.**

### 2.2.3 Scanning PSD analysis

The intricacy of the PSD profiles in Figures 2.1 a) and b) reflects a complex collection of periodic signals contained within different regions of the datasets, the complexity of which is discernable from close study of different sections of the datasets themselves. To characterise and better understand the interaction between these different periodic signals, it is useful to break the datasets into smaller stretches, and recalculate the PSD of the shorter region.

To achieve this, a "scanning PSD" analysis technique was developed. To generate a scanning PSD, a window of sufficient size is used such that periodic signals up to 400 bp can be reliably discerned. Therefore, the choice of the window size reflects a balance between ensuring that it is of sufficient size to identify periodicities in the desired range (up to 400 bp) whilst still allowing the output to be related to local features of the data. If the window is too small, the ability of the Fourier Transform procedure to identify periodicities within the required range is impaired. A window size of 2000 bp was found empirically to be a good compromise and is used throughout the thesis. This window is slid across the dataset being analysed in steps of 100 bp, with the PSD of the data contained within each window calculated. PSDs from consecutive windows are then plotted next to one another in a pseudo-3D surface plot, with the window start position along the x-axis, period on the y-axis, and the PSD on the z-axis. The PSD scale is represented by a colourbar on the right hand side of each plot; the strength of periodicities range from blue to red (lowest to highest). One should note that a window start position of 2000 bp on the x-axis plots the PSD within a window from 2000-4000 bp.

Scanning PSD analyses of the BLG and globin datasets are presented in Figures 2.2 a) and b) respectively. In the former, the strongest periodicities are generally to be found within 190-210 bp, but this is variable across the dataset, with significant periodicities between 150 to 400 bp also identified. One can clearly see the contribution to the peak at 193 bp and the surrounding peaks in the region between 5500 and 9500 bp, the region which contains the coding sequences. Other notable periodic signals of ~290 bp in length are detected within windows between 500 and 3000 bp of the BLG dataset.

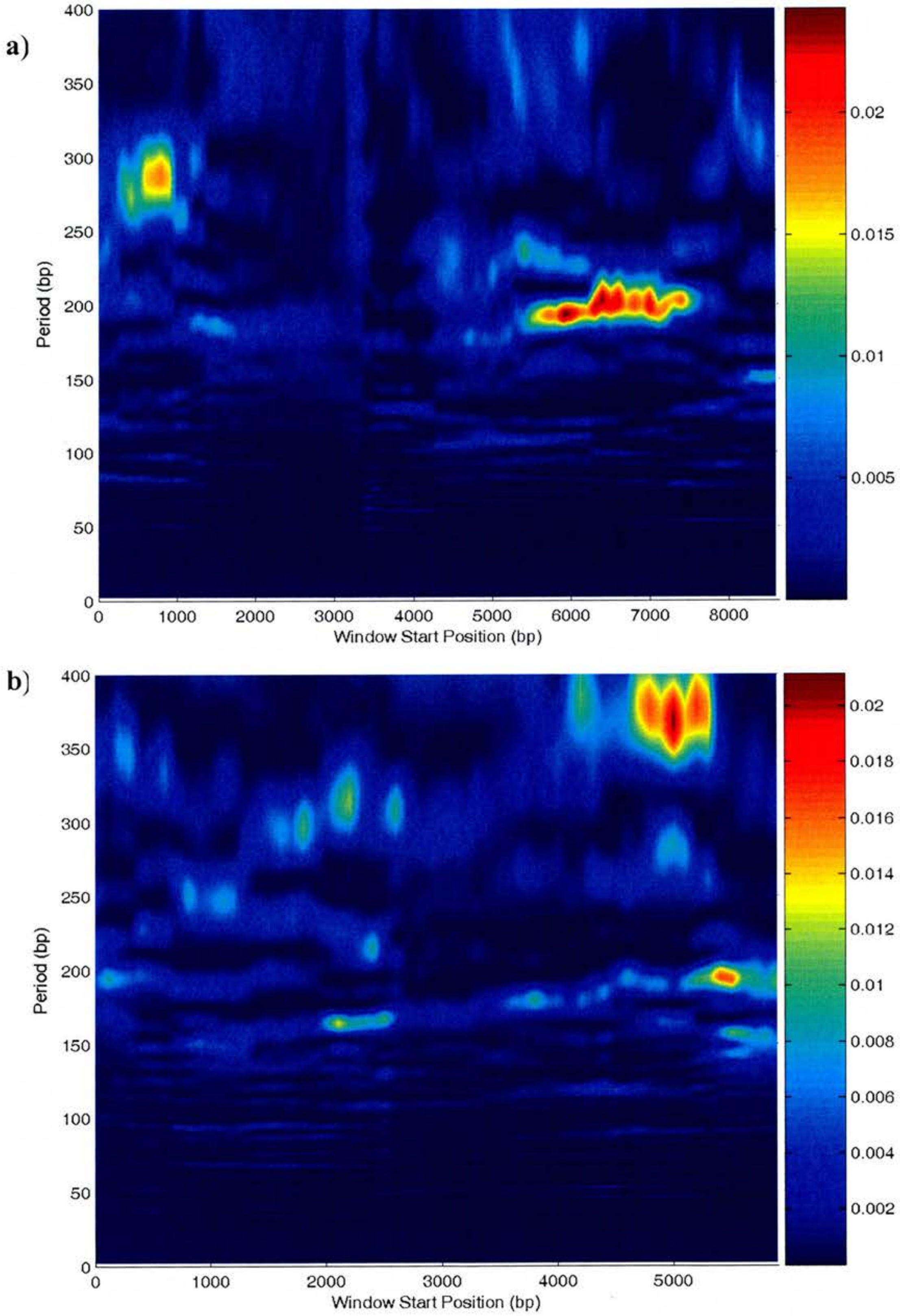


**Figure 2.2: Scanning Fourier analysis of ME assessed *in vitro* nucleosome positioning signals in a) the BLG and b) the globin gene regions.**

Turning to the globin dataset, one can discern more complex variances in different regions of the dataset, which are undoubtedly the cause of the more complex set of periodicities observed around 200 bp in Figure 2.1 b). Interestingly, it is common to find at least 3 strong periodicities in most regions of the globin dataset. For instance, in the region between 2000 and 4500 bp, there are strong periodicities evident at 160, 210 and 300 bp, whereas between 4500 and 7000 bp, periods of around 155, 185, 290 and 385 bp are prominent. A similar pattern of behaviour is not generally found in the BLG dataset which commonly only has two notable periodicities, excepting the window start position of between 5000 to 6000 bp.

### 2.2.4 Summary

The power spectral analyses presented reveal that the two ME datasets display a similar characteristic periodicity, with strong periodic signals in the region of 200 bp. This falls within the range of nucleosome repeat lengths commonly found *in vivo*[3], and it has been argued that this is an indication of the potential influence of *in vitro* nucleosome positioning on higher order chromatin structures (Davey *et al.*, 1995). It is also worth noting that there are considerable fluctuations in the periodic arrangement of nucleosome positioning signals within different regions of the dataset. Also noteworthy is that, although the ~200 bp periodicities revealed are relatively strong, it would not seem to reflect the regular arrays of nucleosomes commonly observed *in vivo* (Van Holde, 1989), especially given the strength of periodic signals outwith the 185-210 bp region in both datasets.

Also intriguing is the seeming absence once again of ~10 bp periodic signals. Given the current consensus on the role of rotational positioning, one would have expected some evidence, especially in regions which contain strong positioning sites. However, from the current analysis, there is little evidence to support rotational settings making a substantial contribution in the BLG and globin ME datasets.

[3] Repeat lengths of ~187 bp are observed *in vivo* in BLG. Globin, on the other hand, displays two distinct repeat lengths: 210 bp when the adult β-globin is expressed and ~187 bp when the embryonic gene is expressed.

## 2.3 Free energy ME comparison with competitive reconstitution

### 2.3.1 Motivation

Two commonly used methods for quantifying nucleosome positioning affinity for DNA are ME and competitive reconstitution (Ellington and Szostak, 1990; Tuerk and Gold, 1990; Widlund *et al.*, 1997). Presented here is an attempt at a semi-quantitative comparison between the relative free energies observed by each technique.

### 2.3.2 Comparison design

Widlund *et al.* (1997) have previously calculated the free energy of nucleosome binding for their competitive reconstitution experiments, based on the work of Shrader and Crothers (1989), and have used the following equation for determining relative free energy:

$$\Delta G^{\circ} = -RT\ln(f_i / f_{ref}) \qquad \textbf{(2.1)}$$

where $\Delta G^{o}$ is the relative difference in free energy, R is the molar gas constant, and T is the temperature in Kelvin. $f_i$ and $f_{ref}$ are respectively the ratios between the band intensities from nucleosomal and free DNA for the particular sequence being evaluated ($f_i$ ) and reference sequence ($f_{ref}$). This ratio gives a quantitative assessment of the relative ability of the sequence to capture an octamer, and therefore is a measurement of the affinity for the histone octamer of the particular sequence being evaluated.

In competitive reconstitution, free energies are therefore calculated relative to a reference DNA sequence, commonly the well characterised 5S RNA gene, which allows the comparison of results from different experiments. However, there is no corresponding reference sequence in ME experiments, as the positioning signals are mapped in the same experiment relative to the other sites being mapped. In addition, there is no direct analogue to $f_i$, the ratio of band intensity between DNA which has bound an octamer and DNA that has remained unbound, which constitutes a problem when attempting to relate the two approaches. In ME, positioning affinities are assessed in the context of the free, unbound DNA of the remainder of the DNA molecule (which is considerably larger than 146 bp), and so it is therefore possible to assume that the intensity of ME bands incorporate this factor. In other words, the intensity of a band for a particular fragment, $I_i$, is equivalent to $f_i$.

If one accepts these arguments, it is possible to define a ratio of relative free energies, relative to an arbitrary positioning site, analogous to Equation (2.1), for ME datasets:

$$\Delta G_i^{\circ} = -RT\ln(I_i / I_{ref}) \qquad \textbf{(2.2)}$$

Where $I_i$ denotes the ME band intensity of positioning site $i$, and $I_{ref}$ represents the band intensity of the arbitrarily chosen reference positioning site. Equation (2.2) therefore represents a measure of relative free energy, relative to the free energy of the reference positioning site.

To make use of Equation (2.2), a subset of the BLG ME data was used: the positioning affinities of 1771 clearly discernable positioning sites were manually quantified by visual inspection (this dataset was prepared by Dr Marieta Gencheva (Gencheva *et al.*, 2006)). This represents only a small subset (~17%) of the 10,640 mapped positioning sites. The intensity of the site with the lowest appreciable affinity was taken to be the reference sequence, $I_{ref}$, with the relative free energies of the 1771 positioning sites calculated using Equation (2.2). The choice of the reference site is arbitrary, but it will only affect scaling of the x-axis. In this sense, it is similar to the choice of the positioning affinity of the 5S RNA sequence as a reference in many competitive reconstitution experiments (Shrader and Crothers, 1989; Widlund *et al.*, 1997; Thastrom *et al.*, 2004b). The experimental temperature was taken to be 293K.

It is possible, by making a few assumptions, to attempt to directly compare the free energy values to those found via the competitive reconstitution assays. Such studies have reported that the free energy value, relative to the 5S gene, for bulk nucleosomal DNA, is approximately 1.3 kcal/mol (Shrader and Crothers, 1989;

Widlund *et al.*, 1997). If one assumes that the effective intensity for these positioning sites which were too low in affinity to have a discernable positioning affinity (83% of the BLG positioning sites), are uniformly distributed between the value for the lowest discernable site and zero, one can calculate a mean free energy for all the BLG positioning sites. It is reasonable to assume that this value should be similar to the free energy of bulk DNA. Hence it is possible to scale the relative free energy with respect to these common values, and so allow semi-quantitative comparison between the two techniques.

### 2.3.3 Results and Discussion

Figure 2.3 demonstrates that the free energies are distributed with an overall range of ~5 kcal/mol. Disregarding outlier data points, the majority of the sites (95%) lie between -1 and -4 kcal/mol. The distribution is approximately Gaussian in overall shape, although it is possible that the distribution is comprised of at least 2 distinct distributions, one centred at around -2.5 kcal/mol, one at around -3.1 kcal/mol.

The top axis denotes the free energies appropriately normalised in the manner suggested in section 2.3.2, and was calculated as follows: the mean of the identified positioning sites plotted in Figure 2.3 a) is -2.80 kcal/mol. After factoring in the remaining positioning sites not included in the 1771 quantified manually, leads to an estimate of the mean free energy for all positioning sites in the mapped BLG gene region of -0.34 kcal/mol.

The overall range of relative free energies in Figure 2.3 a) is similar to that reported in a number of competitive reconstitution studies (Shrader and Crothers, 1989; Lowary and Widom, 1997; Widlund *et al.*, 1997; Lowary and Widom, 1998; Thastrom *et al.*, 1999; Thastrom *et al.*, 2004a). A summary is graphically represented in Figure 2.3 b). Further, in Figure 2.3 b) the strongest positioning sites lie around -3 kcal/mol and the least energetically favourable sites around 1 kcal/mol, which appears to be in reasonable agreement with the strongest and weakest positioning sites in the BLG analysis (upper scale in Figure 2.3 a)).

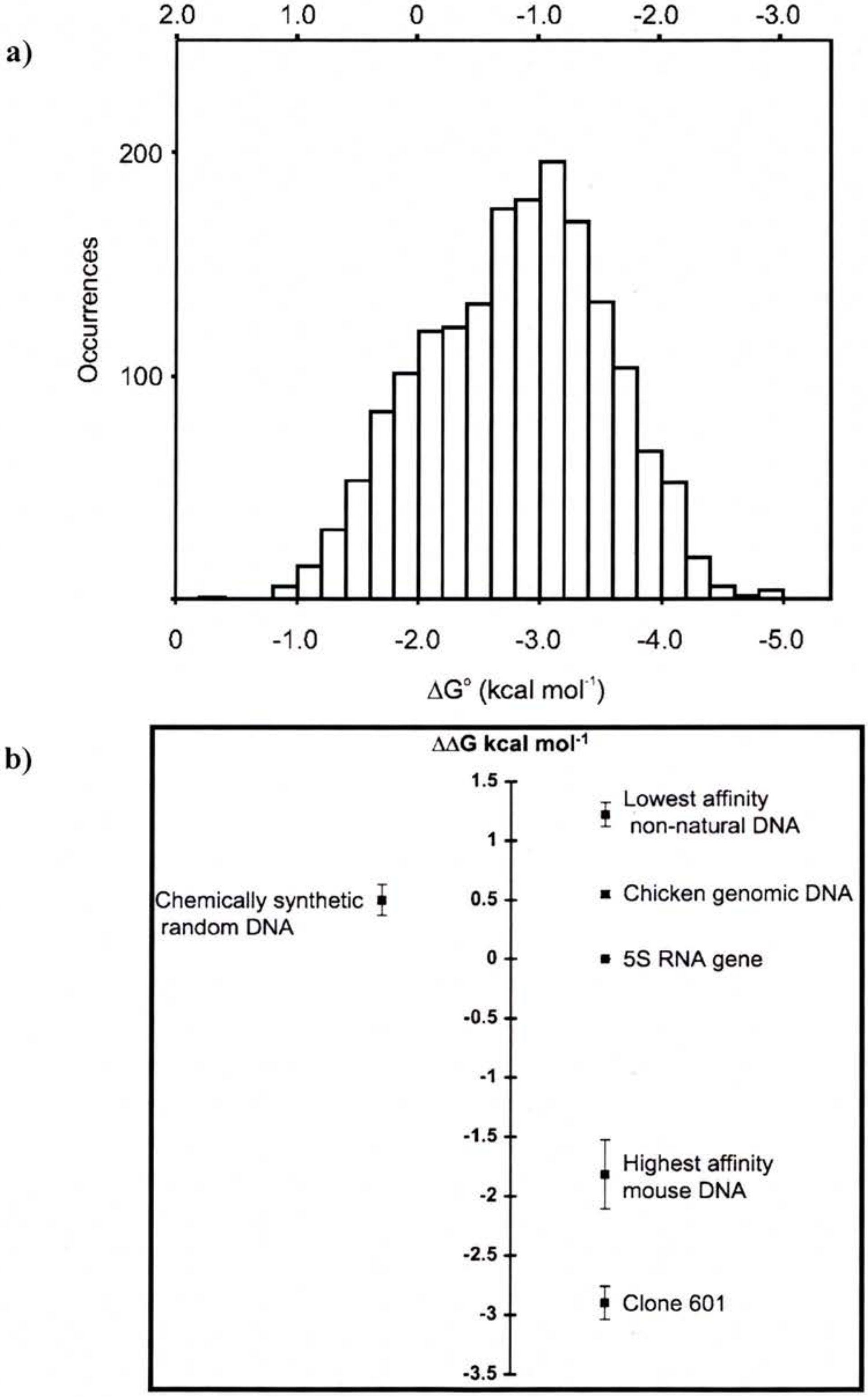


**Figure 2.3: a) Histogram of relative free energies (ΔG°) of 1771 BLG *in vitro* positioning sites, relative to the weakest identified site (lower scale) or to the 5S ribosomal gene sequence (upper scale). b) Spread of free energies from selected natural and artificial DNA sequences. b) is reproduced from (Thastrom *et al.*, 2004b).**

It has previously been suggested that the binding affinity determined by competitive reconstitution experiments are likely to be dominated by the ability of the given sequence to "capture" a histone octamer onto the short sequence of DNA rather than the ability to position (Thastrom *et al.*, 2004a). In addition, there is some uncertainty as to what component(s) of DNA-histone interactions the equilibrium measured by competitive reconstitution actually reflects (Drew, 1991). Recent experimental developments in this area suggest a possible solution to this potential issue using a modified dialysis-based approach (Thastrom *et al.*, 2004b). As the "initial capture" (or binding) of the histone octamer by the DNA occurs at relatively high salt concentrations, the properties of the DNA which influence this binding process may differ from those that determine nucleosome positioning at lower, more physiological, salt concentrations. This raises the prospect that relative affinities (free energies) determined by such processes are assessing the affinity for octamer binding (initial capture) rather than octamer positioning *per se*.

It has also been suggested that the first part of the histone octamer to assemble, the stable H3/H4 tetramer, may be the dominant factor in determining the initial capture of the octamer (Thastrom *et al.*, 2004a), and that therefore the binding affinity may be more reflective of tetramer rather than octamer binding.

On the other hand, given the assumption that the conditions in which ME experiments are conducted are such that the histone octamer is relatively free to translocate itself on the DNA molecule after initial capture (Beard, 1978; Davey *et al.*, 1995) – which is in general significantly longer than the short (<250 bp) molecules used in competitive reconstitution – it is not unreasonable to expect that ME data more accurately reflects the affinity for final octamer positioning.

It follows that the similar range in Figure 2.3 a) and b) offers support to the proposal that factors which influence initial octamer capture by the DNA sequence are related to those that influence the eventual positioning of the nucleosome on the DNA molecule[4]. There is some corroborating evidence from competitive reconstitution experiments that binding affinity appears to be linked to positioning strength for at least some sequences (Lowary and Widom, 1998; Thastrom *et al.*, 1999), although it is clear that the relationship between nucleosome binding and positioning (if indeed there is any) is not currently well understood.

Nonetheless, it is somewhat surprising that the range of relative free energies from competitive reconstitution – which commonly make use of nucleosome SELEX[5] enrichment to select DNAs, including artificial sequences, for their affinity (or lack thereof) for the DNA molecule – would be similar to those in a natural gene region. It would not have been unreasonable to expect, *a priori*, that the range of free energies from the *in vitro* selection pool, including DNAs from chicken erythrocytes, mouse tissue culture cells, and synthetic sequences, would be wider than that found in only 10.7 kbp of the BLG gene region. It is possible that this reflects that the overall range of nucleosome positioning affinities are themselves limited in scope, and/or that even specially selected sequences, including sequences of synthetic origin (which need not be subject to physiological constraints), do not have significantly higher (or lower) affinity for the histone octamer than those found *in vivo*.

Another possible explanation is provided by recent work by Wu and Travers (2005). Here, the authors demonstrate that ability of a particular DNA molecule to capture a histone octamer is dependent on the conditions under which the histone reconstitution occurs. One key result is that the 601 sequence, the highest binding affinity sequence identified prior to this study, has a significantly lower affinity than natural mouse DNA sequences at certain temperatures and histone concentrations. As the conditions under which the ME technique is run are different to those of nucleosome SELEX, this will affect the range of affinities found from both techniques.

[4] Assuming one accepts that free energies determined by competitive reconstitution are likely to be dominated by the capture of the histone octamer whereas ME represents the positioning affinity.
[5] SELEX (Systematic evolution of ligands by exponential enrichment), or *in vitro* selection, is a technique which allows the simultaneous screening of varied pools of different DNA molecules for a particular feature. In the case of nucleosomal SELEX, this feature is the affinity of the sequences in question for the histone octamer.

### 2.3.4 Summary

Although the magnitude and range of free energies determined by the two *in vitro* approaches appear similar, it is difficult to come to any firm conclusions as to whether nucleosome binding and positioning rely on the same parameters of the interaction between the histone octamer and DNA, especially given the number of assumptions on which this analysis is based. There nonetheless remains an intriguing correspondence between the datasets from these two experimental techniques. Competitive reconstitution experiments using DNA fragments from the BLG gene and/or ME assessment of sequences already characterised by competitive reconstitution, would be a substantive step towards resolving these questions.

## 2.4 Comparison of BLG *in vivo* and *in vitro* positioning maps

### 2.4.1 Impact of sequence-dependent nucleosome positioning *in vivo*

The degree to which the factors that affect nucleosome positioning *in vitro* influence long range positioning *in vivo* still remains an open question (Blank and Becker, 1995; 1996; Becker, 2002). However, the existence of an *in vivo* and *in vitro* nucleosome positioning map for the BLG gene region (Gencheva *et al.*, 2006) provides a unique opportunity for comparisons between the two maps, which may shed some light onto this area.

Therefore the following section seeks to explore the potential relationship between the *in vitro* positioning signals and nucleosome positions adopted *in vivo*. To accomplish this, the positioning signals which lie within the approximate location of the 50 identified nucleosome positions *in vivo* were analysed (Gencheva *et al.*, 2006).

### 2.4.2 Design

The *in vivo* mapping technique used in the production of this map, indirect end-labelling (Wu, 1980), has a relatively large associated error, resulting in an uncertainty of approximately ±20-30 bp in the determination of the positions of the nucleosomes. This error is significantly larger than the error associated with the ME technique, and therefore the error associated with ME can be considered relatively insignificant.

To compensate for the uncertainty in the determination of the *in vivo* positions, a rectangular filter was designed to process the intensity of *in vitro* positioning signals that fall within specific distances from the observed *in vivo* nucleosome dyads. This filter summed the positioning signal within windows of various lengths.

The summed intensity within a window is defined as the Observed signal ($\Omega$). To normalise this quantity, it is useful to define a second quantity, the "Expected" signal (E) one would anticipate within the window if the *in vitro* data was completely random. E is calculated by multiplying the window size by the arithmetic mean of all the positioning signals found in BLG. An Ω/E value of 1 indicates that the mean nucleosome positioning signal within the window was identical to what one would expect on average. Values greater than 1 indicate that positioning signals within the window were greater than the amount one would expect. Similarly, a value below one indicates that there was less positioning signal in that window than expected at random. Therefore the mean Ω/E over the windows serves as a normalised estimate of the correspondence between the two nucleosome positioning maps.

### 2.4.3 Relationship between *in vivo* and *in vitro* nucleosome positioning maps

The range of window sizes used was between 1 bp, which is equivalent to a direct comparison between the positioning site directly below the experimentally determined *in vivo* dyad position, to 151 bp, which incorporates the positioning signals ±75 bp either side of each *in vivo* dyad position. Results are shown in Figure 2.4 a).

The relationship between the *in vivo* and *in vitro* datasets is always positive throughout the range of window sizes examined (Figure 2.4 a)). This shows that the *in vivo* nucleosome positions are located within regions containing a greater than average amount of *in vitro* nucleosome positioning signals.

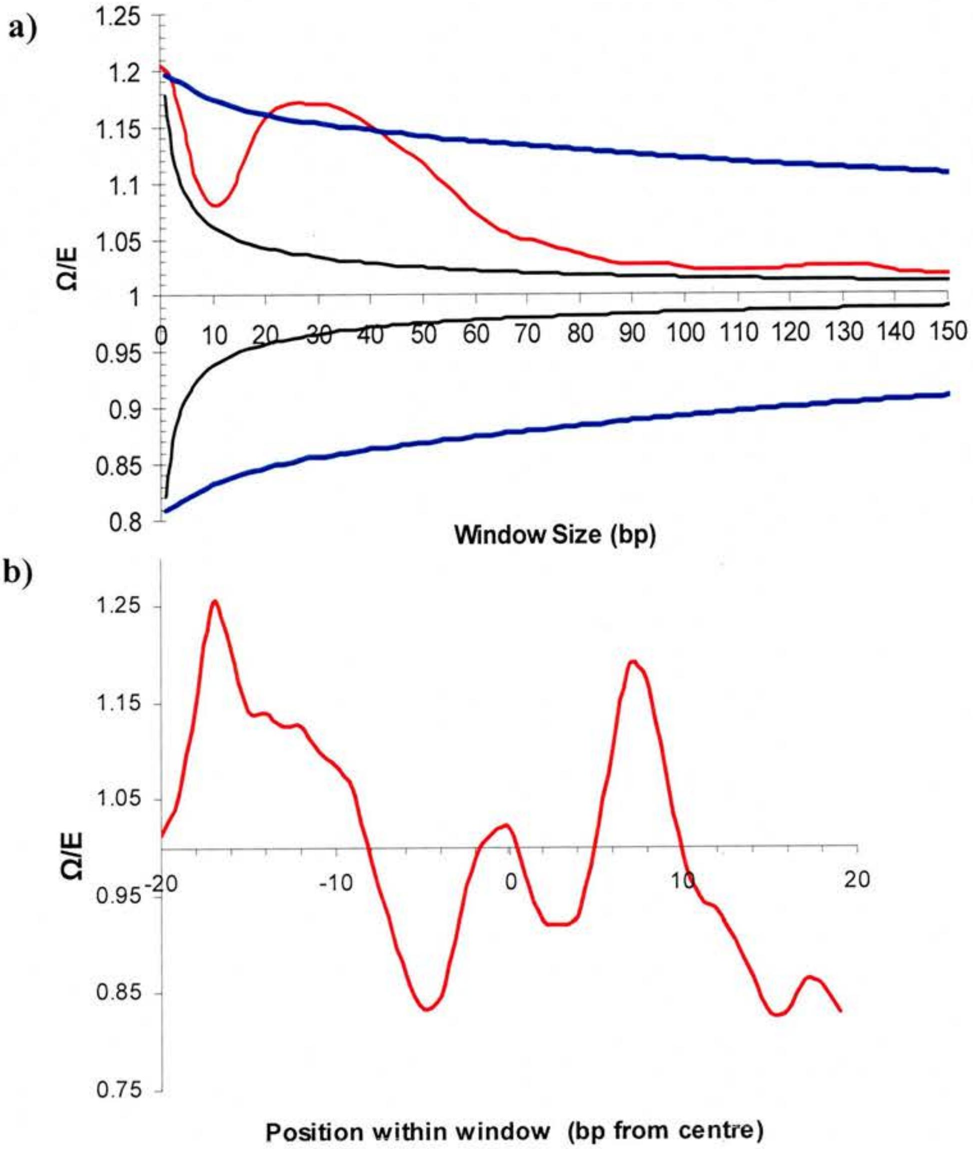


**Figure 2.4: a) Relationship between the ratio of the summed observed (Ω) and expected (E) *in vitro* positioning signal intensities contained within windows centred on the dyads of the 50 *in vivo* nucleosome positioning sites is presented as a function of window size (red). Also shown are the one standard deviation envelopes obtained by randomisations of the *in vitro* (black) and *in vivo* (blue) data sets (see text). b) Relationship between the ratio of the summed observed and expected *in vitro* signal intensities for each nucleotide position within 50 windows centred on the dyads of the 50 *in vivo* nucleosome positioning sites is presented as a function of position within a 41 bp window.**

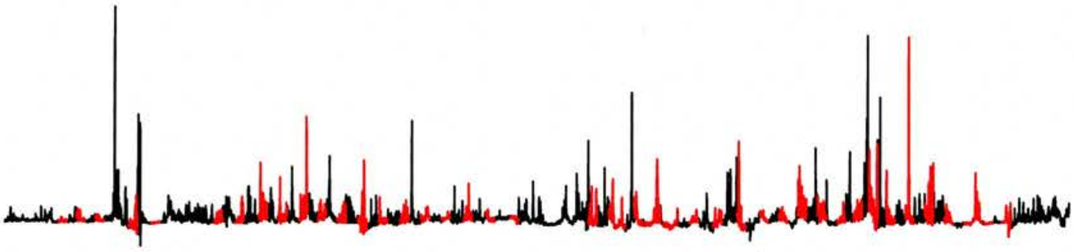


**Figure 2.4c: The location of the identified BLG *in vivo* nucleosome positions relative to the ME *in vitro* positioning map for BLG. The *in vitro* map is in black, whilst overlaid in red are the positioning sites within a window ±20 bp from each of the 50 identified *in vivo* positioning sites.**

Perhaps the most striking feature, however, is the significant variation in Ω/E as a function of window size: the relationship is strongest at very small window sizes, falling as the window size approaches 10 bp, only to rise again as the window size approaches 20 bp. This second peak in the Ω/E profile extends over a range of window sizes from approximately 20 to 60 bp, with the peak located at a window size 31 bp (±15 bp each side of the *in vivo* dyad). Ω/E should tend towards 1 as window size is increased, as when the window size is large enough such that the 50 combined windows encompass the entire sequence, the value of Ω/E must by definition be 1. This behaviour is visible as the window size approaches 151 bp.

To gain an understanding of the significance of this profile, two separate controls were undertaken. The first, "randomisation of *in vitro*" in Figure 2.4 a), involved 100,000 different shuffles (randomisations) of the *in vitro* dataset whilst keeping the *in vivo* positions stationary. This control represents a lower bound on values that may be significant using this technique, as the unique distribution of nucleosome positioning signals contained within the *in vitro* map were destroyed by the randomisation. The resulting Ω/E profiles for these 100,000 randomisations were normally distributed around unity, and the envelope plotted in Figure 2.4 a) represents one standard deviation. As a result, this envelope encompasses two thirds of all values, and therefore any value of Ω/E above this envelope indicates that the value is within the top 17% of possible values.

The second control, "randomisation of *in vivo*" in Figure 2.4 a), involved a randomisation of the 50 *in vivo* positions. The minimum dyad-dyad separation in the randomised nucleosome configurations was always >168 bp. As the properties of the *in vitro* dataset were maintained this is therefore a more reliable indicator of the significance of the correlation profile in Figure 2.4 a). As with the Randomised *In vitro* data, the resulting values were normally distributed around a mean Ω/E value of 1. Any value above the envelope plotted is within the top sixth of the possible Ω/E values for each window size.

Compared to the randomised *in vitro* control, the original correlation profile is seen to be significant at one standard deviation throughout all window sizes plotted, whilst being highly significant (greater than 2 standard deviations) at window sizes between 21 and 61 bp. Generally, randomisation of the *in vitro* data does not produce *in vitro* maps that improve the relationship to the *in vivo* data. In particular, it does not reproduce the broad peak in the 21-61 bp range, with the upper Ω/E envelope monotonically decreasing as window size increases, indicating that some property of the original *in vitro* data is responsible for this distinctive feature of the Ω/E profile.

The second control, randomising the *in vivo* positions, represents a control on the significance of the features of the *in vitro* map. The resulting envelope, at one standard deviation, is substantially greater than in the *in vitro* randomisation control, which reflects the unique properties of the *in vitro* dataset. Compared to the randomised *in vivo* analysis, the original correlation profile is not significant at small window sizes, except at a window size of 1, but is significant at window sizes between 21 and 61 bp. This control demonstrates that randomisation of the *in vivo* dataset will more often produce nucleosomal arrangements which display an improved relationship to the *in vitro* data, by optimising coincidence with the periodic strong positioning sites in the ME data set. Although many strong *in vitro* positioning sites are found within ±20 bp of the *in vivo* positions, there are a number of strong sites that are not contained within these windows (Figure 2.4 c)). This implies that the positions adopted by nucleosomes in liver nuclei are not dictated solely by the strength of the positioning information inherent in the DNA sequence, as there are more optimal nucleosome configurations which contain more ME positioning signals than the identified *in vivo* positions.

A significant aspect of Figure 2.4 a) is the hump located between window sizes of 20 and 60 bp. This suggests that there is relationship between *in vivo* and *in vitro* positioning sites located between ±10 to 30 bp of the identified *in vivo* nucleosome positioning sites. There are a number of possible explanations for this peak within the Ω/E profile. For instance, if the *in vivo* positions are located in regions of the *in vitro* map where strong positioning sites occur with a 10 bp periodicity, this would explain the observed shape. Positioning sites where alternative overlapping positions share the same rotational setting have frequently been viewed as characteristic of nucleosome positioning sites (Simpson, 1991; Thoma, 1992). If this were the case, as the window size is increased, the 10 bp periodicity in the *in vitro* data should result in peaks in the Ω/E profile at window sizes in multiples of 20 bp. As the profile only shows the hump from window sizes of 21 to 61 bp, this suggests, if this is a factor, that it is only significant within the first ±30 bp from the *in vivo* positioning sites. There is, though, a lack of any corroborating 10 bp period detected by the BLG PSD analysis (section 2.2.2).

However, a more direct demonstration of the ~10 bp weakly periodic nature of the *in vitro* dataset in those regions centred on the *in vivo* positions is provided by considering Ω/E for a fixed window size (40 bp) as a function of position within the window. It is notable in the analysis in Figure 2.4 b) that although there is a peak in the *in vitro* positioning data at the centre of the window, the peaks that occur upstream and downstream, with an approximate 10 bp spacing, are of greater amplitude. This suggests that on average the *in vivo* positioning sites are not necessarily aligned with the strongest available *in vitro* positioning site but tend to be located either 10 bp upstream or downstream of these sites.

Previous research has indicated the possibility that the centre of chromatosome may be shifted by approximately 10 bp relative to the positioning site where the histone octamer is in fact located (Muyldermans and Travers, 1994; Travers and Muyldermans, 1996; Travers and Drew, 1997). It is therefore possible that the *in vivo* nucleosome positions could be shifted in a similar fashion, which would explain, at least in part, the elevated Ω/E distribution observed in the Figure 2.4 b).

### 2.4.4 Summary

The results presented within this section demonstrate that the relationship between the *in vivo* positions and *in vitro* positioning signals is always positive, even though the *in vivo* positions are not located close to some of the highest affinity sites (Figure 2.4 c)). Further, it seems that, within the associated errors of the technique, there is an explanation for the unique shape of profile seen in Figure 2.4 a). Collectively, these observations strongly suggest that there is a quantitative relationship between the *in vivo* and *in vitro* nucleosome positioning maps. This adds confidence to the proposition that global sequence dependent *in vitro* nucleosome positioning signals play at least some role in determining *in vivo* positions.

## 2.5 Computational Prediction of Nucleosome positioning

### 2.5.1 Introduction

Over the years since the discovery of the nucleosome, a large number of methods to analyse and predict the distribution of positioned nucleosomes on DNA have been proposed. Some notable examples have been published over the past 25 years (Trifonov and Sussman, 1980; Mengeritsky and Trifonov, 1983; Calladine and Drew, 1986; Satchwell *et al.*, 1986; Drew and Calladine, 1987; Uberbacher *et al.*, 1988; Fitzgerald *et al.*, 1994; Staffelbach *et al.*, 1994; Ulyanov and Stormo, 1995; Baldi *et al.*, 1996; Ioshikhes *et al.*, 1996; Levitsky *et al.*, 1999; Stein and Bina, 1999; Levitsky *et al.*, 2001; Levitsky, 2004). The resulting prediction programs of two of these approaches will be compared with the nucleosome positioning signals determined *in vitro* by ME.

Sequence motifs, as well as di- and trinucleotide composition, have been used in many analyses of nucleosome positioning. Such factors are known to have an influence over the intrinsic curvature and bendability of DNA, and hence are thought to be important determinants of sequence driven nucleosome positioning (Satchwell *et al.*, 1986; Simpson, 1991; Thoma, 1992; Widom, 1998; Widom, 2001; Kiyama and Trifonov, 2002).

The analyses presented here will concentrate on the BLG sequence, as previous studies have analysed theoretical predictions for rotational and translation positioning on the promoter of the chicken $\beta^A$ globin gene (Kefalas *et al.*, 1988), as well as a 66 kbp region of the human β-globin locus (Wada-Kiyama *et al.*, 1999b), including the β globin promoter region mapped by ME (Yenidunya *et al.*, 1994). Neither study shows particular resemblance to the globin ME positioning map, although the resolution of the prediction is very low in the Wada-Kiyama *et al.* (1999b) study.

## 2.5.2 Drew-Calladine algorithm

### 2.5.2.1 Introduction

Drew and Calladine (1987) proposed a method for predicting nucleosome positioning, which is reported as a refinement of their previous attempt (Satchwell *et al.*, 1986). The algorithm relies on a matrix of rotational preferences for specific dinucleotides situated in certain positions relative to the outwards facing point of the minor groove of the double helix when aligned on the histone octamer. The likelihood of finding a nucleosome in a defined rotational position, on a given 124 bp sequence, is calculated as the sum of the probabilities of finding each of the dinucleotides in the angular orientation defined by the position of each in the 124 bp window. The data presented in this section was generated using the Patterton and Graves (2000) implementation of the algorithm.

### 2.5.2.2 Results and Discussion

The Drew-Calladine prediction for nucleosome positioning in the region around the BLG promoter is plotted in Figure 2.5 a), whilst Figure 2.5 b) charts the predicted nucleosome positioning affinity for the entire BLG gene region mapped experimentally by ME. The resulting prediction is then directly compared via a scatter plot with the nucleosome positioning signal assessed by ME for each individual positioning site (Figure 2.5c).

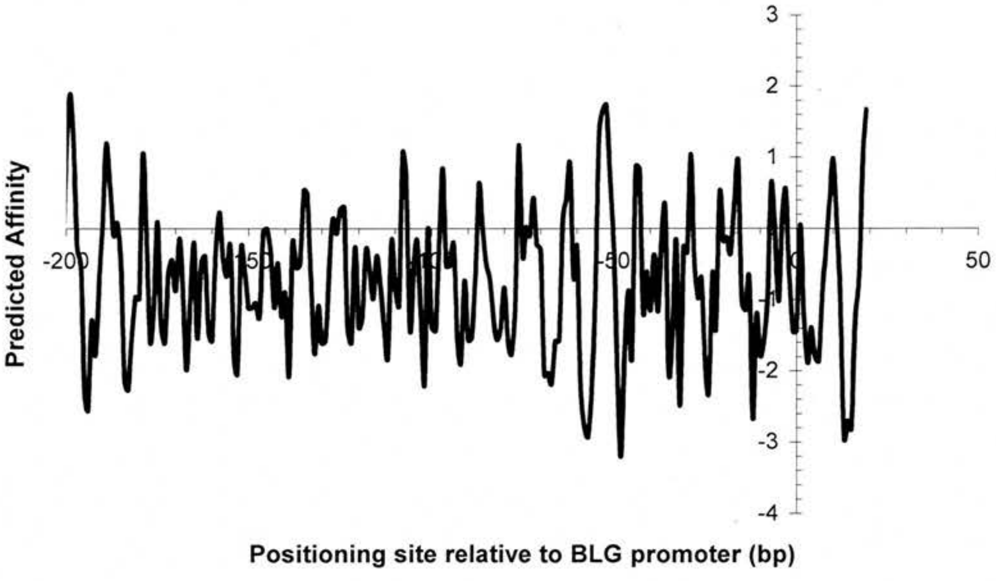


**Figure 2.5a: Predicted positioning affinity using the Drew-Calladine algorithm for sequences in and around the BLG promoter region. +2 represents a good fit to nucleosome positioning sites, whereas -2 represents a poor fit.**

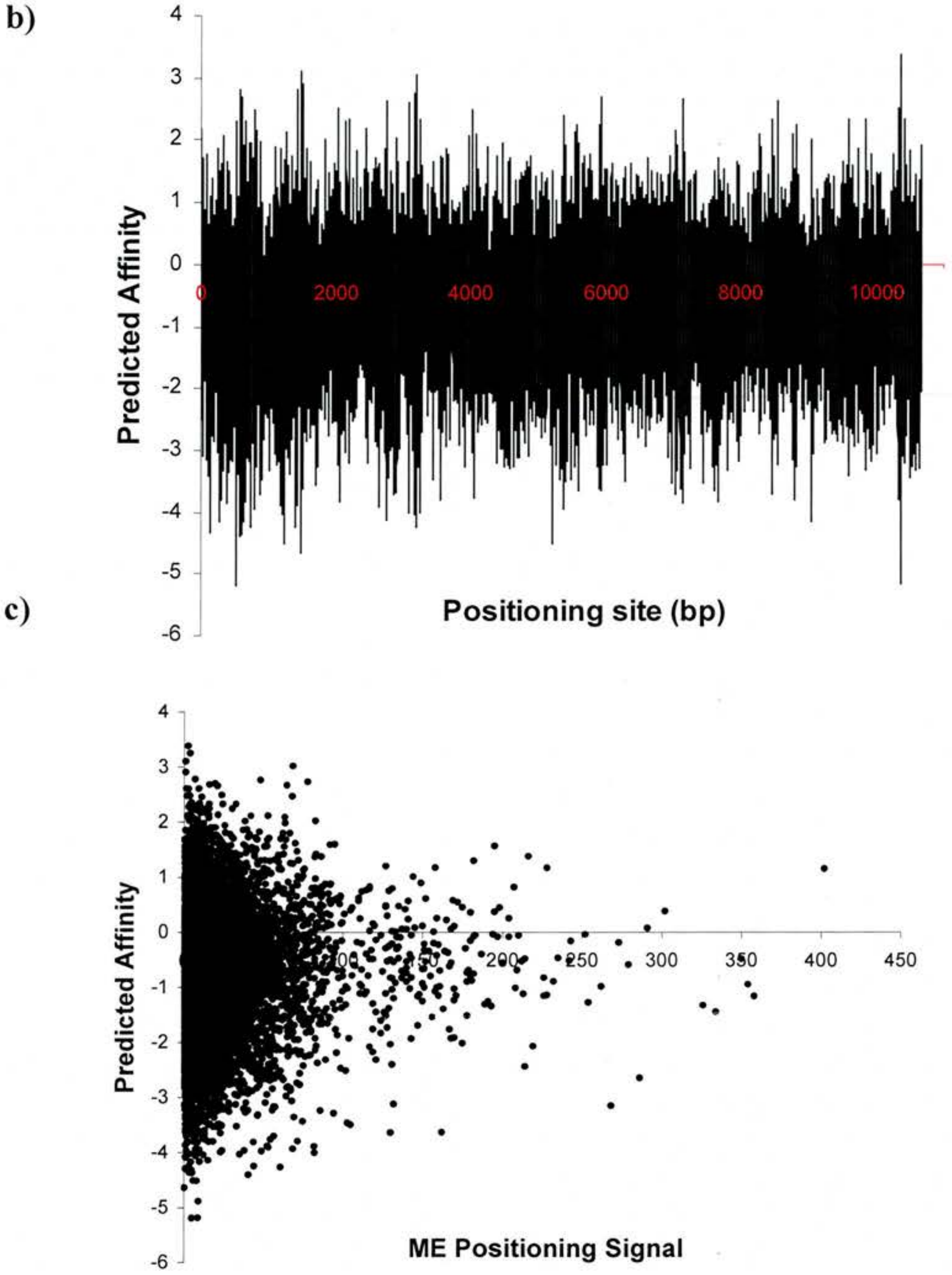


**Figure 2.5: b) Predicted nucleosome positioning affinity (by the Drew-Calladine algorithm) for DNA contained within the BLG gene region. c) Scatter plot of ME assessed positioning signal and the corresponding positioning affinity predicted by the Drew-Calladine algorithm. Correlation coefficient = 0.0009**

Perhaps the most fundamental difference between the ME positioning signals and the predicted nucleosome affinities lies in their distinctly different periodic nature. The Drew-Calladine algorithm assigns preferences for nucleosome placement based on angular orientation of sequence motifs relative to the outwards facing point of the minor groove. Therefore, the predicted affinity will be cyclical with integer multiples of the DNA helical repeat. Consequently, a 10-11 bp periodicity is effectively built into the prediction, which is logical given the design and theoretical underpinning of the algorithm. However, there is not a detectable periodicity of 10 bp evident in the BLG and globin ME datasets. Rather, as shown in section 2.2.2, periodic signals observed within in the ME datasets tend to be considerably longer in range, typically >100 bp.

These fundamental differences help explain the lack of any discernable correlation between the Drew-Calladine prediction and the ME datasets in Figure 2.4 c).

#### 2.5.2.3 Summary

It is apparent that the Drew-Calladine prediction for nucleosome positioning does not resemble the nucleosome positioning signals found experimentally by ME. Indeed, in the ~22 kbp mapped by ME, only one short region containing what appears to be clear 10 bp repeats has been noted, in the Human H19 gene (Davey *et al.*, 2003). The fact that ME detected such repeats where they exist in the H19 gene strongly suggests that similar features are not present in either the BLG or globin gene regions.

However, there is evidence to suggest that translational positioning signals are encoded by sequence elements in addition to those specifying rotational positioning (Negri *et al.*, 2001). It may be that nucleosome positioning signals unique to translational positioning are stronger than those shared with rotational positioning, leading to the inaccuracies in the predicted likelihood of positioning for the majority of sequences.

Therefore, if one assumes that the ME nucleosome positioning signals encode at least some aspects of histone octamer positioning affinity, then these results further call into question the role of rotational settings being the sole, or even dominant, determinant of long range translational nucleosome positioning in general, and in the BLG and globin gene regions in particular.

Approaches using DNA structural parameters to predict nucleosome positioning have also previously been made using the BEND program (Goodsell and Dickerson, 1994), for example (Wada-Kiyama *et al.*, 1999a). Previous studies have resulted in a similar lack of correspondence between the predicted positioning affinity and the experimental ME data. Similarly, there is no correspondence between the ME positioning signals and base-stacking interactions (Gardiner *et al.*, 2003) (data not shown).

### 2.5.3 RECON: Predicting nucleosome formation potential based on dinucleotide abundance

#### 2.5.3.1 Introduction

Levitsky *et al.* (2001) have proposed a method to calculate the "nucleosome formation potential" (NFP) of a given DNA sequence. The nucleosome prediction algorithm, named RECON, obtains a function which optimally discriminates between 86 DNA sequences which have been observed experimentally to strongly position nucleosomes (Widlund *et al.*, 1997) and 40 DNA sequences which nucleosomes avoid (Cao *et al.*, 1998), using the dinucleotide frequencies within a 160 bp sliding window which is partitioned into 14 separate windows. The partition itself is designed to optimise the discrimination in dinucleotide space between the two datasets.

The authors present evidence that nucleosome formation potential varies depending on the class of the gene, with genes only expressed in specific tissues having a significantly higher NFP than more widely expressed genes. Further, they report that exons have lower nucleosome formation potential than introns and Alu repeats.

The BLG and globin genes are classified as tissue-specific genes, as they are only expressed in mammary gland and blood cells respectively. An analysis of the promoters contained within the ME mapped gene regions would therefore be an interesting comparison. Unfortunately, portions of the adult β-globin promoter sequence were rejected by the RECON program for having dinucleotide content which differed too significantly from the integrated sequence derived from the training dataset and were therefore rejected by the program for having "abnormal dinucleotide content" characteristic of "artificial sequence". However no such problem existed for the BLG and embryonic ε-globin promoters.

### 2.5.3.2 NFP promoter regions of tissue-specific and housekeeping genes

One of the most striking results from RECON is the difference between the nucleosome formation potential (NFP) found in the promoter region of 3 gene classes: housekeeping, widely expressed, and tissue-specific (Levitsky *et al.*, 2001). The promoter regions of tissue-specific genes are found to have significantly higher NFP than more widely expressed genes, with housekeeping genes (which are always being expressed and therefore undergo limited regulation) having the lowest (Figure 2.6a). This is an interesting result as nucleosome placement has long been suspected of playing a significant role in gene regulation. It would be preferable to have a large degree of control over nucleosome positioning within the promoter region as access to the DNA sequence within a nucleosome is inhibited (Anderson *et al.*, 2002; Li and Widom, 2004; Li *et al.*, 2005).

In the following analyses, the investigation of NFP found within the promoter regions was repeated with a different set of 54 housekeeping and 34 tissue-specific genes (Figure 2.6b). These sequences were obtained by extracting entries which pattern matched the keywords "housekeeping" and "tissue-specific" from the Eukaryotic Promoter Database (Bucher and Trifonov, 1986; Schmid *et al.*, 2006).

For the 88 matching sequences used in the current repeat study, the discrimination between the tissue specific and housekeeping genes is not as successful as reported by Levitsky *et al.*, with the mean NFP of the promoter regions of the new dataset being closer to 0 than the ~-1 reported by the authors from their sequence selection. Nonetheless, the program is still able to reliably discriminate between gene types for a collection of promoters, although the results of Figure 2.6 b) do suggest that the ability to do so for individually examined promoter sequences is questionable. In this sense, RECON is not suited for automated classification of gene types based on an analysis of promoter regions.

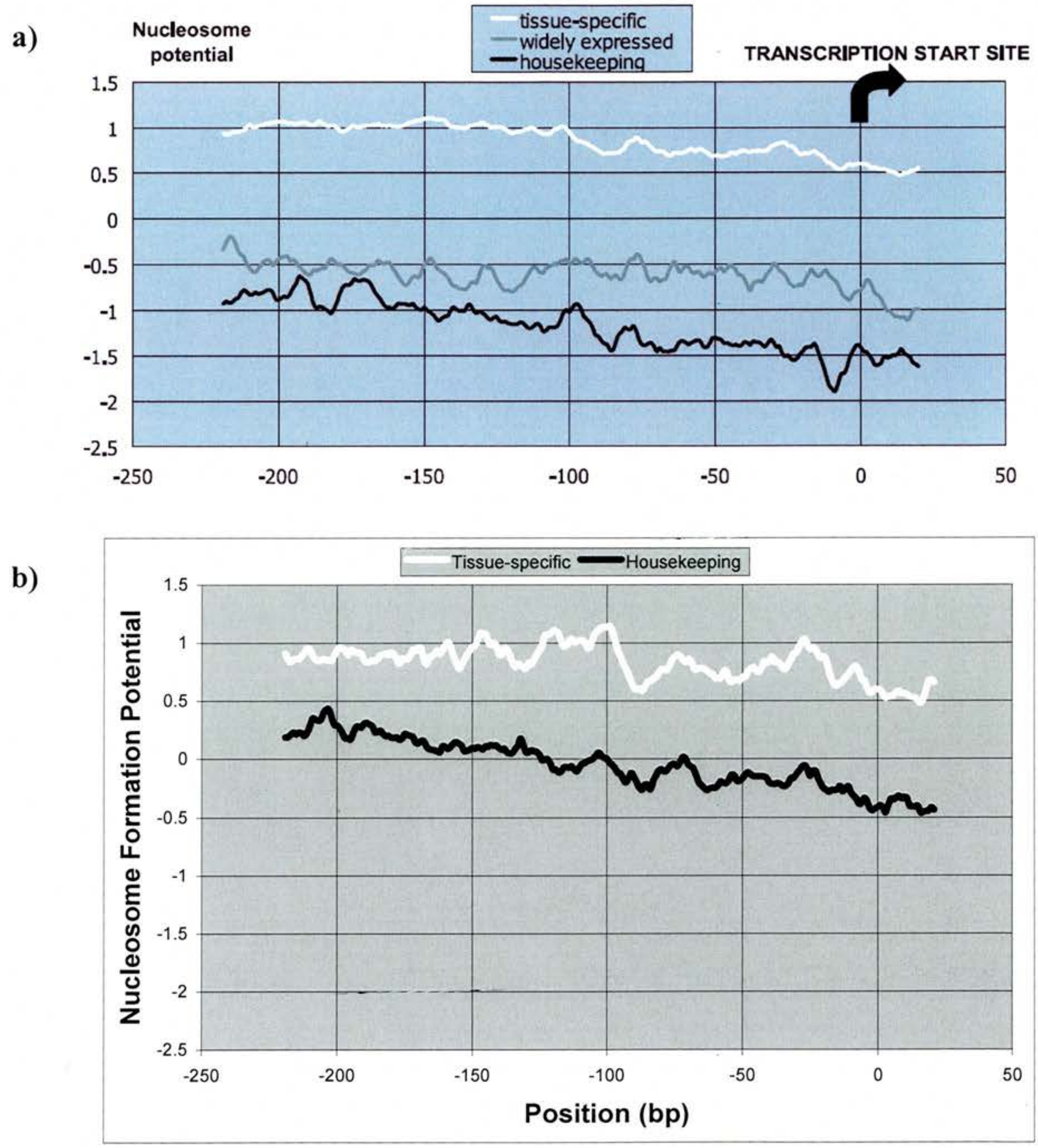


**Figure 2.6: a) Average of the NFP within the promoter regions of 23 housekeeping, 30 widely expressed, and 141 tissue-specific genes. Reproduced from (Levitsky *et al.*, 2001). b) Average of the NFP within the promoter regions with a different set of 54 housekeeping and 34 tissue-specific genes. The authors scaled NFP so that a value of +1 represents a close match to the positive training set (nucleosome positioning sequences), whereas -1 represents the NFP found in random sequence.**

### 2.5.3.3 NFP within the promoter regions of the BLG and ε-globin genes

Both BLG and β-lactoglobulin genes are classified as tissue-specific. Despite the misgivings expressed above, an examination of the promoter regions of the promoters found within the BLG and ε-globin genes serves as two individual examples of the variance in NFP found within promoter regions.

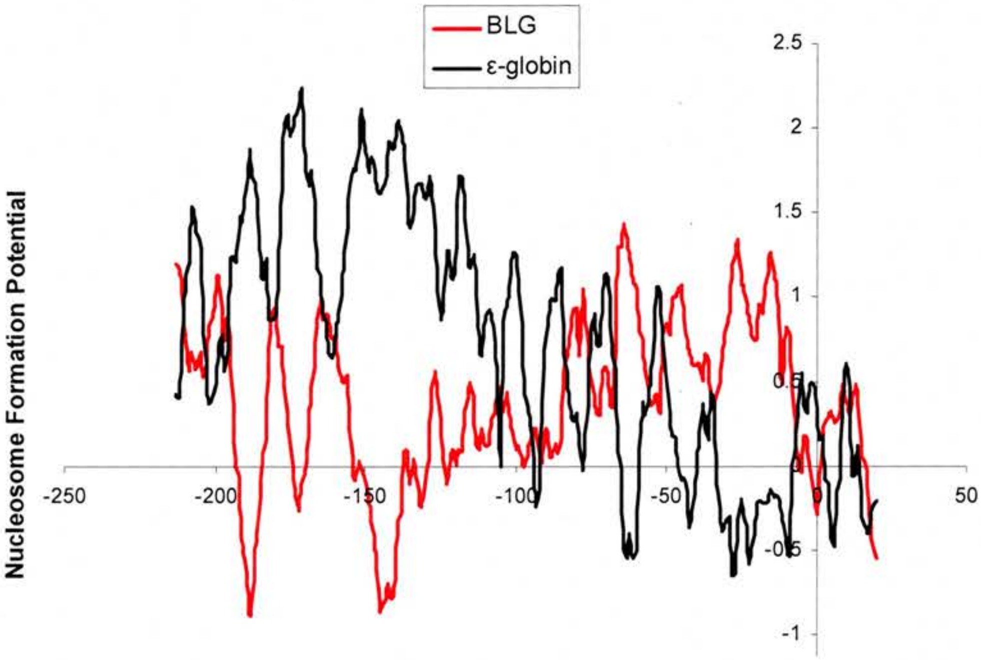


**Figure 2.7: NFP within the promoter regions of the BLG (red) and ε-globin (black) genes, both of which are classified as tissue specific genes. Positioning sites are scaled relative to transcription start site**

Both promoter regions plotted in Figure 2.7 are broadly within the reported range of tissue specific genes in Figures 2.6 a) and b), although it would be difficult to firmly classify either gene as tissue-specific or housekeeping solely on the basis of the NFP contained within its promoter region. The NFP within the BLG promoter oscillates between positive and negative values within the -200 to -150 bp region, after which the NFP becomes strongly positive until the transcription start site. On the other hand, the ε-globin gene has strongly positive NFP until around -60 bp from the transcription start site, whereupon it falls below 1 and oscillates ±0.5. Throughout both profiles, a periodic signal of ~10 bp is noticeable from auto-correlation analyses (data not shown).

### 2.5.3.4 Nucleosome Formation Potential with the BLG gene region

The analysis was then extended to the entire BLG gene region mapped by ME (Figure 2.8). The RECON program rejected the sequence between 8036 and 8351 bp for abnormal dinucleotide content.

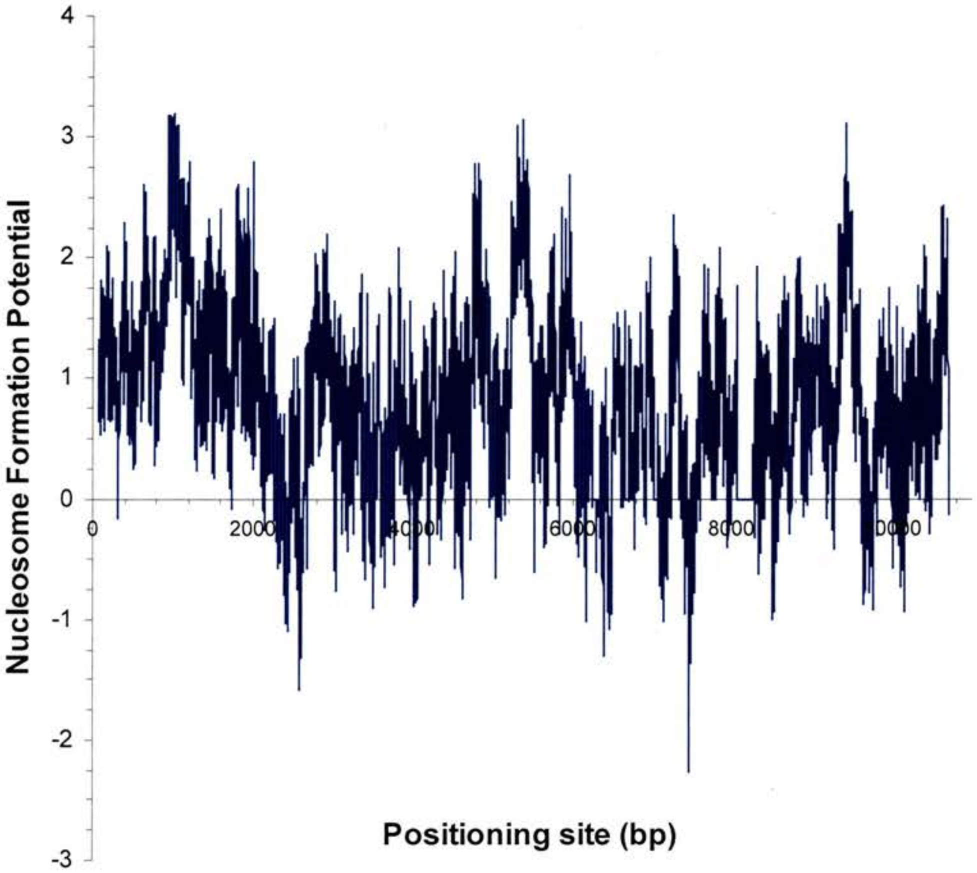


**Figure 2.8: RECON prediction of the nucleosome formation potential within the BLG gene region mapped by ME. A value of +1 corresponds to the mean NFP of the nucleosome positioning training set and -1 to the mean NFP in random sequence.**

Similar to the prediction by the Drew-Calladine algorithm, the overall properties of the prediction appear to be inconsistent with the basic structure of BLG ME dataset. For instance, significant variances in affinity within a few base pairs, which is a notable feature of the ME datasets, does not appear to be a feature of the RECON prediction, which appears to be smoothed in comparison.

Interestingly, most of the gene region appears to have somewhat similar NFP values to those found within the BLG promoter region.

### 2.5.3.5 Relationship between NFP and ME positioning signals for BLG

As previously demonstrated, scatterplots are a useful method for graphically assessing potential correspondence between two (or more) datasets. Plotted below, in Figure 2.9 is the scatterplot of NFP against ME assessed positioning signal for the BLG sequence.

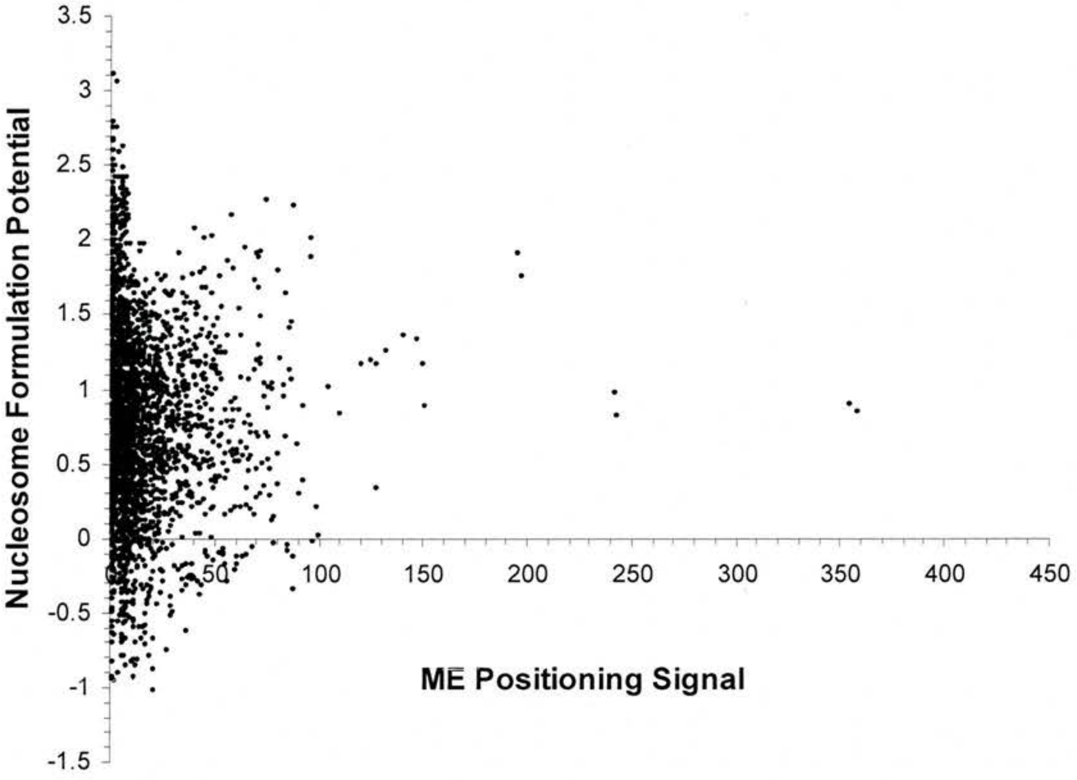


**Figure 2.9: Scatter plot of the ME positioning signals and the corresponding nucleosome formation potential assessed by the RECON algorithm for the BLG sequence. Correlation coefficient is 0.014.**

As with Figure 2.5 c), there is no discernable relationship between the two sets of data, although 488 BLG positioning sites (~5%), including some of the strongest sites, were rejected by RECON as having "abnormal dinucleotide content".

In its favour, the NFP of the 6 strongest positioning sites are all assessed to have a NFP of around 1 or greater. The average for the entire gene region is 0.785.

Of the top 300 BLG positioning sites, 246 were assessed as having a NFP greater than zero. On the other hand, of the 5000 lowest affinity sites, only ~10% are given a NFP of below 0, whilst half are given an NFP value of greater than 1; recall that +1 is the mean value of the (positive) nucleosome positioning sequences used to train RECON. The two sites with the strongest NFP (> 3) were also assessed by ME to be two of the *lowest* affinity positioning sites. The peak at 5350 bp in Figure 2.8 is located within a relatively rich A+T region (Figure 2.11 b)).

### 2.5.3.6 Summary

It is evident that the dinucleotide content within promoter regions appears to be generally similar to that found in the database of sequences which form stable nucleosomes. However, whatever features of the dinucleotide content of the training datasets that RECON represents, it would appear that there is little or no relationship to translational nucleosome positioning. Understanding why such a promising method such as this should fail is therefore of some importance.

There are three main possibilities as to why RECON fails to predict ME positioning signals: a deficiency in the discriminant analysis method (and/or the current underlying tenets of sequence dependent nucleosome positioning); a deficiency in the training dataset; or finally that the current interpretation of ME positioning signals is not accurate.

A potential issue with the training sequences is that there is not a quantitative assessment of the positioning affinity for the sequences observed to preferentially position nucleosomes *in vivo*. For instance, an average positioning site located within a region of weak nucleosome positioning signals could well be observed as preferentially positioning nucleosomes even though it does not have a particularly strong positioning affinity. As a result, it is possible that not particularly strong positioning sequences could have been incorrectly identified as forming stable nucleosomes. The converse effect could also have resulted in a misclassification of "anti-nucleosome" sequences used in the RECON training phase.

Further, it is possible that training sequences were highly similar to each other, which would limit the variability of the set the program was trained with. Only 193 sequences were used, with the majority of these from mouse (41%) and yeast (20%). This could potentially limit the applicability of the program to other DNA sequences. It is therefore possible that the program fails to accurately predict DNA sequences found in the highly tissue-specific gene regions of BLG and globin. It should be noted that the program rejected the dinucleotide content of significant stretches of sequence from both genes, which indicates the lack of similarity to the training sequences used.

### 2.5.4 Comparison between Drew-Calladine and NFP

Two prediction algorithms have been tested against the ME data, and neither has demonstrated any particular correlation with nucleosome positioning signals from ME experiments. Nonetheless, an appealing comparison would be to contrast the prediction of the two approaches. Presented in this section is a direct comparison of the prediction of the two methods used in the preceding sections. The circular cluster distribution in Figure 2.10 indicates that there is no relationship between the Drew-Calladine predicted affinity and Levitsky RECON.

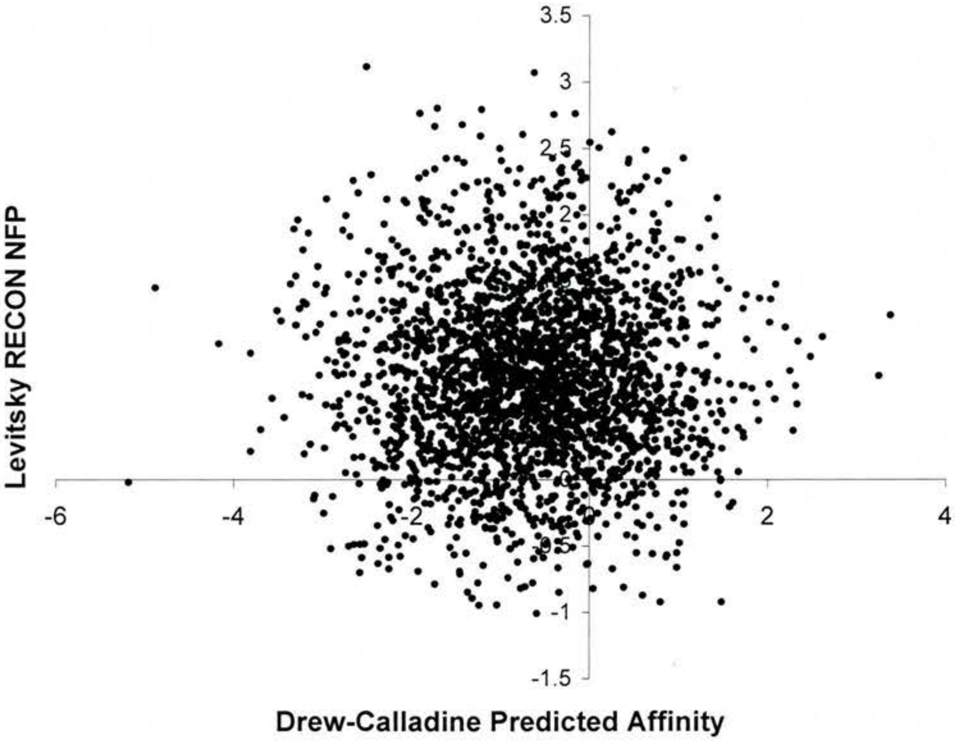


**Figure 2.10: Scatter plot of predicted affinities for BLG between the Levitsky RECON program and the Drew-Calladine algorithm.**

### 2.5.5 BLG Sequence Analysis

Both of the prediction methods used in this section rely on the dinucleotide content of the sequences analysed. Given the failure of both methods to reliably predict nucleosomal affinity, it is appropriate to take a closer look at BLG sequence to gain some insight as to why these two methods are not working as intended.

An examination of the dinucleotide frequencies within the strongest and weakest positioning sites sheds some light on why both approaches seem not to be successful. Firstly, it is useful to examine the dinucleotide content of the strongest positioning sequences. Counter-phase oscillation of AA and TT dinucleotides has commonly been regarded as a prominent (if not the principal) determinant of nucleosome positioning (Widom, 2001; Cohanim *et al.*, 2006). Consequently, the second analysis will examine the location of AA and TT dinucleotides with respect to the nucleosome dyad, to identify any patterns within the strongest positioning sites which are perhaps not present in the weakest.

### 2.5.5.1 Sequence Composition of the 100 strongest nucleosome positioning sites

The 146 bp sequences of the 100 highest affinity positioning sites were analysed. The 100 top positioning sites are C+G rich in comparison to the gene region as a whole (Figure 2.11a), which is itself C+G rich (Figure 2.11b), leading to a conclusion that the highest affinity sites are highly C+G rich.

Compared to the average within the gene region as a whole, there is approximately half the number of AAs and TTs in the 100 top binders than the rest of the positioning sites. Indeed, even in C+G rich regions, one would have expected, if the prevailing theories on sequence dependent nucleosome positioning are accurate, that there would be an increase in the number and periodicity of AA/TTs in stronger positioning sequences. These dinucleotides have often been linked with nucleosome positioning, as they confer increased DNA flexibility, which in turn lowers the energy cost required to bend the DNA around the histone octamer to form a nucleosome. On the other hand, GC, reported to be ~3 times less prevalent in nucleosome positioning sequences (Satchwell *et al.*, 1986), are in fact enhanced in the strongest positioning sites. However, this result may have been affected by the differing base composition of the nucleosome positioning sequences. In particular, the C+G richness of the BLG gene region may skew the observed frequency of the GC dinucleotide, although the occurrence of any given dinucleotide within the top 100 nucleosome positioning sequences is normalised with respect to its occurrence within the BLG gene region. This should minimise this potential effect.

If nucleosome positioning does play an important role in gene regulation, and given the strongly tissue-specific nature of BLG expression, one would expect strong nucleosome positioning signals to be contained within the gene. Indeed, a positioning site with 5000 times the positioning affinity of the lowest site was characterised by visual identification and integration (Gencheva *et al.*, 2006) .

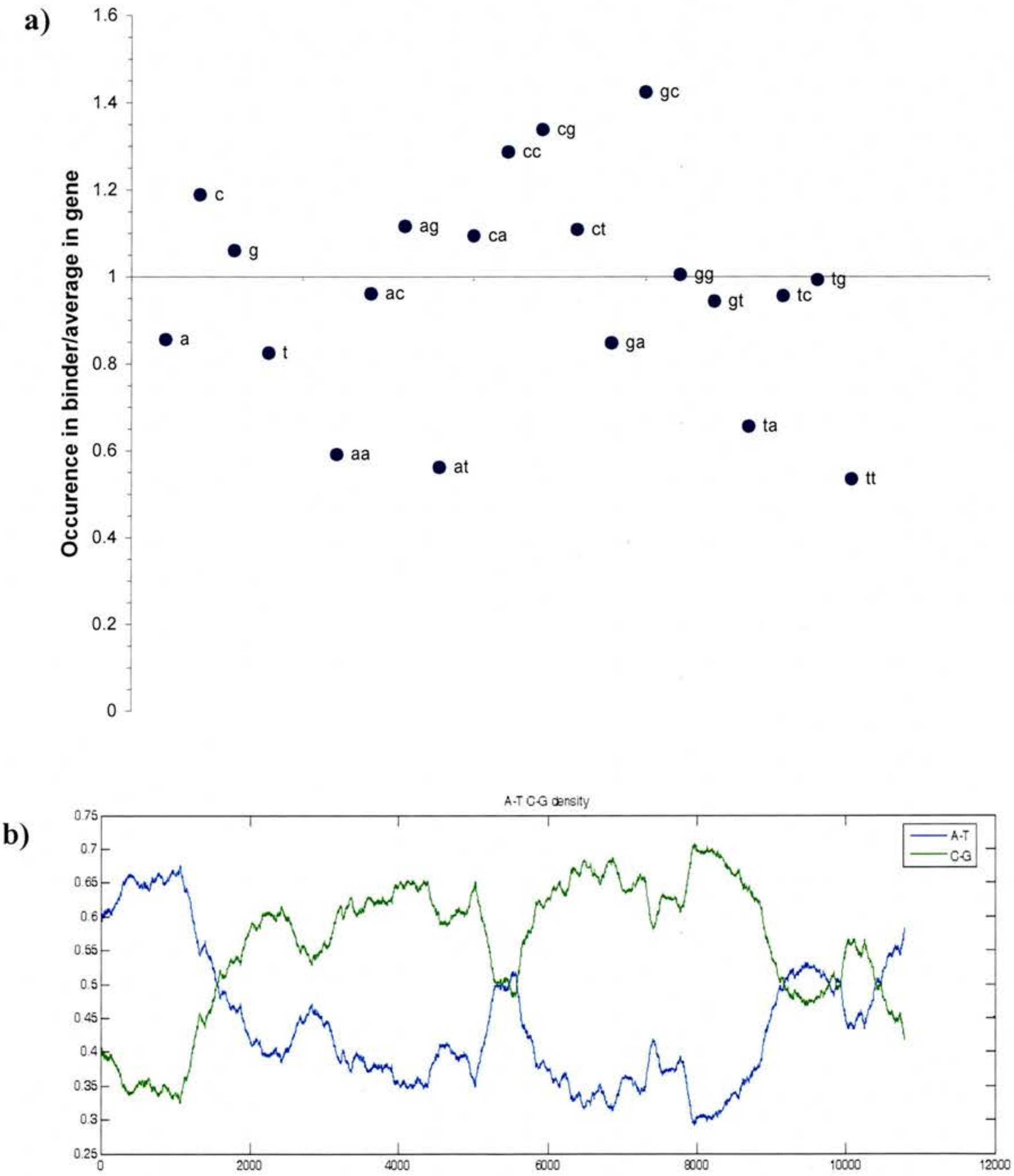


**Figure 2.11: a) Mono- and dinucleotide composition of the top 100 highest affinity BLG ME sites. b) A+T and C+G composition within the BLG gene region.**

### 2.5.5.2 AA/TT Locations within strong and weak ME positioning sites

There is a large body of literature connecting periodic occurrences of AA and TT dinucleotides phased by the DNA double helical turns and nucleosome positioning (Trifonov and Sussman, 1980; Ioshikhes *et al.*, 1996; Widom, 1996; Herzel *et al.*, 1998; Lowary and Widom, 1998; Herzel *et al.*, 1999; Tomita *et al.*, 1999; Schieg and Herzel, 2004; Cohanim *et al.*, 2005). The body of evidence that supports this AA/TT[6] pattern is such that it has recently been called the firmest indicator yet identified of strong nucleosome positioning (Cohanim *et al.*, 2006). It follows that an analysis of the locations within the highest and lowest ME affinity sites may be fruitful for explaining the inability of current algorithms to accurately predict ME nucleosome positioning signals. Figures 2.12 and 2.13 a) and b) graphically represent the occurrences of AAs, TTs and AA/TT, relative to the core particle dyad for BLG and globin respectively. The y-axis is not scaled: it represents the absolute number of occurrences.

It is not surprising that the profiles in Figures 2.12 and 2.13 bear a significant resemblance to Figure 2.5 a), as Calladine-Drew likelihood matrixes have a significant contribution from AA/TT locations.

Using autocorrelation analysis, there is some evidence of a 10 bp periodic arrangement in localised areas of both Figure 2.12 a) and b). While there is certainly evidence of a periodic arrangement within the highest affinity sequences, autocorrelation analysis reveals only a relatively weak periodicity around 16 bp. Interestingly, the set of sequences with the clearest ~10 bp AA/TT periodicity are the weakest globin positioning sequences.

More generally, it is notable that AA, TT and AA/TT appear more prevalent on the 3' end of the highest affinity site but on the 5' end of the lower affinity sites. This appears to be either a coincidence or a feature specific to the BLG positioning map, as this behaviour is not observed in globin (Figure 2.13).

[6] AA and TT are reverse complements of each other.

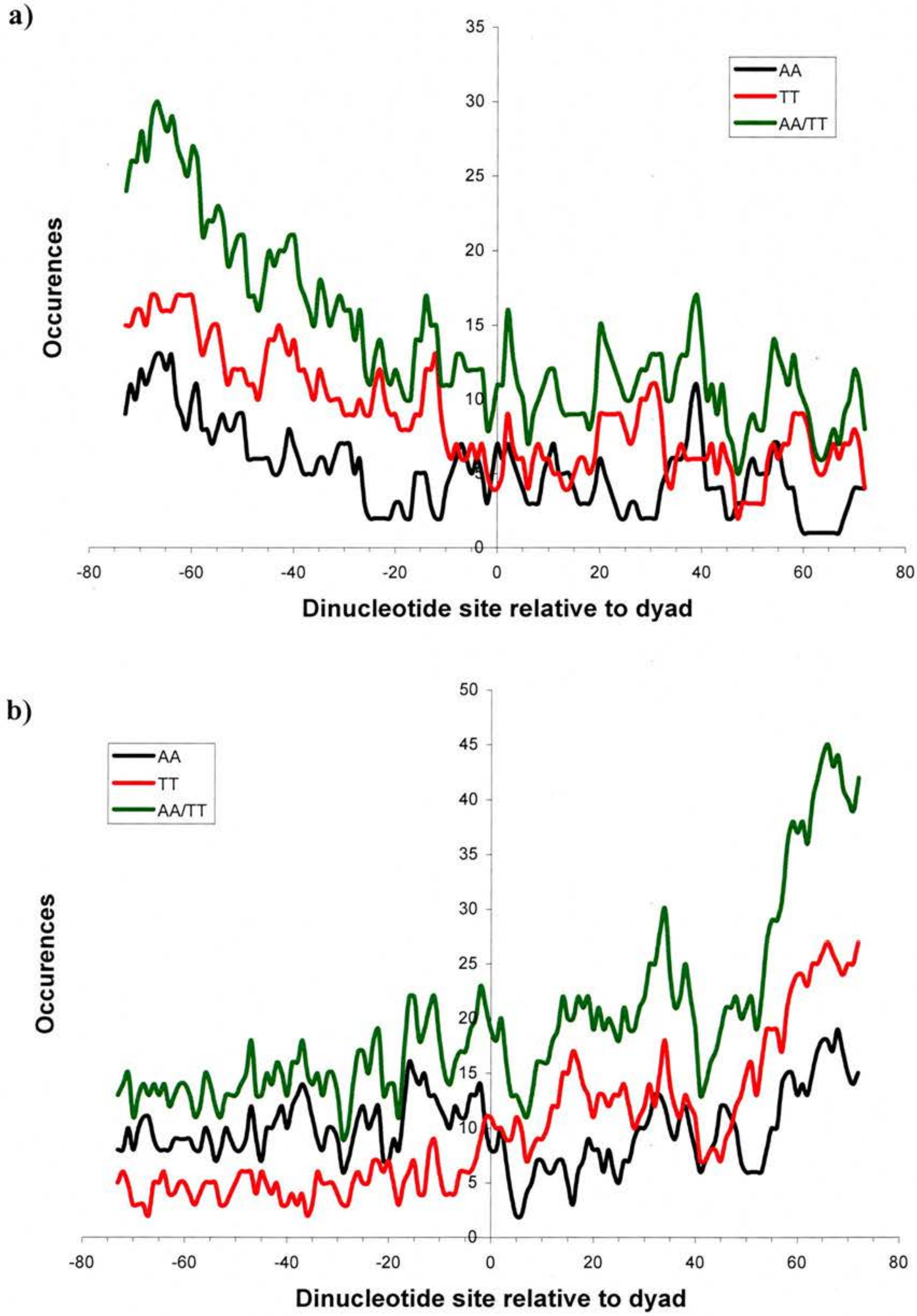


**Figure 2.12: AA, TT and AA+TT occurrences within the 200 highest (a) and lowest (b) BLG relative affinity sites . The origin of the x-axis is located at the dyad.**

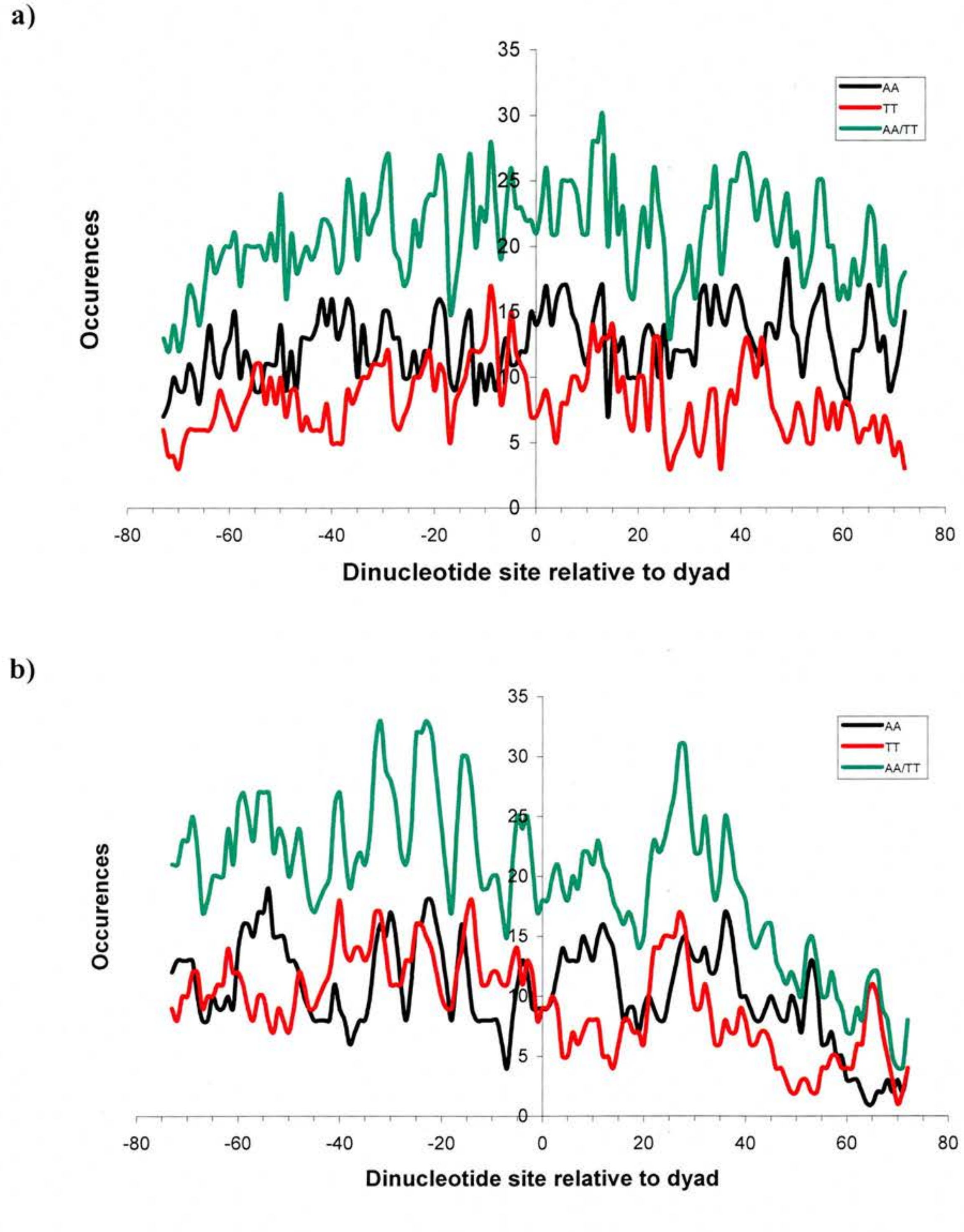


**Figure 2.13: AA, TT and AA+TT occurrences within the 200 highest (a) and lowest (b) globin relative affinity sites . The origin of the x-axis is located at the dyad.**

### 2.5.6 Summary

Given the apparent failure of both the approaches used within this section to reliably predict ME affinities, the most obvious question to ask is why they fail. Is it a consequence of a deficiency in current understanding of factors involved in positioning nucleosomes, a deficiency in the methods and/or nucleosome positioning data on which the algorithms are trained, or perhaps a misinterpretation of the ME dataset?

It certainly seems apparent that currently identified sequence motifs and dinucleotide composition-based nucleosome positioning predictions are not capable of differentiating between a strong and weak ME positioning affinity site. Further, there is little evidence to support a strong 10 bp periodic arrangement of AA/TTs within the strongest ME affinity positioning sites.

It is possible, however, that it is the current understanding of the ME datasets that is lacking, in that some of the experimental and data analysis procedures involved in generating the ME datasets currently do not work as expected. However, if this is not the case, the results presented within this section raise questions about the role of rotational positioning in general, as well as the utility of examinations of dinucleotide content and other sequence motifs for predicting nucleosome positioning both *in vivo* and *in vitro*.

One possible avenue for exploration could be to examine di- and tri- nucleotide arrangements which are spatially separate, particularly by ~80 bp as these dinucleotides will be located close to each other (and perhaps interact) when wrapped around the nucleosome. Correlated nucleotides symmetrically disposed with respect to the dyad axis have been observed to have biological implications: a nucleosome "supergroove" has been recently identified as being used as a molecular recognition site (Edayathumangalam *et al.*, 2004).

CHAPTER 3

# Simulating higher order chromatin structure via Monte Carlo methods

## 3.1 Introduction

### 3.1.1 Motivation

The ME datasets provide an experimental assessment of the affinity for the histone octamer to position itself on specific DNA sequences. The ME technique minimises possible inter-nucleosome interactions during reconstitution by having an excess of DNA to core histones (approximately one histone octamer per 500 bp). As the ratio of DNA to histone octamers is kept low, the positioning information remains unbiased by nucleosome-nucleosome interactions. However, *in vivo*, nucleosomes have a higher density and are therefore likely to interact and compete with each other for the most favourable binding sites. *In vivo* the most energetically favourable nucleosome configurations, for instance, may not include the most favourable individual site if, in so doing, that requires neighbouring nucleosomes to position on unfavourable sites.

The principal motivation behind the technique proposed here was to attempt to generate and explore the properties of credible configurations of nucleosomes based upon nucleosome positioning signals determined by ME; the goal is to bridge the gap between the *in vitro* datasets and physiological nucleosome densities by using *in silico* techniques.

To achieve this objective, a simplified 1D model was designed to simulate the competition of nucleosomes for binding sites on the DNA molecule. Such a simulation is fraught with difficulty as, using biologically relevant parameters, there are an exceptionally large number of possible nucleosome configurations. Exploring all possible configurations is therefore impossible. In addition, current understanding of the dynamics involved in positioning nucleosomes is insufficient to construct an adequate physical model for deterministic simulation techniques.

To surmount these problems, a stochastic simulation based upon Metropolis Monte Carlo methods has been developed. The only prerequisite for such a simulation is that the system can be described by at least one known probability density function (pdf).

The principal assumption of the model is that the ME nucleosome positioning maps provide quantitative positioning information directly proportional to the probability of occupancy of each binding site *in vitro*. This is a reasonable assumption, as the intensity of the bands as determined by densitometry is proportional to the amount of monomer DNA in a band, which is proportional to the population of histone octamers that were bound to that particular 146 bp of DNA sequence. Using this assumption, the probability of each site being occupied by a nucleosome should be proportional to the nucleosome positioning strength of the site. The generation of the necessary pdf is therefore straightforward, as the ME positioning maps, properly normalised, can serve as the required pdf. This approach is justified as it has been demonstrated that, if nucleosomes are given the opportunity to explore different positioning sites on the DNA molecule prior to final positioning, the positioning sites will be occupied in agreement with their Boltzmann probabilities (Lowary and Widom, 1998).

It is therefore possible, using the Boltzmann distribution, to calculate a thermodynamic quantity similar to “free energy” for each positioning site or combination of positioning sites. This allows for the generation of a discrete energy “landscape” or “lattice” for the simulation, where each discrete positioning site is assigned an energy value which represents its affinity for positioning nucleosomes as assessed by ME.

To eliminate configurations that are inadmissible on structural grounds, a minimum nucleosome-nucleosome separation constraint was imposed on the system. This constraint is variable, and various minimum separation distances were experimented with. However, in the analyses presented hereinafter, the commonly accepted value for the minimum length of DNA wrapped around the octamer in the nucleosome (168 bp) was used. 168 bp was chosen as it allows for the greatest degree of flexibility whilst still respecting the physical constraints of the biological system. Data from simulations with a nucleosome-nucleosome constraint of 146 bp (the length of DNA in the nucleosome core particle) and 180 bp (a separation commonly observed *in vivo* (Van Holde, 1989)) were also used extensively in preliminary simulations. These preliminary simulations, run at extremes of the biologically acceptable limits, demonstrate that such large-scale changes in the minimum nucleosome spacing can have an impact on the simulation output, although this is only significant in simulations where the mean nucleosome spacing (the density of simulated nucleosomes) approaches the minimum separation. More pertinently, the results presented in Chapter 4, which use the 168 bp constraint, are robust to small changes in the minimum nucleosome separation distance.

### 3.1.2 Sequence-dependent prediction of nucleosome positioning

One area that has attracted much research interest has been the effort to predict nucleosome positioning *in silico*, reducing the need for expensive and time-consuming experiments. The RECON program (Levitsky *et al.*, 2001; Levitsky, 2004), introduced in chapter 1 and explored in more detail in chapter 2, is a prominent recent attempt which has been used in recent studies to predict nucleosome positioning, for example (Wasserman and Sandelin, 2004). A major obstacle, given the apparent relative weakness and/or redundancy of nucleosome positioning signals *in vivo*, has been the lack of reliable, quantitative nucleosome positioning data. Whilst, to the author's knowledge, the dataset analysed here is presently the largest of its type (3.4 Mbp), there are several reasons why the dataset is not ideally suited for this purpose. Most importantly, whilst ME has a resolution down to the base pair level, it does have known, and as yet unresolved, experimental errors which may introduce an uncertainty of up to several bp in the determination of the precise sequence responsible for the assessed positioning signal (section 1.3). Consequently, there is a possibility that the positioning information will be shifted with respect to the underlying sequence. Other techniques, such as competitive reconstitution (Shrader and Crothers, 1989), can more reliably identify the sequence responsible for the positioning observed as they directly sequence the DNA. Secondly, unlike the competitive reconstitution method, which assesses affinity relative to a reference sequence (commonly the well characterised 5 S gene positioning sequence), ME can only assess histone octamer affinity relative to other sites on the DNA molecule being mapped. This limits the ability to make quantitative comparisons between the relative affinities of positioning sites from different experiments. Thirdly, two factors limit the scope of the data: as the data is of continuous natural DNA sequences, the sequences of two neighbouring positioning sites are offset by only 1 base pair, limiting the diversity of the sequences of the nucleosome sites mapped. In addition, the nucleosomal DNA sequences mapped come exclusively from natural gene regions, which whilst being the most relevant type given the nature of the problem, is a limited subset with respect to the overall set of possible sequence combinations.

Despite these limitations with the ME datasets, some preliminary efforts have been made in this area using machine learning pattern recognition techniques (Fraser, Allan, and Simmen, unpublished)

## 3.2 Introduction to Monte Carlo Methods

Monte Carlo (MC) methods are a class of computational algorithms used to simulate the behaviour of a wide range of physical and mathematical systems. They have been used in a wide variety of fields, from physics to economics.

Although the term "Monte Carlo methods" is relatively new (Metropolis and Ulam, 1949), the fundamental concepts underpinning the technique have been used for over a hundred years under names such as "statistical sampling". Only in the last 60 years, with the advent of computer technology, has it become possible to implement these methods on a wider, more rigorous scale.

MC methods fall under the umbrella of stochastic (nondeterministic) simulation methods, and usually involve the use of random numbers (or more often pseudo-random numbers). A stochastic process is one whose behaviour is non-deterministic in some fashion, such that the next step of the process is not fully determined by the properties of the previous step. Other common simulation methods, such as molecular dynamics (MD), use deterministic algorithms rather than stochastic methods (Alder and Wainwright, 1957). Classical (or empirical) MD simulations use Newton's equations of motion on a model of a molecular system to study the behaviour of molecules over time, whereas *ab initio* (first principles) methods employ quantum mechanics to calculate the potential energy of the system. The MD technique has been successful applied in the study of small molecules, e.g. (Karplus and McCammon, 2002), but extending these techniques to large macromolecules is exceptionally challenging, the main obstacle being the computational time required for complex simulations of this type. In addition, current knowledge of the structure of nucleosomes, and the interaction between the core and linker histones with both the DNA and with each other is at a basic level, sufficient for only small scale simulations (Bishop, 2005).

Monte Carlo methods are most often used, when simulating a thermal system, to calculate the expectation value <Q> of an observable quantity Q. In the following analysis, the observable quantity is some property of the configurations of nucleosomes, which varies with different densities of simulated nucleosomes.

### 3.2.1 Equilibrium and the Boltzmann Distribution

In 1902, the American mathematical physicist JW Gibb, demonstrated that the equilibrium occupation probabilities, $p_\mu$, for a system in thermal equilibrium with a reservoir at temperature T (in Kelvin) were given by the following equation:

$$p_\mu = \frac{1}{Z} e^{-\beta E_\mu} \tag{3.1}$$

where $E_\mu$ is the energy of state $\mu$. Note that it is convention that the symbol $\beta$, known as the "thermodynamic beta", is used to represent $(k\mathrm{T})^{-1}$, where $k$ is the Boltzmann constant. The $e^{-\beta E_\mu}$ term is called the Boltzmann factor. The "partition function" Z serves as a normalising constant, ensuring that the probabilities sum to one:

$$Z = \sum_\mu e^{-\beta E_\mu} \tag{3.2}$$

The partition function plays a key role in statistical physics, as it encodes the statistical properties of a system in thermodynamic equilibrium (a more detailed account of its properties is outwith the scope of this thesis).

Equation (3.1), an important result from statistical physics, is the probability distribution commonly referred to as the Boltzmann distribution, and this result will be taken, without proof, as the starting point for the review of basic MC theory in the next section. The characteristic shape of the Boltzmann distribution will become important later on in this chapter (Figure 3.1).

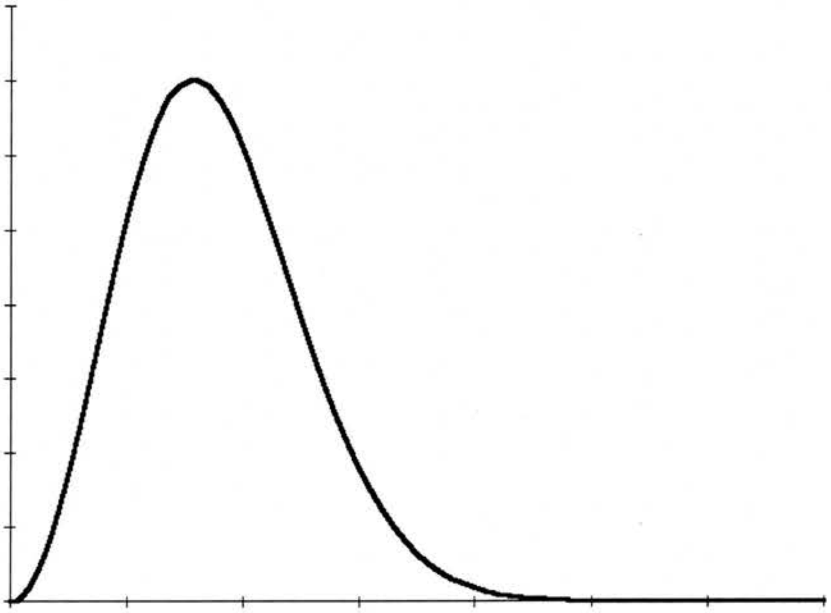


**Figure 3.1: Characteristic shape of the Boltzmann distribution.**

In statistical physics, the properties of a given state are encoded in a set of weights, $w_\mu(t)$, which represent the probability that the system is in state $\mu$ at time $t$. The goal is to calculate some observable macroscopic property, $Q$, which takes the value $Q_\mu$ in state $\mu$. The expectation value of $Q$ at time $t$, $<Q>$, can be expressed as:

$$\langle Q \rangle = \sum_\mu Q_\mu w_\mu(t) \tag{3.3}$$

If one then assumes that the system being simulated has reached an equilibrium state, that is the rate of change of all weights $w_\mu(t)$ will be zero and hence the weights will be invariant as $t \to \infty$, then the equilibrium probabilities, $p_\mu$, can be defined as:

$$p_\mu = \lim_{t \to \infty} w_\mu(t) \tag{3.4}$$

which leads, from Equation (3.3), to:

$$\langle Q \rangle = \sum_\mu Q_\mu p_\mu \tag{3.5}$$

### 3.2.2 Importance Sampling

The ideal method of calculating this expectation value is to average the observable quantity Q over all possible $\mu$ states, with each state weighted by its own Boltzmann probability.

$$\langle Q \rangle = \frac{\sum_{\mu} Q_{\mu} e^{-\beta E_{\mu}}}{\sum_{\mu} e^{-\beta E_{\mu}}} \tag{3.6}$$

Unfortunately, it is only practical to calculate this in systems with relatively small number of possible states; it becomes impractical in almost all real life situations. However, recognising that only a subset of states, M, can be sampled, an estimator of Q, $Q_M$, can be defined from Equation (3.6) as:

$$Q_M = \frac{\sum_{i=1}^{M} Q_{\mu_i} e^{-\beta E_{\mu_i}}}{\sum_{i=1}^{M} e^{-\beta E_{\mu_i}}} \tag{3.7}$$

Here, as M increases, $Q_M$ becomes an increasingly accurate estimate of <Q>. What is required is an efficient method for selecting a subset of M states to sample from in such a fashion that it minimises the error in the expectation value of our desired observable, Q, whilst being attainable in a feasible amount of computational time. It would be possible to sample configuration states completely at random, in an undirected search of configuration space, but this is not an efficient method for reliably approximating <Q>.

To achieve this is, a technique called "importance sampling" is used. The key idea is to direct the sampling of possible states, $\mu$, by recognising that some states will be more important to the sum in Equation (3.7) (as the states with low energy dominate) than others and to therefore bias the sampling towards these more important states, drawing states from a specified probability distribution, $g_\mu$. Rather than choose each state with equal probability, we set the probability of selecting each state as in Equation (3.1):

$$g_\mu = p_\mu = \frac{e^{-\beta E_\mu}}{Z} \qquad \textbf{(3.8)}$$

This choice reduces the complexity of <Q>, cancelling out the Boltzmann factors. $Q_M$ reduces to:

$$Q_M = \frac{1}{M}\sum_{i=1}^{M} Q_{\mu_i} \qquad \textbf{(3.9)}$$

This allows for a more efficient selection of the subset of states, M, to sample from. However, these steps are not sufficient to ensure that the states are selected according to their Boltzmann probabilities. The standard method of achieving this is a Markov process.

### 3.2.3 Markov Processes and Markov Chains

A Markov process, as applied to MC methods, is a mechanism for generating a new state of a system $v$ from an initial state $\mu$ in a non-deterministic fashion. That is, given the same initial state $\mu$, the process will generate a different new state v every time. The probability of generating a new state $v$ from the initial state $\mu$ is defined as the transition probability, $P(\mu \rightarrow v)$. In a Markov process, the transition probabilities should be invariant with time and should only depend on the properties of the states $\mu$ and v, and not on any other state the system has passed through or will pass through. These two conditions ensure that the transition probabilities from state $\mu$ to state v are invariant, no matter what the circumstances.

One further condition is imposed by necessity upon the transition probabilities:

$$\sum_{\nu} P(\mu \to \nu) = 1 \tag{3.10}$$

This condition simply states that the Markov process must be able to generate at least one new state ν given any initial state $\mu$. If this were not so, the simulation could become trapped in a particular state.

Markov processes are used repeatedly in MC simulations to generate a Markov "chain" of states, from initial state $\mu$ to state ν onto state ξ and so on. The Markov process is specifically designed so that, when run for a sufficient length, it will generate new states with probabilities given by the Boltzmann distribution. The process of achieving the Boltzmann distribution is known as "coming to equilibrium", so named as this process is analogous to the process a thermal system undergoes to achieve equilibrium at a given temperature. To ensure that this occurs, however, two further conditions must be imposed on the Markov process.

### 3.2.4 Ergodicity and detailed balance

To make use of MC methods, the system is required to adhere to the principles of "ergodicity" and "detailed balance". The former requires that the probability of moving from a state to any other state be non-zero. That is, every possible state ν must be accessible from any state $\mu$, in a finite number of steps.

The second condition, "detailed balance", requires that the probability of moving from one state, $\mu$, to any other arbitrary state ν, is equal to the probability of reversing the move.

Mathematically, detailed balance can be expressed thus:

$$p_{\mu} P(\mu \to \nu) = p_{\nu} P(\nu \to \mu) \tag{3.11}$$

where $\mu$ and ν represent any two possible configurations. As before, P($\mu$→ν) represents the transition probability from state $\mu$ to state ν, whilst P(ν → $\mu$) denotes the reverse transition probability. Recalling that the intention is to generate a chain of states according to their Boltzmann probabilities, one sets the equilibrium probabilities, $p_\mu$ and $p_\nu$ as in Equation (3.1), the equilibrium occupation probabilities. Equation (3.11) can therefore be rearranged as:

$$\frac{P(\mu \rightarrow \nu)}{P(\nu \rightarrow \mu)} = \frac{p_\mu}{p_\nu} = e^{-\beta(E_\nu - E_\mu)} \tag{3.12}$$

This equation, together with Equation (3.10) and the condition of ergodicity, constitute the constraints on the selection of transition probabilities. Any transition probabilities that satisfy these conditions will guarantee that the equilibrium distribution of states from the Markov process outlined above will be the Boltzmann distribution. It therefore follows that, once the simulation has come to equilibrium, and the generation of states is sufficiently close to the Boltzmann distribution, the simulation will begin to average the desired observable, Q.

### 3.2.5 Acceptance Probability

The transition probability, P($\mu$→ ν), is a product of two constituent probabilities:

$$P(\mu \rightarrow \nu) = g(\mu \rightarrow \nu)A(\mu \rightarrow \nu) \tag{3.13}$$

where g($\mu$→ ν) is known as the "selection probability", the probability of state ν being selected as a possible new configuration from state $\mu$. A(μ→ $\nu$) represents the acceptance probability, the probability of accepting the newly generated state, ν, given initial state μ. Should the potential move be rejected, the simulation will remain in state μ.

The acceptance probability can theoretically range from 0 to 1, although zero is not desirable, as the resultant simulation would never move from the initial state μ. In practical terms, having an acceptance probability that is too low will result in inefficient use of computational resources, as the simulation will not sample many states. Ideally, the chosen acceptance probability (or probabilities) should be as close to 1 as is practical, so the simulation can sample as many states as possible.

### 3.2.6 The Metropolis algorithm

The oldest and still most common practical solution, which satisfies all the conditions mentioned hitherto, is the Metropolis algorithm (Metropolis *et al.*, 1953). Briefly, the acceptance criteria are:

- If $E_\nu$-$E_\mu$ ($\Delta E$) is less than or equal to zero, that is the change in energy is favourable, the new configuration is always accepted.
- If the change in energy is unfavourable, $E_\nu$-$E_\mu$ >0, the new configuration is accepted probabilistically.

The acceptance criteria can be expressed mathematically:

$$A(\mu \to \nu) = \begin{cases} e^{-\beta(\Delta E)} & \text{if } \Delta E > 0 \\ 1 & \text{otherwise} \end{cases} \tag{3.14}$$

When using the Metropolis algorithm, the selection probabilities, $g(\mu \to \nu)$, for all possible states $\nu$ are equal. So $g(\mu \to \nu) = 1/N_\nu$, where $N_\nu$ is the number of possible states we could reach from any state $\mu$.

A more thorough treatment, including the derivations of the results used here, can be found in many texts on MC methods, such as Hammersley and Handscomb (1965), Rubinstein (1981), Binder (1986) and Newman and Barkema (1999), which were used extensively throughout the development of the simulation. The notation used in this thesis is consistent with Newman and Barkema (1999).

## 3.3 Implementation

### 3.3.1 State Representations

A state of the simulation, μ, can be represented by a "state vector", $\underline{s}$, which defines the current configuration:

$$\underline{s} = (s_1, s_2, \ldots s_N) \tag{3.15}$$

where $s_i$ denotes the binding site at which nucleosome $i$ is located, and N is the number of nucleosomes simulated.

In a similar fashion, the configuration can also be represented as a binary site vector, $\underline{b}$:

$$\underline{b} = (b_1, b_2, \ldots b_{nbps}) \tag{3.16}$$

where $b_j$ takes the value 1 if there is a nucleosome currently positioned at site $j$, otherwise it is zero, and index $j$ ranges from 1 to the number of base pairs simulated (nbps). It follows that $\Sigma b_i = \mathrm{N}$.

For example, given a situation where N=2 and nbps =5, where the 2 nucleosomes are located at positioning sites 2 and 4, it follows that $\underline{s}$=[2,4] and $\underline{b}$=[0,1,0,1,0].

The former definition is preferable for use with the coding of the simulation, as it only requires a vector of length N, as opposed to the length of the simulated sequence, which is significantly longer (10640 for BLG). The latter definition, however, is preferable for generating maps of the frequency of occupation of each individual site, as generation of the site occupancy maps is reduced to a superposition of the binary sites vectors. However, the conversion between the two configuration state vector representations is straightforward.

### 3.3.2 Overview of the Simulation

The simulation is initialised by placing the desired number of nucleosomes at specific positions onto the discrete energy lattice. This constitutes the initial state, μ, of the simulation. Once the initial energy is calculated, the nucleosomes are allowed to randomly move within a specified local neighbourhood, subject to the structural constraint, and a new state *v* is produced. Figure 3.2 below illustrates one such variation:

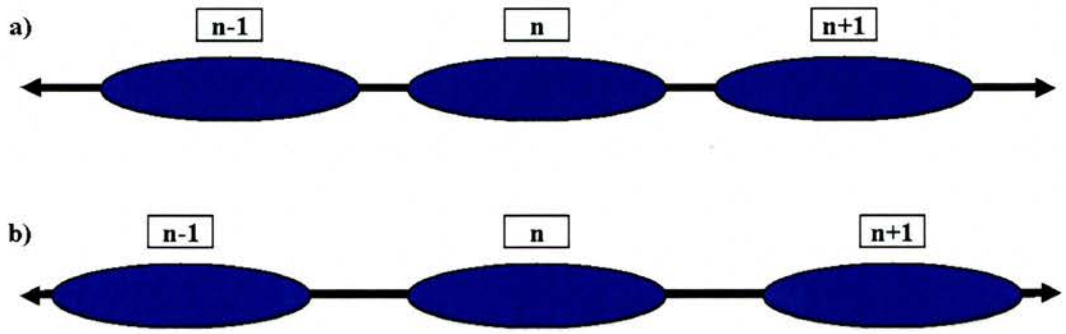


**Figure 3.2: A closer look at two states of the simulation, μ and v, with 3 simulated nucleosomes on ~600bp of DNA within a larger simulation. a) initial state μ b) subsequently generated state v. The first nucleosome, n-1, has been given an opportunity to move, a new position has been generated, and the proposed move to a positioning site several base pairs towards the 5' end has been accepted. Similarly for the third nucleosome, n+1, except it has been offered and accepted a position several bp toward the 3' end. The central nucleosome, n, was either not offered a new position in this sweep or its proposed move was rejected by the Metropolis acceptance criteria.**

The system was sampled at set intervals, known as Monte Carlo Steps (MCS), and the position of each nucleosome was recorded. One MCS is deemed to have past when each nucleosome has been given, on average, one chance to move to a non-constrained site. However, as with nucleosome n in Figure 3.2, the nucleosome is not required to move, even if it is given the opportunity, as the new site may be rejected by the Metropolis selection criteria. The nucleosome density is a primary determinant of the amount of computational time required to complete one MCS for a given length of DNA.

Unless otherwise stated, each simulation was run for 10 million MCS, not including an initial number of MCS required to ensure the simulation had reached equilibrium prior to sampling. Simulations of this length were found to give a good balance between the need for sufficient sampling to obtain reliable results and the computational time required for each simulation.

### 3.3.3 Generation of the Probability Density Function

To obtain the appropriate pdf, it is assumed that the relative affinity for the histone octamer as assessed by ME gives an indication of the proportion of histone octamers bound to a known 146 bp stretch of DNA. This allows one to assume that the probability of each site being occupied by a nucleosome, $p_\mu$, is proportional to the assessed strength of the positioning site. In other words, it is possible to regard the intensity of each ME band, as assessed by phosphorimaging, as proportional to the frequency of occupation of each site.

It should be noted that in this formulation, a site cannot have a negative positioning affinity. Therefore every positioning site that had an assessed intensity value of <1 was set to 1. Such sites can arise from experimental errors, especially from normalisation errors, when the ME positioning affinity map are being constructed. This adjustment is justified by the following argument: within the ME scheme, a negative value for the intensity has no meaning as there cannot be less than 0 octamers bound to a particular positioning site. It should also be noted that, in theory, every natural binding site should be able, in appropriate conditions, to position a nucleosome. That is to say, the positioning of nucleosomes is a stochastic process, where the probability of positioning on any given sequence is always non-zero (Widom, 1998).

### 3.3.4 Energy Landscape

Assuming thermal equilibrium, the pdf can then be converted into a discrete energy landscape (a 1-dimensional lattice) by making use of the Boltzmann probability distribution (Equation (3.1)). The assumptions presented in the last sections allow the direct relation of quantitative band pixel intensity from a ME gel to the free energy associated with positioning on that site:

$$I_i \sim e^{-\beta E_i} \tag{3.17}$$

where $I_i$ is the pixel intensity from ME, and $E_i$ is the positioning energy of site $i$. This can be rearranged for $E_i$, the energy of site $i$:

$$E_i = -kT \ln I_i \tag{3.18}$$

Using Equation (3.18), it is straightforward to generate the energy landscape from the ME datasets. Figure 3.3 below shows the generated energy landscape for the globin dataset.

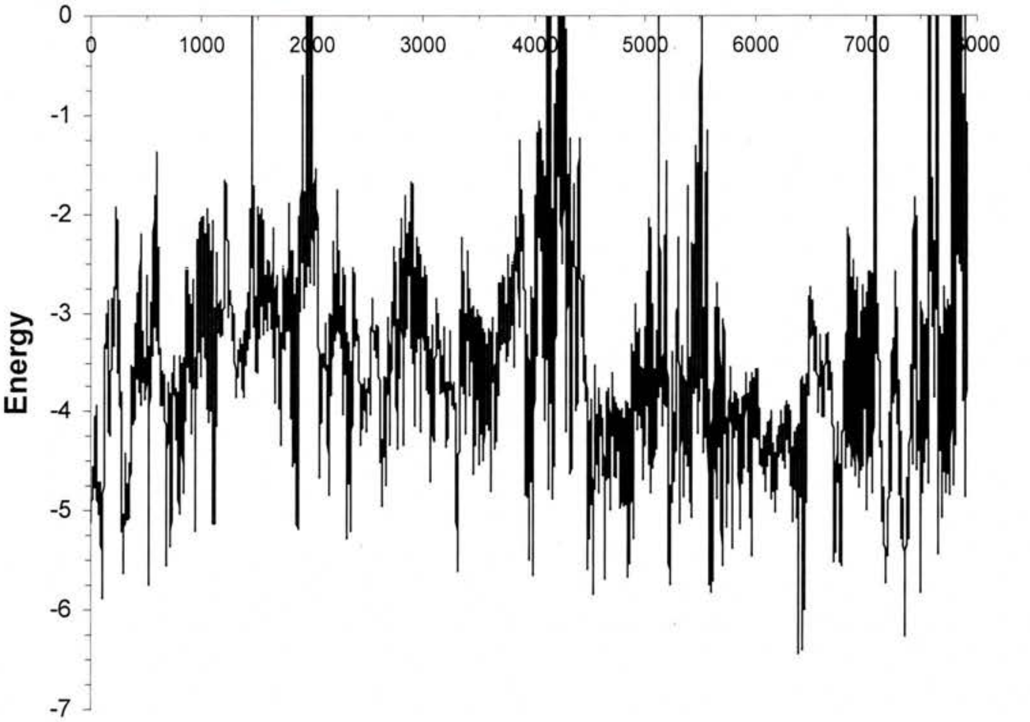


**Figure 3.3: The generated discrete energy landscape used for the globin simulations**

### 3.3.5 Code outline and flow diagram

The simulation was given a set number of nucleosomes (N) to simulate and the minimum nucleosome-nucleosome separation, goes through the following steps:

1. Generate the energy landscape from the ME positioning map.
2. Initialises the system by placing N nucleosomes evenly onto the energy landscape
3. Randomly varies the configuration, subject to the nucleosome separation constraint
4. Accepts or rejects this new configuration according to the Metropolis acceptance criteria.
5. Samples the system per MCS, recording the configuration for analysis.
6. Returns to step 3 until the specified number of MCS is reached (usually 10M MCS).
7. Generates the site occupancy maps.

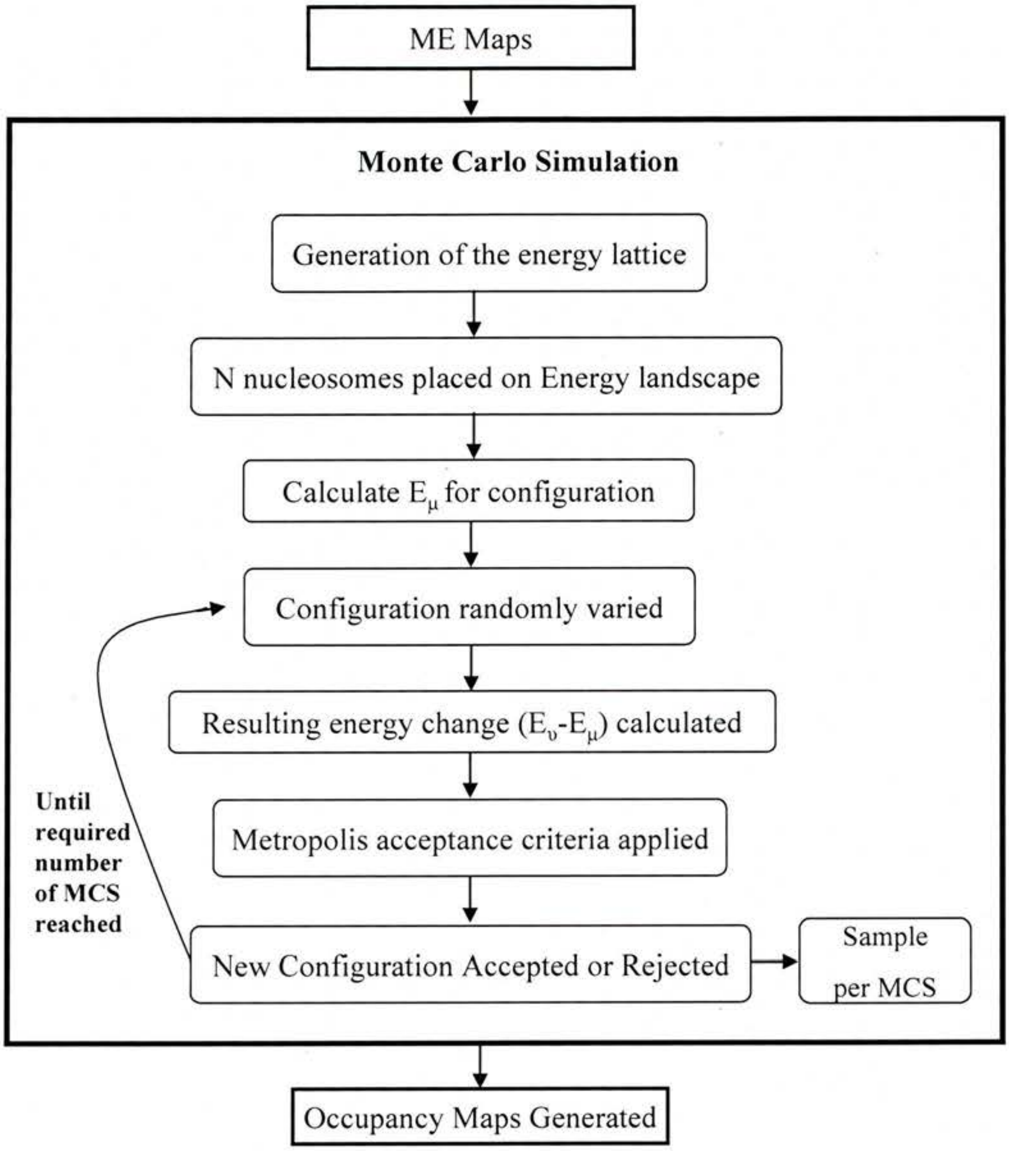

ME Maps
Monte Carlo Simulation
Generation of the energy lattice
N nucleosomes placed on Energy landscape
Calculate $E_\mu$ for configuration
Configuration randomly varied
Resulting energy change ($E_\upsilon$-$E_\mu$) calculated
Metropolis acceptance criteria applied
New Configuration Accepted or Rejected
Sample per MCS
Until required number of MCS reached
Occupancy Maps Generated

**Figure 3.4: Basic constituent steps of the MC simulation.**

The simulation produces results in two forms. Firstly, it produces a site occupancy map, which records the frequency of occupancy for each of the possible nucleosome positioning sites. This is analogous to the ME positioning maps, where the dataset represents the frequency of occupation of every nucleosome positioning site, and as such invites direct comparisons between the two maps. However, to enable a reliable comparison, both maps need to be appropriately normalised. For the ME datasets, this is achieved by dividing the positioning signal for each site by the total positioning signal within the dataset. The resulting normalised ME positioning map therefore sums to 1.

A similar procedure is carried out on the positioning site occupancy maps. Recalling the binary site vector from Equation (3.16), one can define the total occupancy of site *j*, after a simulation of *M* MCS (where *k* represents an individual step), as:

$$f_j = \sum_{k=1}^{M} b_j^{(k)} \tag{3.19}$$

The normalising factor can be calculated by multiplying the total number of MCS of the simulation (*M*) by the number of nucleosomes simulated (*N*). The normalised relative occupancy of site *j*, $O_j$ is therefore:

$$O_j = \frac{f_j}{MN} \tag{3.20}$$

The simulation also stores the state vector, $\underline{s}$, every MCS. As the simulations are usually run for 10M MCS, this produces a very large dataset for each simulation[7]. At present, this data is primarily used for consistency checking, although there is potential for future analysis of this data, such as analysing the correlation between occupancy of specific binding sites and for generating models of chromatin fibres.

[7] A typical BLG simulation, simulating 53 nucleosomes (201 bp/nuc), stores 530 million nucleosome positions. Even heavily compressed using the bzip2 compression algorithm, this requires approximately half a gigabyte of storage.

## 3.4 Program Validation and Control Analyses

To establish that the simulation is working as intended and to gain an appreciation of any potential simulation artefacts, a number of control simulations were carried out. The first two analyses presented are consistency checks to determine if the simulation is working as intended, whereas the latter three controls are necessary to gain insight into the possible influence of sequence ends on the results presented in the next chapter. It is important to pay attention to possible periodic disruptions within the site occupancy maps, as periodicity analyses are used extensively in the following chapter. One can also directly compare site occupancy plots for simulations with varying nucleosome densities (nucleosome repeat lengths), so it is consequently essential to determine the regions of the occupancy map where one can have confidence that any potential simulation artefacts are not having a significant impact on the reliability of this data.

### 3.4.1 Program Validation

#### 3.4.1.1 Ensuring Equilibrium

It is vital that the simulation is run for an initial period, where no sampling takes place, to allow the simulation to "come to equilibrium" such that the Markov process is generating states according to the Boltzmann distribution. This can only be determined empirically (Newman and Barkema, 1999). Figure 3.5 is a histogram of the nucleosome configuration energies (the combined sum of the energy for each simulated nucleosome) of the first 100,000 configurations, one sample per MCS, from a typical simulation.

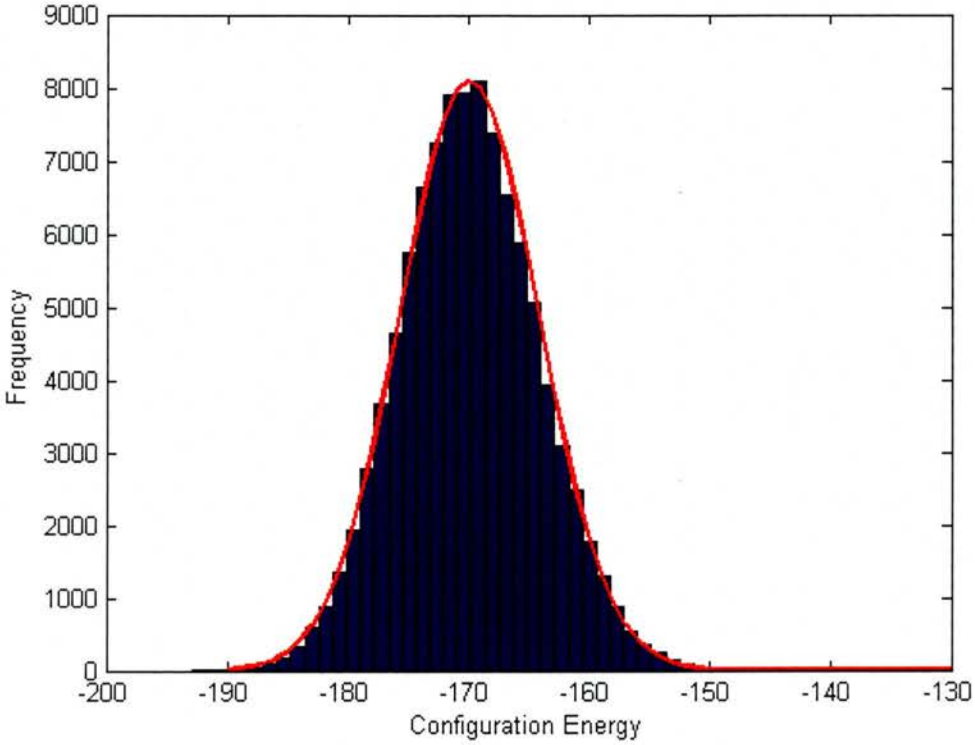


**Figure 3.5: Histogram of the total nucleosome configuration energies for the initial 100,000 MCS for an N=53 BLG (201 bp/nucleosome) simulation. Plotted in red is the best fit to an appropriately parameterised Boltzmann distribution.**

The histogram of total configuration energies in Figure 3.5 is a Boltzmann distribution, as demonstrated by the fitted curve in red (c.f. Figure 3.1). The simulation has therefore come to equilibrium by this point. The initial equilibration period, where no sampling of the state vector takes place, was therefore set to 100,000 MCS for all simulations. Consequently, a 10 million MCS simulation was in practice run for 10.1 MCS, discarding the state vectors sampled in the first 100,000 MCS.

### 3.4.1.2 Low Nucleosome Density

Given the design of the simulation model, one would expect the derived site occupancy map of a simulation run with a nucleosome density low enough that there will be virtually no excluded sequence interactions between the nucleosomes (one nucleosome will not commonly be moved to within the sequence around another nucleosome by the structural constraint) to replicate the pdf used to generate the energy landscape. This is, therefore, an important check of the consistency of the simulation.

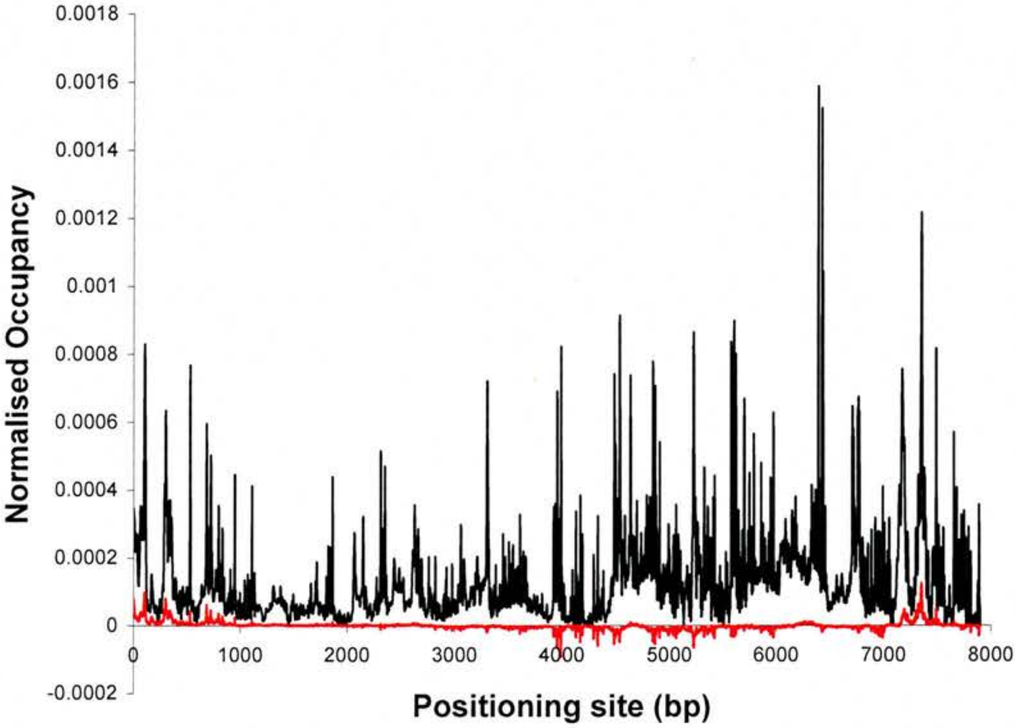


**Figure 3.6: The black line is the binding site occupancy map for a globin simulation with 5 nucleosomes. The red line is a difference plot between the simulation and the equivalent normalised globin ME nucleosome positioning map. Sites where the red line is positive represent sites which are more favoured in the ME map.**

Figure 3.6 is the occupancy map of a globin simulation with 5 nucleosomes, which corresponds to more than 2500 bp/nuc. The density of nucleosomes in this simulation is such that excluded sequence interactions will be minimal. The differences between the normalised ME globin map and the MC simulation are negligible, as represented by the data plotted in red.

### 3.4.2 Characterising "end effects"

One potential issue that requires exploration is the effect of the finite length of the simulated sequence, especially given its 1-dimensional nature. Such effects could have a significant impact on the frequency of occupancy of each binding site, particularly towards the ends of the simulated sequence. It is important, therefore, to determine the impact of such "end effects", to ensure the reliability of any results from the simulation. Several simulations were devised to gain an understanding of end effects, the most important of which are presented here.

#### 3.4.2.1 "Flat" Energy Map

To ascertain the maximum likely impact of such effects, a simulation was run with a flat energy map, 7906 bp in length (the size of the globin dataset). The size of the globin dataset was used, as this is the smaller of the two datasets analysed, and therefore the more likely to suffer from end effects. In this scheme, each nucleosome positioning site is deemed to have equal positioning affinity for the histone octamer.

As expected, given the 1D nature of the simulation model, the end effects are noticeable over approximately 1000 bp on each end of the simulation, causing a ripple effect throughout these regions (Figure 3.7). The affected region corresponds to the first and last 5 simulated nucleosomes, as the simulation was run at ~200 bp/nuc. Crucially, even with this unrealistic energy landscape, at 10M MCS[8] the effect is negligible within the middle of the region being studied. The fact that the occupancy maps for 10M and 100M MCS are effectively indistinguishable is additional strong evidence that the simulation has converged to a stationary occupancy distribution after 10M MCS.

[8] 10 million Monte Carlo Steps

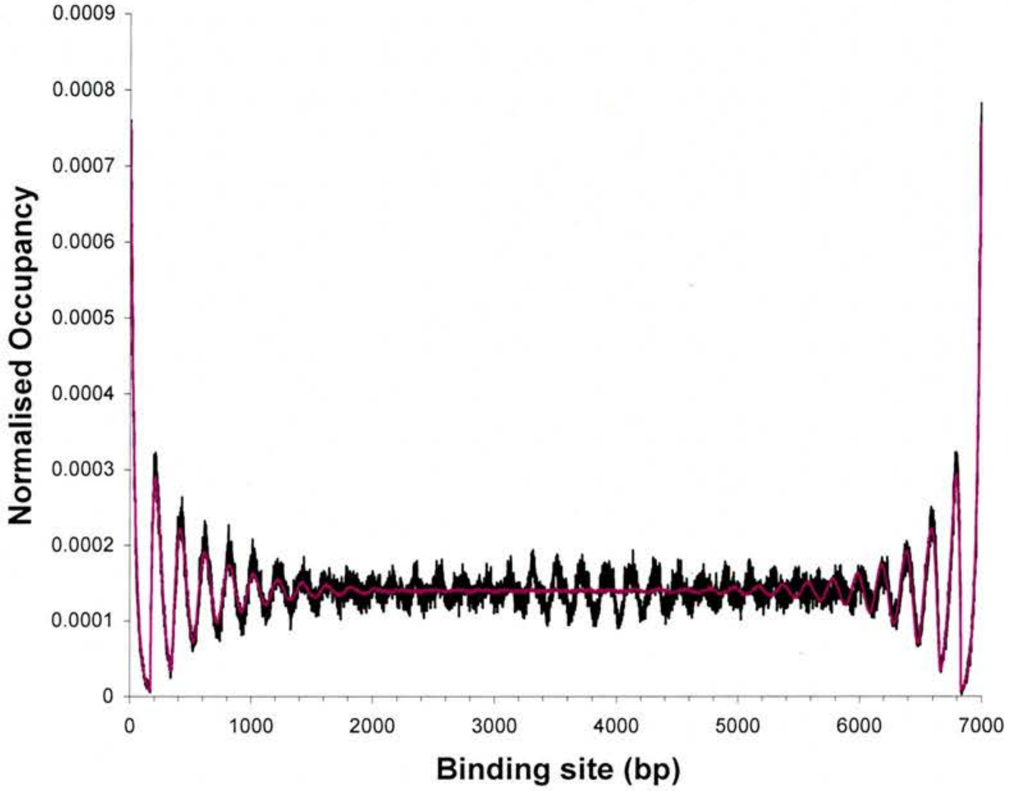


**Figure 3.7: Occupancy maps for 7906bp simulation with 39 nucleosomes using a flat (constant) energy map (203 bp/nucleosome). The simulations were run for 1, 10 and 100 million MCS (black, red and purple respectively). The site occupancy for the 10 and 100 million simulations are virtually indistinguishable.**

### 3.4.2.2 Shuffled Positioning Map

The purpose of this control is similar to the previous one, although here the positioning signals are randomised by shuffling the order of the positioning sites. This is, however, a more realistic control set as it more closely reproduces the simulation conditions used for next two chapters. A shuffled positioning map has the advantage over a random positioning map, where the intensity of each site is completely random, as it retains the first order statistical features of the individual positioning sites.

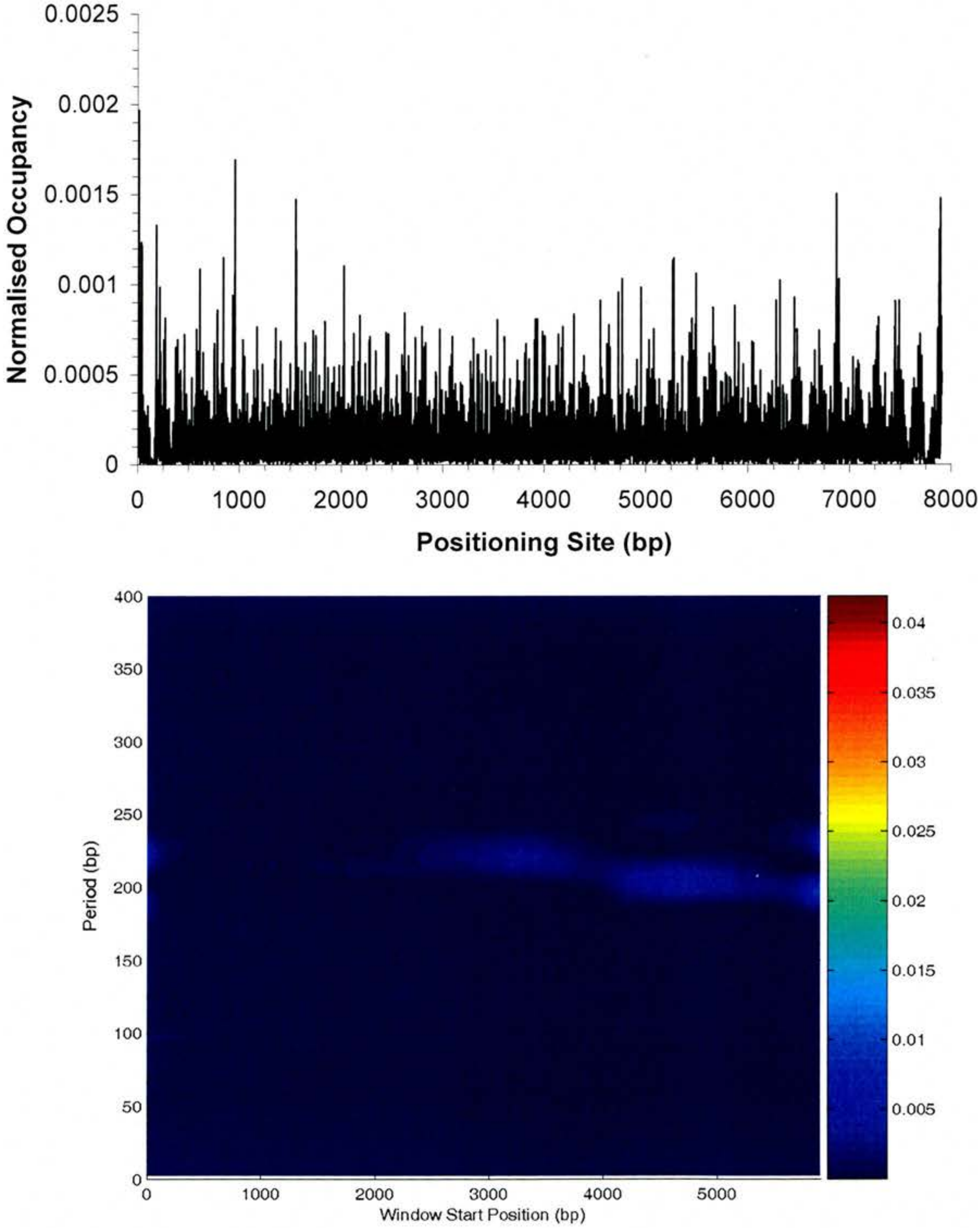


**Figure 3.8: a) MC occupancy map for a shuffled globin map.simulation with 39 nucleosomes b) scanning Fourier analysis of a)**

Scanning PSD analysis of the occupancy map in Figure 3.8 a) demonstrates that, at physiological nucleosomal densities (200 bp/nuc), the simulation does not demonstrate strong periodicities, although some periodicity is noted. It is not until the average nucleosome spacing approaches the excluded sequence constraint that strong periodicities in the occupancy maps are noted (data not shown).

### 3.4.2.3 Padded Positioning Map

To better understand the possible end effects, 1000 bp of randomised positioning data was added onto each of the ends of the energy maps. This artificially extends the ends of the simulation, thereby allowing the examination of the occupancy map without rippling effect demonstrated in Figure 3.7.

There are two points of *prima facie* concern with this type of analysis. Firstly, by creating further positioning sites with random energies, the fidelity of the energy landscape is inevitably compromised. Any such effects should, however, be localised to the crossover region between the globin map and the added sites, and therefore should not significantly impact the end effect in question. The second possible complication with the padded simulation is that the nucleosome density will tend to vary as nucleosomes can enter and leave the padded regions. This is not a significant problems as on average 39 nucleosomes will be located within the non-padded region. Given that, and that changes in nucleosome density are likely to be small, the impact of the changes in nucleosomal density are likely to be inconsequential.

Presented in Figure 3.9 is one such padded simulation for globin.

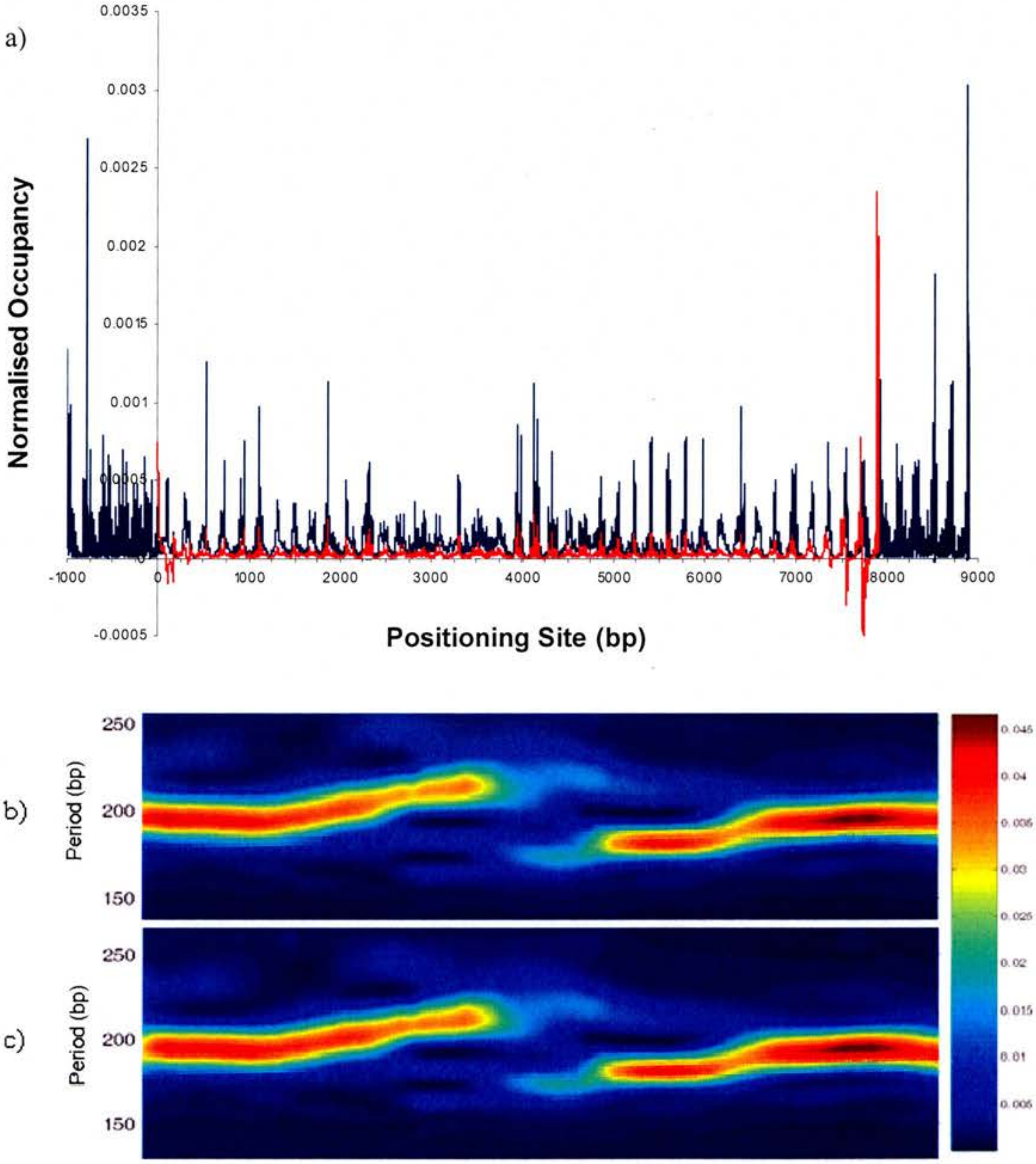


**Figure 3.9: Padded Globin occupancy maps for a ~200bp/nuc simulation. a) the black profile is resulting occupancy map for a ±1000bp padded simulation, whilst the red plots the difference between the padded simulation and a standard unpadded simulation at an equivalent nucleosome density. b) and c) are a side-by-side comparison of the PSD of periods between 150 and 250 bp for the padded and unpadded simulations respectively**

As the difference plot line in Figure 3.9a) demonstrates, with the exception of the end 2-3 nucleosomes, the padding makes no significant difference to the shape of the MC occupancy maps, although it does has some effect on the relative occupancies. The variance seen is likely to be due to fluctuations of nucleosome density over the non-padded energy landscape, which is caused when one or two simulated nucleosomes are located within the padded region and not on proper positioning sites.

This is further confirmed by the negligible changes in the periodicities, as demonstrated by scanning Fourier analyses in Figure 3.4 b) and c).

### 3.4.3 Summary

The validations demonstrate that the simulation is working as intended. Further, the controls indicate that, outwith the edge of the maps, the simulations are not adversely affected by the finite nature of the maps, giving confidence in the validity of the results presented in the next chapter.

CHAPTER 4

# Monte Carlo Simulation Analyses

## 4.1 Overview

The simulation method proposed in the last chapter was principally designed to study the effect of varying histone octamer density on the occupancy of individual positioning sites. As such, the number of nucleosomes initially placed on the energy landscape will be the main parameter for the simulations presented. Other simulation parameters included the structural constraint (the minimum nucleosome dyad-dyad distance), which was set at 168 bp, and the simulation "temperature" (the value of $kT$ used in the generation of the energy landscape in Equation (3.18)) which was kept constant at the temperature at which the ME experiments were undertaken (293K).

The "occupancy maps" are a macroscopic measure of the frequency of occupation for each positioning site, in that they record the total occupancy of each site during the simulation. A higher occupancy indicates a higher prevalence for the simulation to place an octamer on that site. At present, only preliminary analyses and consistency checks have been performed with the sampled nucleosome configurations per MCS. This is due to the prohibitive amount of computational resources required to work with such a large dataset.

The results presented concentrate principally on occupancy maps from the BLG simulations, which is the longer and more accurate of the two ME maps compatible with the simulation[9]. Another consideration is the availability of an *in vivo* positioning map for the region mapped by ME, as the existence of this data provides a unique opportunity for comparison between *in vitro* and *in vivo* behaviour.

[9] The Igf2r, Human and Mouse H19 ME maps are each only ~2000 bp in length, and would therefore suffer significantly from end effects.

Supporting results and observation from the globin simulations are presented where appropriate.

Before presenting the results, it is worth briefly examining the known limitations of the simulation model. ME, for instance, only provides a metric for determining the sequence affinity for the core histone octamer. As such, the dataset takes no account of any effect of the linker histone, H1, which is normally a prerequisite for the formation of higher order chromatin structures such as the 30 nm fibre. The treatment of the excluded sequence interaction between the nucleosomes is also of a rudimentary nature, only specifying that the simulation maintains a minimum nucleosome-nucleosome separation. Recent work in this area by Mergell *et al.* (2004) has made some progress in expanding current understanding of these interactions. Further, the model does not take into account steric effects (hindrance or attraction) due to the physical location of each adjacent core particle and any potential interactions between them. The orientation of neighbouring nucleosomes with respect to each other strongly depends on the length of the linker DNA, as this length will dictate the rotational relationship for adjacent nucleosomes. The simulation also takes no account of necessary distortions of the linker DNA that, with their intrinsic energy cost, will be required to form a higher order structure from a simulated nucleosome configuration. Prior work has demonstrated a relationship between the DNA helical twist and the length of linker DNA, which suggests that linker lengths may be restricted to integer multiples of the helical repeat due to structural requirements necessary to form the 30 nm fibre (Widom, 1992). The simulation model takes no account of the possible effects of linker length quantisation.

## 4.2 Positioning Site Occupancy Maps

### 4.2.1 General Observations on Occupancy Map Features

It is possible to normalise the experimentally derived ME positioning affinity and the *in silico* derived MC positioning site occupancy maps such that the total sum of the affinities/positioning site occupations is equal to unity. So normalised, the occupancy maps for ME and MC can be viewed as the probability that the site will be occupied *in vitro*/*in silico*. This allows for a direct comparison between the MC occupancy maps and the ME maps, and for comparisons between simulations run at different nucleosome densities, independent of the number of MCS the simulation is run for (which obviously affects the absolute number of times a positioning site is occupied). This normalisation will be used throughout the results presented in this section.

BLG simulations, for the full 10M MCS, were run from N= 45 to 63 (236 to 169 bp/nuc) which covers the range of nucleosome spacing observed to be most prevalent *in vivo*. Similarly, globin simulations were run from N= 30 to 47 (264 to 168 bp/nuc).

Figure 4.1 a) is a plot of the BLG positioning site occupancy map for N=53 (201 bp/nuc), a typical nucleosome spacing found in various tissue types, whilst Figure 4.2 a) displays the occupancy map from a N=57 (187 bp/nuc) simulation, the approximate physiological spacing observed in BLG. For indicative purposes, the occupancy map is lined up to *in vivo* nucleosome positioning map (b), and a schematic representation of the gene structure (c) in both figures. The BLG *in vivo* map and the schematic gene representation will be included where appropriate in all further BLG Figures.

For the equivalent globin occupancy maps, there are two nucleosome densities of biological interest: N= 38 (208 bp/nuc) (Figure 4.3 a)), the approximate density when the adult $\beta^{A}$-globin is expressed, and N= 42 (188 bp/nuc) (Figure 4.4 a)) when the embryonic ε-globin gene is expressed. For globin, however, there is not an equivalent *in vivo* nucleosome positioning map. The schematic gene representation, Figure 4.3 b) & 4.4 b) will be included in all further globin related figures in this chapter.

Perhaps the most striking feature of the positioning site occupancy maps is the visibly periodic arrangement of high probability positioning sites, which is a general feature of all maps in the physiological range of nucleosome densities. For the BLG simulations, this periodicity is evident throughout most regions of any map, but is particularly prominent in positioning sites between 1600 and 3600 bp and 6000 and 8200 bp. In terms of the gene structure, the former region corresponds to the flanking region of the BLG promoter whilst the later encompasses exons III, IV, V. Such obvious regular arrangements of high probability sites are considerably less prominent in the original ME BLG map (Figure 1.5)

Close inspection of the region lying between 3600 and 6000 bp in the BLG maps in Figures 4.1 a) and 4.2 a) highlights a tendency towards more periodic behaviour, as the regularity of high occupancy sites is less clear in the lower nucleosome density simulation (Figure 4.1 a)). This is not the case, however, for the region between 7800 and 9200 bp, where the opposite applies: the periodicity in the lower density simulation is clearer than the higher density simulation.

Similar changes in periodic behaviour can be seen in the globin occupancy maps, with Figure 4.4 generally demonstrating more periodic behaviour than Figure 4.3. This is exemplified within the region between 4200 and 6000 bp, where the distribution appears to be significantly more periodic in the higher nucleosome density simulation.

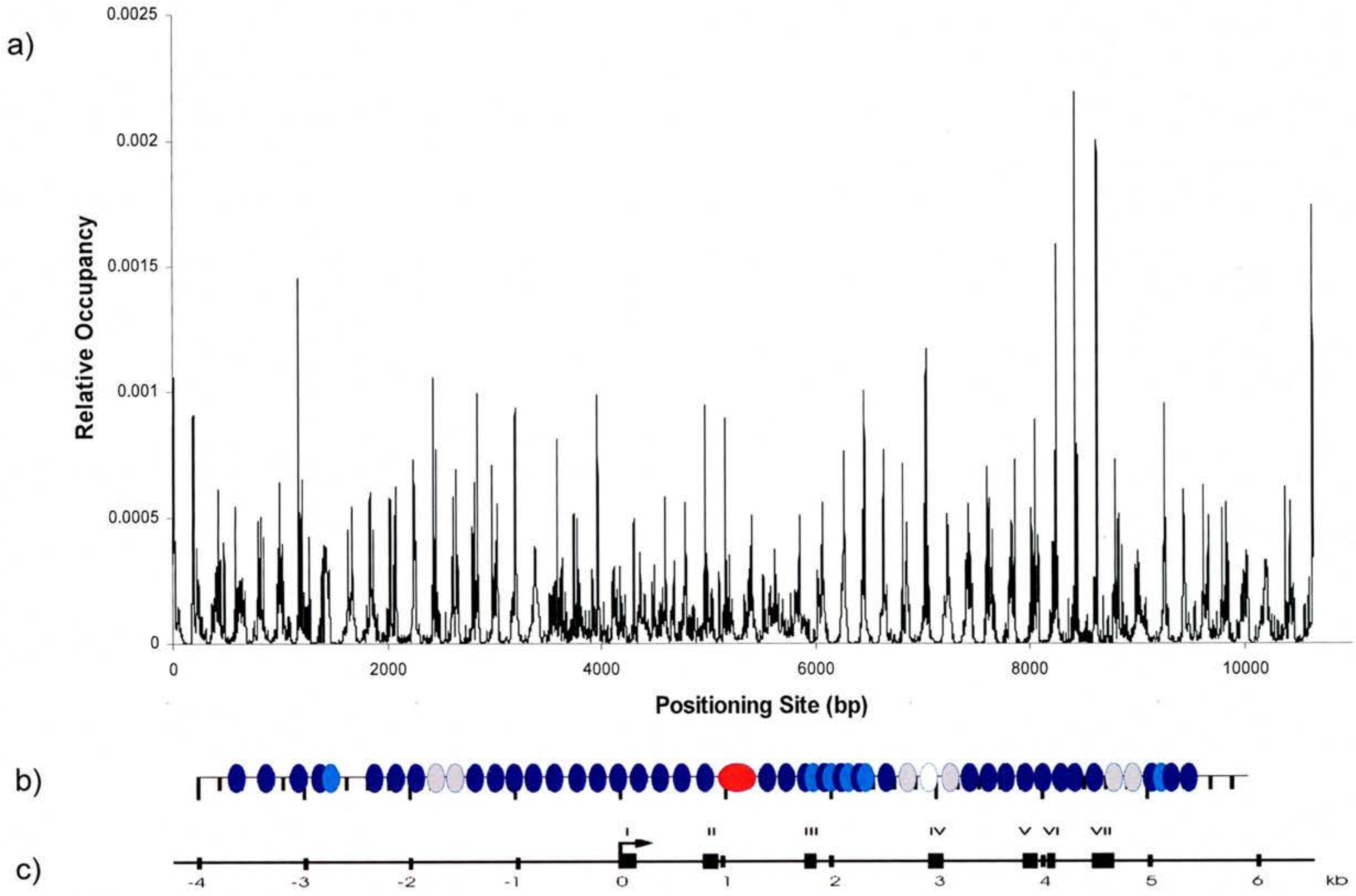


**Figure 4.1: a) Normalised occupancy map for a N=53 (201 bp/nuc) BLG simulation. b) *In vivo* positioning map. c) Location of the promoter and exons 1-7.**

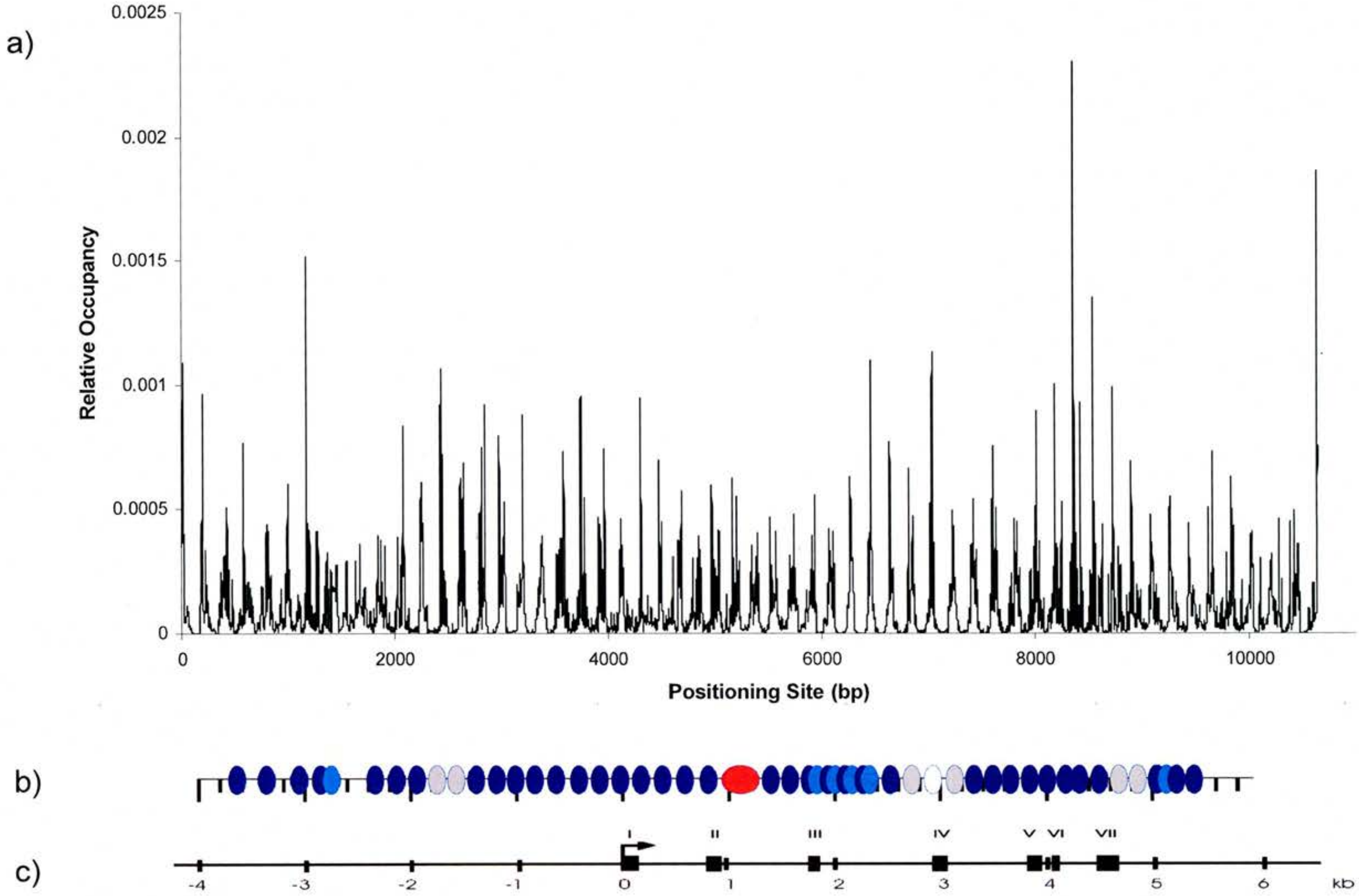


**Figure 4.2: a) Normalised occupancy map for a N=57 (187 bp/nuc) BLG simulation. b) *In vivo* positioning map. c) Location of the promoter and exons 1-7.**

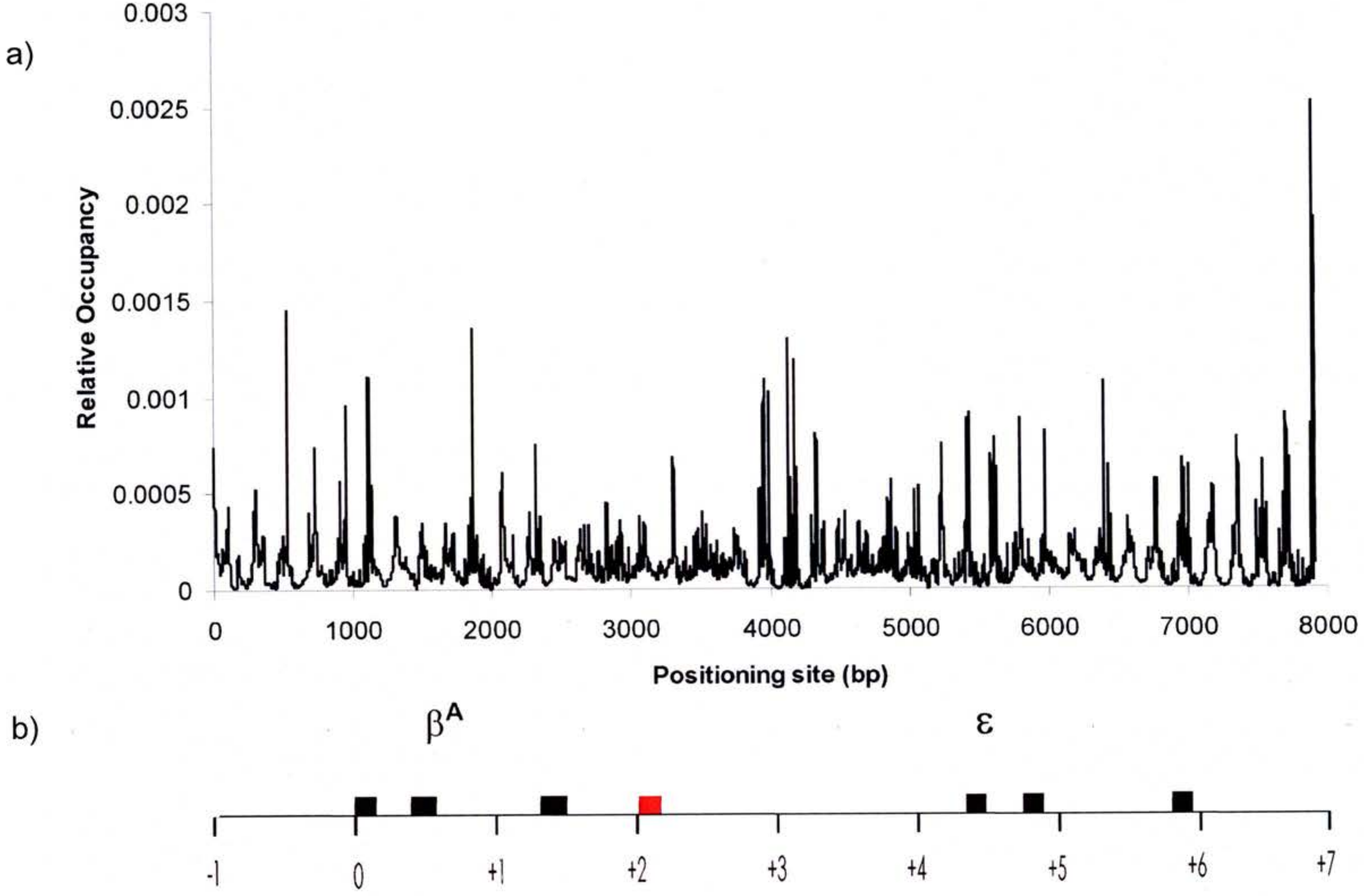


Figure 4.3: a) Normalised occupancy map for a N=38 (201 bp/nuc) globin simulation. b) Schematic gene structure map, indicating locations of the $\beta^A$- and ε-globin gene regions, their exons (black) and enhancer (red).

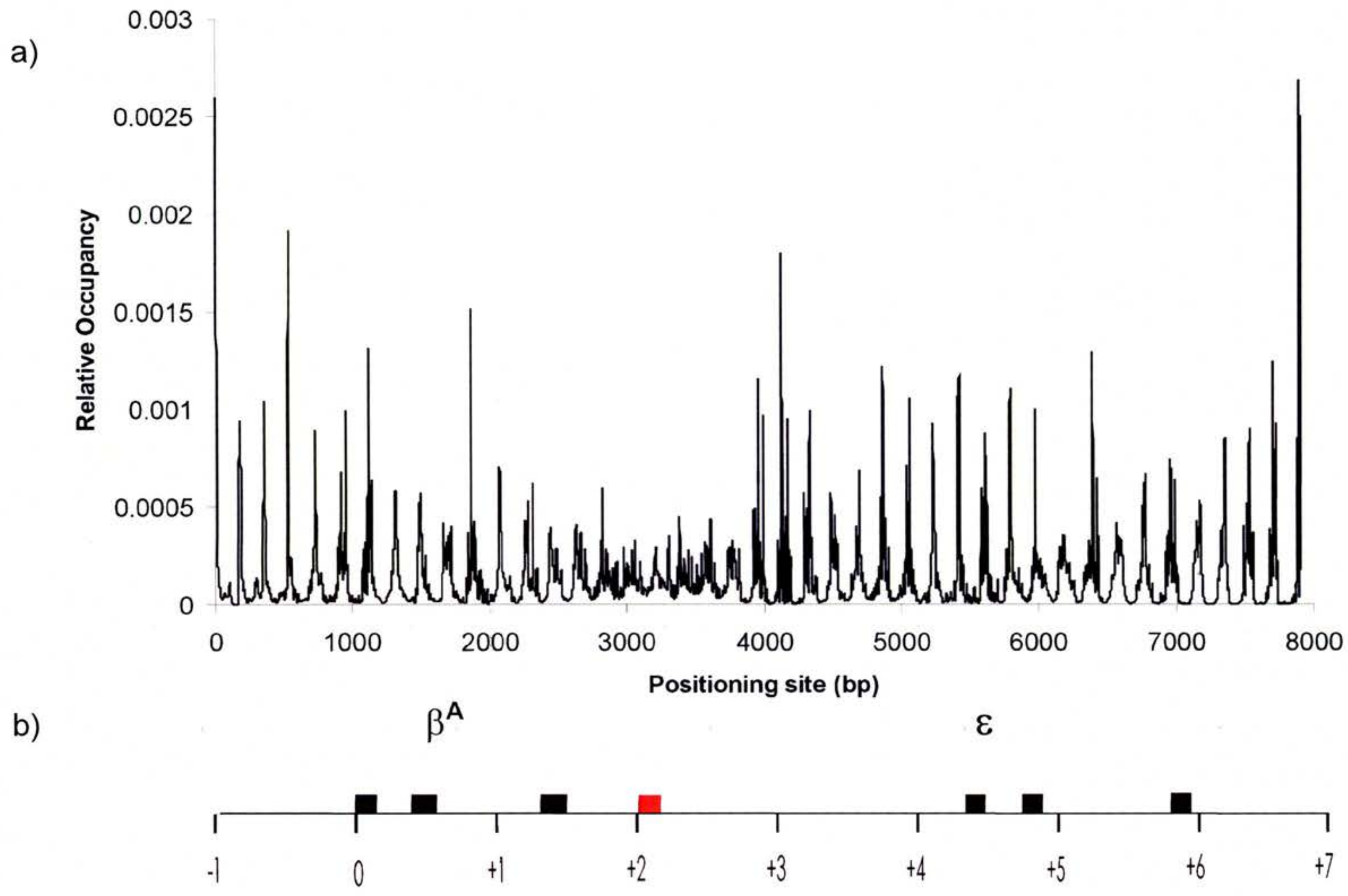


**Figure 4.4: a) Normalised occupancy map for a N=42 (188 bp/nuc) globin simulation. b) Schematic gene structure map, indicating locations of the $\beta^A$- and ε-globin gene regions and their exons (black) and enhancer (red).**

### 4.2.2 Comparison with ME Affinity Maps

In order to more reliably ascertain the scale of the changes in relative occupancy between the ME maps and the simulated positioning site occupancy maps, it is useful to plot the differences between the normalised MC occupancy maps and the normalised BLG ME map. In Figure 4.5, the differences between the BLG ME map and the positioning site occupancies of an N= 57 simulation are presented. In this analysis any positive signal symbolises a site which is bound more often than the original ME data would predicate. That is, the probability of this site being occupied in the simulation was greater than in the ME experiment. Conversely, a negative signal indicates that the simulated occupancy of that site was less than would have been expected based on the inherent affinity of that sequence.

Significant deviations from the ME dataset are noticeable at this nucleosome density. This is, in general, due to the occupancy maps having wider peaks than the ME maps. This effect will tend to happen when the simulation is exploring sites away from a local minimum, oscillating back and forth between the local high affinity site and its neighbours. This is a direct consequence of the simulation being forced by the structural constraint at high nucleosome densities to organise the nucleosomal arrangement into a quasi-regular array. This can result in areas with relatively low positioning affinity being far more regularly occupied than one would expect, as the simulation has little option but to position a nucleosome somewhere in this area. An excellent example of this is the stretch of positioning sites between 6100 and 6300 bp in Figure 4.1 a). This region has a general paucity of ME affinity (Figure1.5) but at this nucleosome density[10], the simulation must place a nucleosome somewhere in this large region.

[10] With N= 57 and nbps = 10640, the average length of linker DNA with the simulation will only be approximately 20 bp.

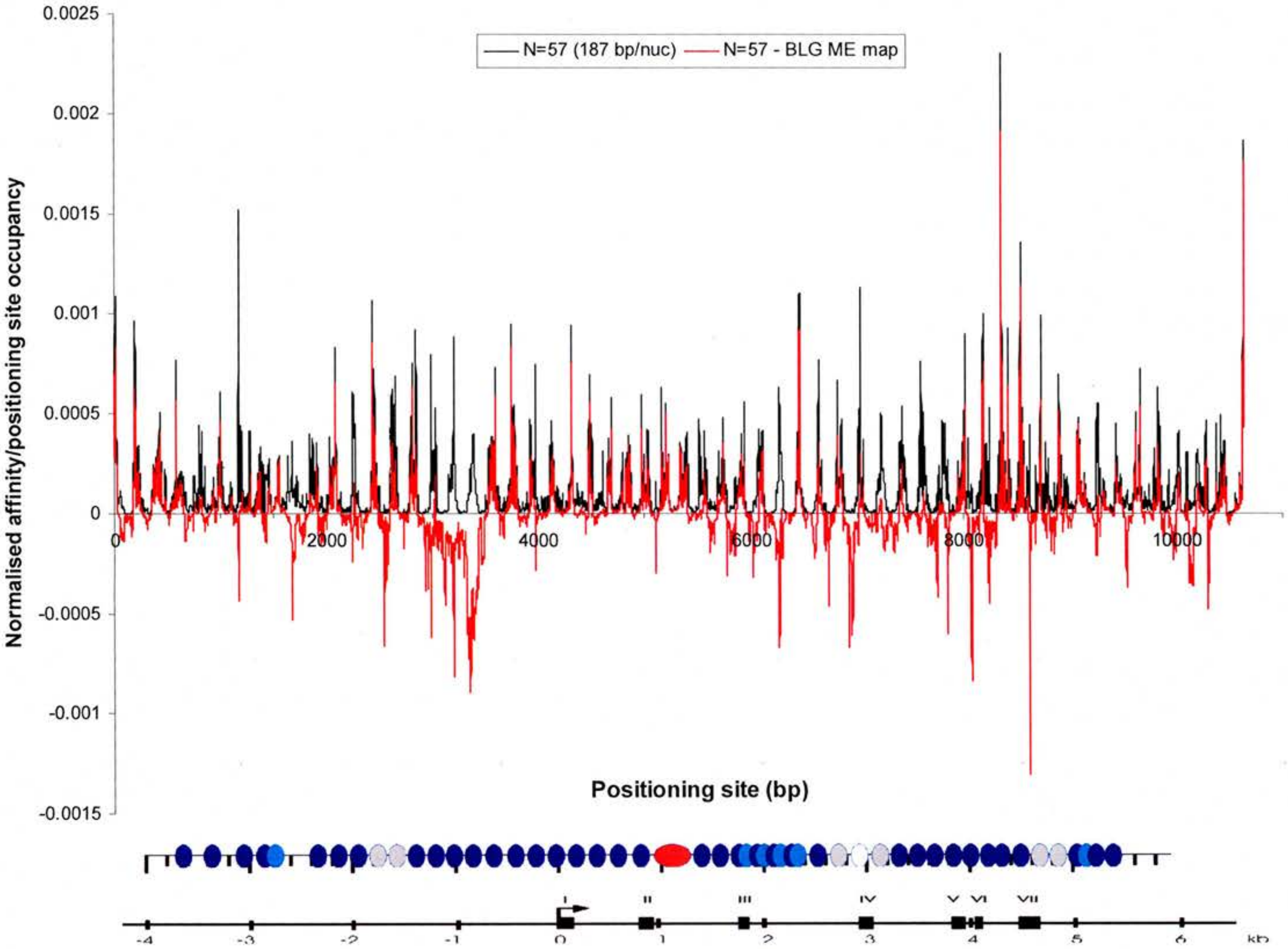


Figure 4.5: Comparison between occupancy map for an N=57 (187 bp/nuc) BLG simulation (plotted in the background in black for comparison) and ME affinity map. In red is the difference plot between the MC and ME normalised occupancies/ affinities, MC minus ME, so a profile positive indicates sites where occupancy in the simulation was more probable than in the ME dataset.

However, this pattern of enhancing the occupancy of positioning sites is not a general phenomenon. Within the region of 3000 to 3500 bp the simulation places nucleosomes three times less often than one would expect based on the sum total of the nucleosome positioning signals within this region (the red profile in Figure 4.5 is highly negative). There are numerous other examples of this type of response scattered throughout the map, particularly between 6000 and 8000 bp in Figure 4.5.

This effect will tend to reduce the range of site occupancies versus ME affinity, as at this nucleosome density and with a minimum nucleosome dyad-dyad separation of 168 bp, the simulation is required to place nucleosomes within regions of relatively low affinity positioning rather than in the regions which are rich in nucleosome positioning "information". This suppresses the relative occupancy within richer regions. The most dramatic effect can be found at 8627 bp, the second highest peak site in the BLG ME map. Here, occupancy is reduced by a factor of 4 in favour of two flanking sites ±~100 bp away (at 8538 and 8727 bp respectively). The positioning site located at 1169 bp towards the 5' edge of the BLG ME map, which is also one of the highest affinity in the dataset as a whole, is still frequently occupied, although the probability of occupancy has been somewhat reduced in comparison to the ME dataset.

The appearance of the ME map is generally one of a few high peaks, with much of the affinity located within clusters of positioning sites (Figure 1.5). However, the MC occupancy maps tend to have a more regular, periodic structure, although there are still sharp peaks evident. The other side of the coin is that some positioning sites have a far higher probability of being occupied than their ME positioning affinity would suggest. Nowhere is this more apparent than at the most occupied site in the simulation, at 8361 bp, which is only a mid range affinity site in the ME map, with a fifth of the affinity of the highest ME affinity site at 1169 bp, and the joint 340th highest affinity site overall. In the occupancy map for the N= 57 simulation, however, this site is easily the most commonly occupied, 50% higher than the next most occupied site[11] located at 1169 bp, which as previously mentioned is the site with the highest positioning affinity in the ME datasets.

Whilst the arrangement of the nucleosomes is still dependent on the energy landscape (the randomised *in vitro* control in chapter 3 demonstrated this), clearly the density of nucleosomes, and structural constraint imposed on the system, are having a significant impact on the general occupancy of positioning sites.

[11] Recall that the occupancy patterns within the first and last 1000 bp of the maps must be treated with some caution, as there is a tendency for occupancy within positioning sites to be influenced by the end effects explored in the controls section of chapter 3.

## 4.3 Effects of Varying Nucleosome Density

### 4.3.1 Introduction

Whilst it is possible to extract interesting general characteristics from close inspection of the occupancy maps on their own, and in comparison to the ME affinity maps, the most interesting properties of the occupancy maps emerge when comparing positioning site occupancies from simulations carried out at different nucleosome densities. There are a number of ways to achieve such comparisons, and this section will explore three distinct methods.

### 4.3.2 Occupancy Map Vertical Alignment Graphs

The most straightforward approach is a direct visual comparison between the total site occupancy maps for simulations at different densities. To accomplish this, occupancy maps from seven different BLG and globin simulations were selected and vertically aligned with respect to the sequence (Figures 4.5 and 4.6 respectively). These plots highlight general behaviour of the occupancy maps as the nucleosome density is varied.

From the BLG alignment graphs (Figure 4.6), it can be seen that one of the most noteworthy changes in site occupancy occurs with the region between 8000 and 8600 bp. Throughout the range of nucleosome densities, there is a marked change in the probability of finding nucleosomes positioned at sites within this region. One can observe significant overall differences in positioning site occupancy from N = 45 to N = 50, a change of average nucleosome spacing from 236 to 213 bp. The overall impression is that the more complex pattern of positioning in the N= 45 density plot is lost in favour of the 3 local strongest sites, which dominate the local landscape by N= 50. Similar behaviour is also apparent between 3700 and 4300 bp.

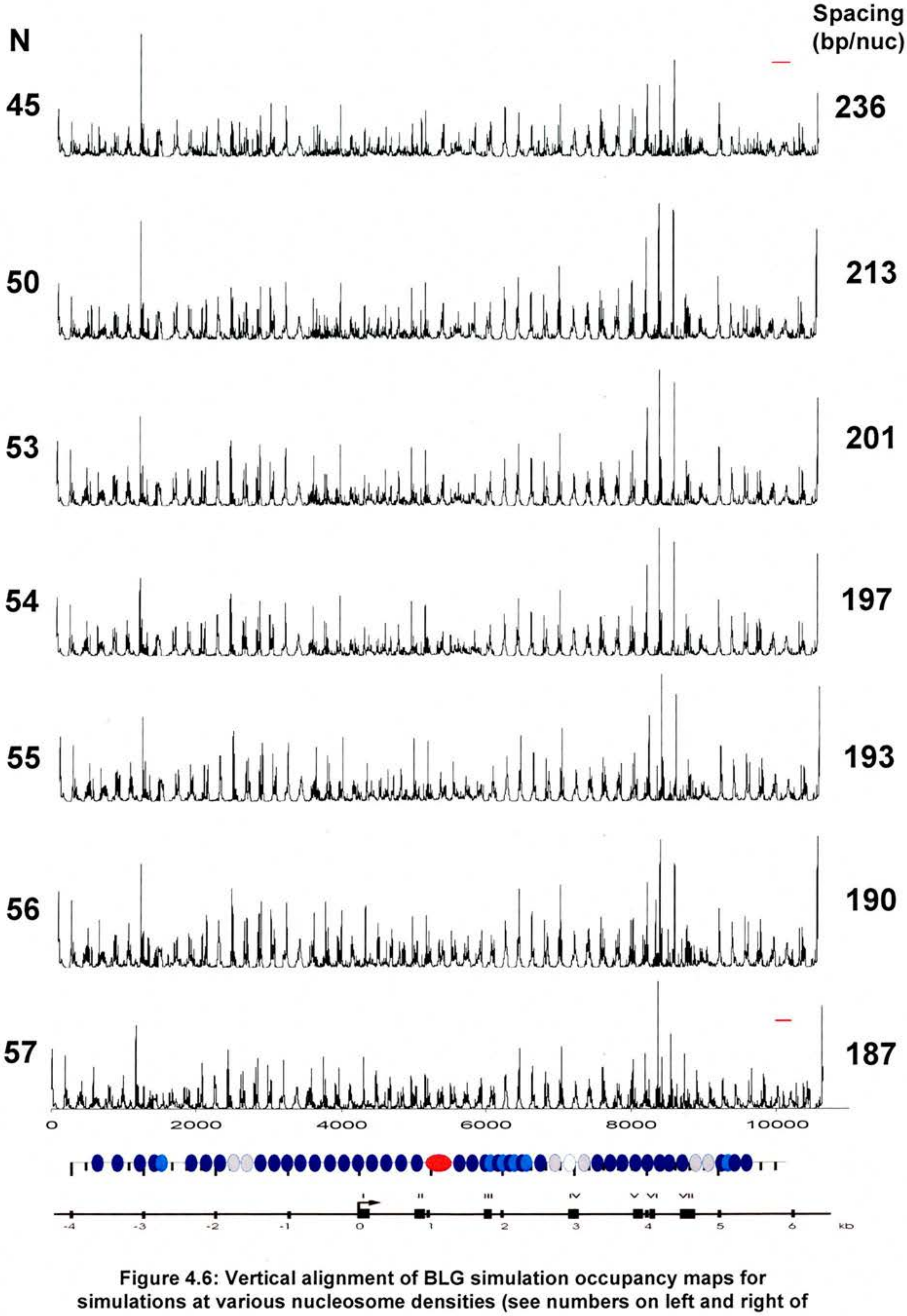


**Figure 4.6: Vertical alignment of BLG simulation occupancy maps for simulations at various nucleosome densities (see numbers on left and right of graphs for details). The red bars in the top and bottom plots indicate the average nucleosome spacing (236 and 187 bp respectively).**

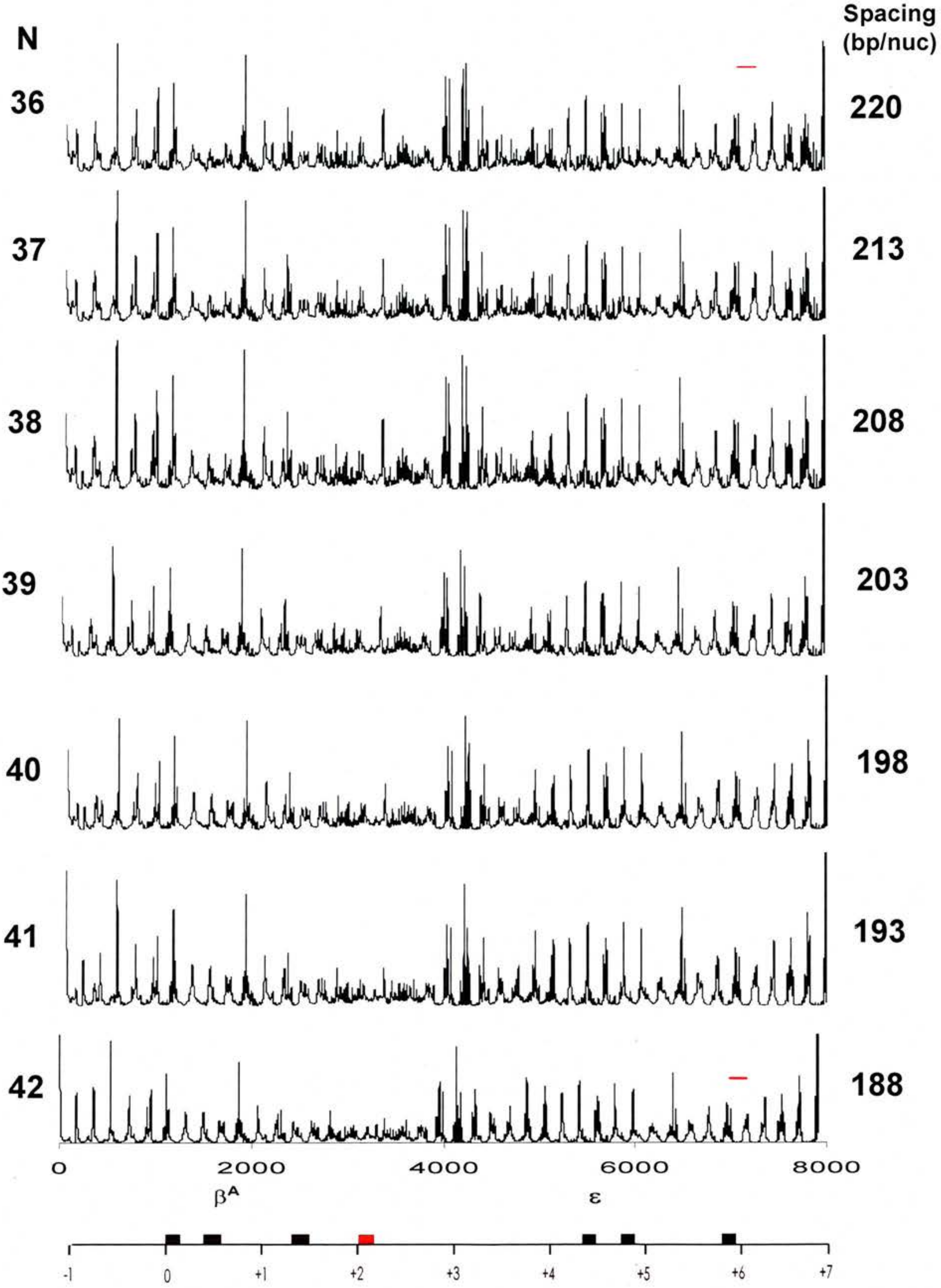


**Figure 4.7: Vertical alignment of globin simulation occupancy maps for simulations at various nucleosome densities (see numbers on left and right of graphs). The red bars in the top and bottom plots indicate the average nucleosome spacing (220 and 188 bp respectively).**

However, there are stretches of positioning sites which maintain the same overall features throughout the range of simulations. The most notable of these areas neighbours the section where some of the most variability is observed. Between 6000 and 8000 bp, the overall pattern of occupancy within the region appears to be relatively invariant to changes in nucleosome density. There are however exceptions: between simulations at 50 and 53 nucleosomes, there does appear to be some changes in positioning probability throughout this region. This exception will be explored in detail in the coming section (Figure 4.8 a)).

Another more localised example is centred at the set of sites surrounding the positioning site at 3383 bp, which appear to be almost completely invariant to changes in nucleosome density. While interesting in itself, it is more curious that this is the approximate shape of the region in the BLG ME map (Figure 1.5), although the relative size of the distribution is much reduced.

At 1169 bp, the highest affinity nucleosome positioning site in the BLG ME map, is another area where some remarkable patterns appear, with the relative occupancy of the site itself fluctuating, but also variations at different nucleosome densities observed in the neighbouring positioning sites

Different features, however, can be observed in the vertically aligned globin maps (Figure 4.7). Although the globin occupancy maps undergo some nucleosome density-dependent variations, the overall tendency appears to indicate smaller differences.

The region of the globin occupancies maps between 2200 and 3800 bp is of significant interest. Unlike the other regions in both the globin and BLG datasets, there seems to be a curious tendency towards lower, less regular arrangements of higher occupancy sites in this region, as if the positioning is smeared across the region rather than being consolidated into a smaller number of sites, as happens in other regions. Quite why such behaviour should be observed in this region is uncertain, although it could perhaps be functionally significant: the $\beta^A$- and ε-globin enhancer lies within this region between the two genes. The enhancer plays a crucial role in determining whether the embryonic or the adult gene is currently active (Choi and Engel, 1988; Foley and Engel, 1992). It is therefore possible that this positioning site occupancy behaviour is relevant to switching the embryonic gene off in favour of the adult gene, as the decrease in nucleosome density which accompanies this change *in vivo* increases the influence of individual positioning sites over the array of nucleosomes *in silico*.

Another area of interest in the globin analysis concerns the strong positioning sites located in a ~600 bp stretch starting at 3900 bp from the 5' end of the map. This region contains three distinct groups, similar in shape to the corresponding groups of positioning affinity in the globin ME map (Figure 1.6). In the globin ME map, the group of sites towards the 5' end is the most dominant of the three, with the other two being of comparable affinity. However, throughout the physiological range of simulated nucleosomes, the central group is always dominant over the other two. Indeed, at N= 42, the 5' and 3' groups appear to be approximately equal, with the central group containing a site almost twice the probability of occupancy *in silico*.

Thus, even using this rather rudimentary type of comparison analysis, one can detect regions of labile nucleosome arrangements which are sensitive to density-dependent remodelling.

### 4.3.3 Occupancy Map Difference Plots

The vertical alignment plots presented in the previous section are helpful for an overview of general properties in changes in positioning site occupancy within the MC simulations. However, it can be useful to analyse more closely the changes between specific nucleosome densities using positioning site difference plots, similar to those used in the comparison of the MC and ME affinity maps in section 4.2.2. There are occupancy maps from 19 BLG and 18 globin simulations (see section 4.1). Difference maps are generated by subtracting the normalised relative occupancy of one simulation from another simulation at a different nucleosome density.

#### 4.3.3.1 BLG Difference Plots

Four such difference plots for BLG are presented, in Figures 4.8 and 4.9. The plots in Figure 4.8 demonstrate the behaviour common when changing the number of simulated nucleosomes by two or three, whereas the two plots in Figure 4.9 demonstrate the differences between two simulations accompanying the addition of only one extra nucleosome.

Firstly, what is perhaps most intriguing is not necessarily the positions of the variations, but the overall nature of the pattern changes in the labile regions. The positioning site occupancy appears to vary in a concerted, periodic fashion. Thus a local increase of one positioning site occupancy is usually accompanied by an increase in occupancy of the flanking sites ±~200 bp away. This phased increase is usually paired with a decrease in occupancy ±~100 bp (although the decrease is not necessarily congruent with the increase). This behaviour, where nucleosomes adopt alternative ~200 bp spaced arrays, is somewhat analogous to the current interpretation of overlapping nucleosome positioning sites found *in vivo* and represented by the alternating dark and light blue ovals on the *in vivo* positioning map used throughout this thesis (see section 1.4 and Figure 1.5 for details).

The difference plot of Δ(N=53, N=50), 201-213 bp/nuc (Figure 4.8) displays sizable variations in the region just before the BLG promoter (3800 ±100 bp) and in the intron between exons III and IV (6700 ±300 bp). Minor variations are noticeable in other regions, in particular in and around the cluster of exons V, VI and VII.

Figure 4.8 b) charts the differences between simulations with 56 and 54 nucleosomes (Δ(N=56, N=54), 190-197 bp/nuc). The occupancy sites show significant variation within a broad affected region: the 3000 bp immediately upstream and within exons I-III (from 3000 to 6000 bp), with the 200 bp phased pattern of increase/decrease discussed earlier being particularly noticable. Interesting, however, there is no appreciable difference in occupancy within the other four exons (6000-9000 bp). Thus the effect seems to be localised around the promoter and the first three exons.

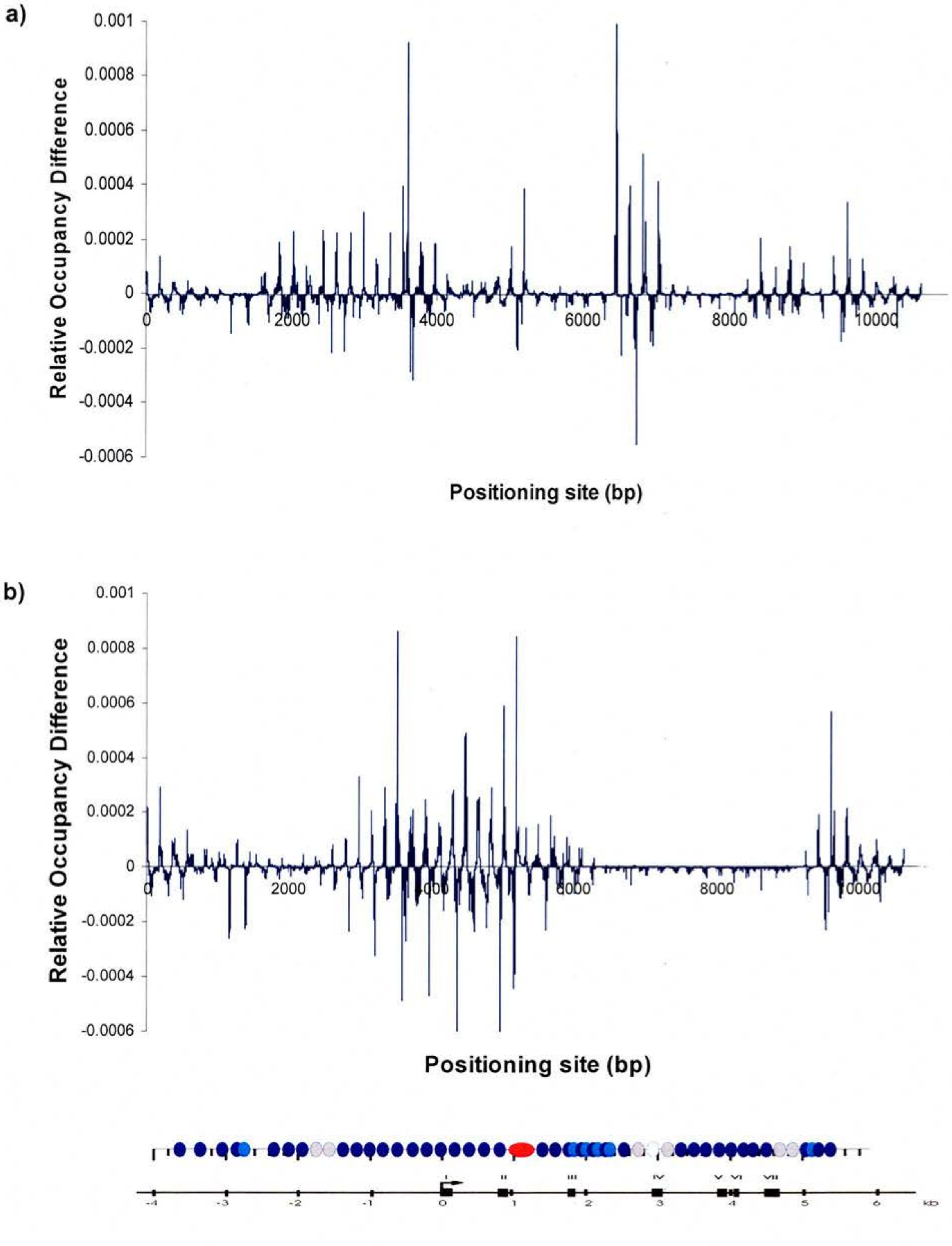


**Figure 4.8: Relative occupancy difference plots for BLG simulations.**
**a) Δ(N=53, N=50) (201-213 bp/nuc). b) Δ(N=56, N=54), 190-197 bp/nuc.**

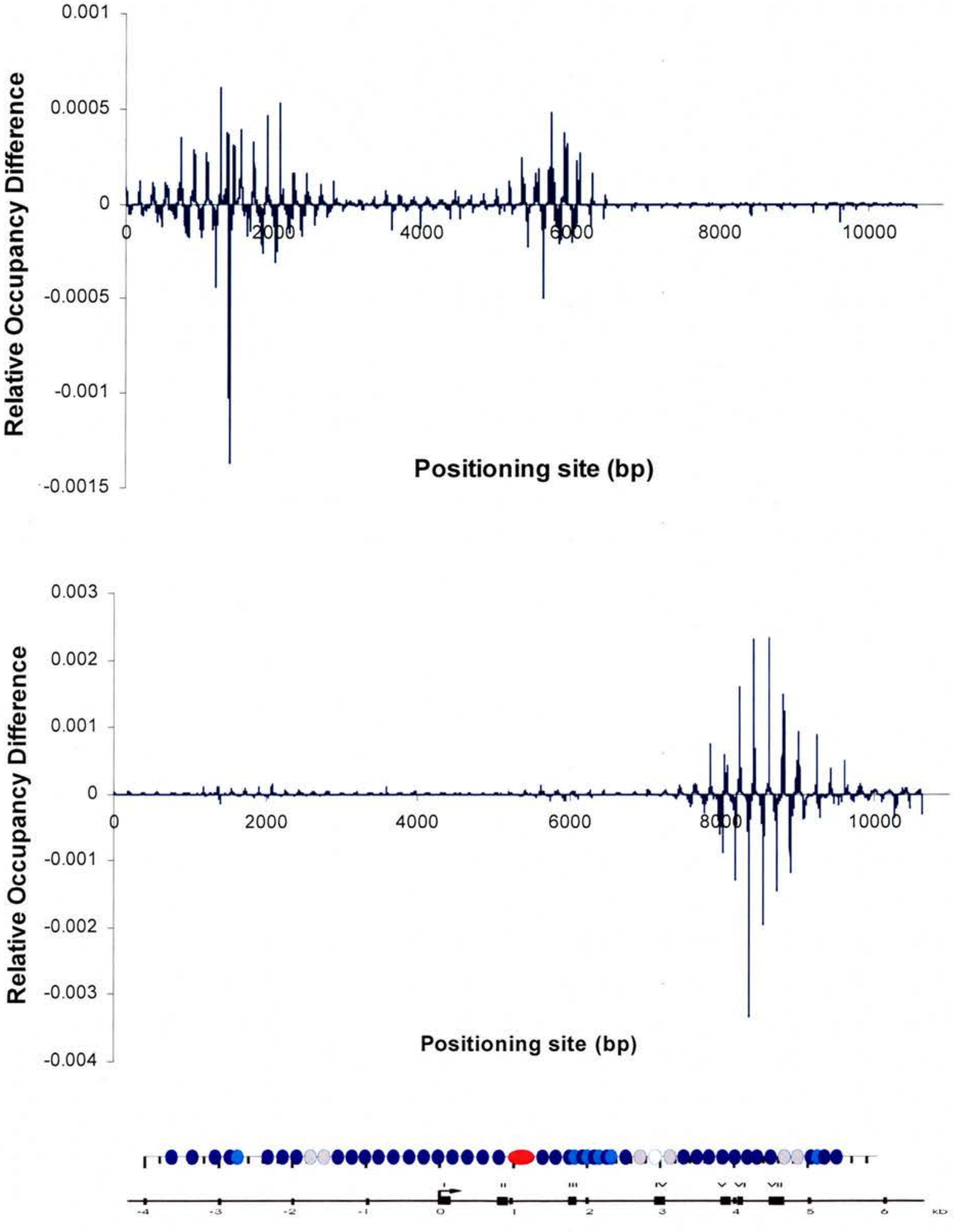


**Figure 4.9: Relative occupancy difference plots for BLG simulations. a) Δ(N=56, N=55) (190-193 bp/nuc), b) Δ(N=57, N=56) (190-187 bp/nuc).**

Thus far, the discussion has been based around density differences equivalent to accommodating two or three extra nucleosomes over the 10.7 kbp BLG gene region. Figure 4.9 a) and b) consider the smallest change currently possible within the simulation design, an alteration in density of only one nucleosome (equivalent to a change of approximately 3 bp in the average repeat length). The two cases presented compare N= 56 and 55 (190-193 bp/nuc) and N= 57 and 56 (187-190 bp/nuc).

In Figure 4.9 a) the analyses again highlight the significant differences occurring within that region surrounding exon III, although there is some smaller variability noted within exons I, II, and the BLG promoter region. These changes are thus focussed around the region containing the three unusual *in vivo* chromatin structures, as denoted by the red circle on the BLG *in vivo* map.

In Figure 4.9 b) the analysis once more demonstrates that the addition of a single nucleosome on the 10.7 kbp stretch of DNA covering the BLG gene region gives rise to a significant deviation in positioning site occupancy. Moreover some of the variations recorded are three times the magnitude of the largest differences discussed previously in Figures 4.8 a) and b) and Figure 4.9 a) (note change in y-axis scale in Figure 4.9 b)). In other words, there are significantly larger differences between the simulations than presented previously, or indeed observed thus far in any of the difference plots.

What makes this analysis even more intriguing is that this distinctive behaviour should be observed once one reaches the physiological density of nucleosomes observed in BLG. Further, there is virtually no other change in positioning site occupancy throughout the rest of the gene region. It would seem that changes required in positioning site occupancy to accommodate the extra nucleosome are focussed within this relatively short region.

#### 4.3.3.2 Globin Difference Plots

The globin simulations appear to exhibit different behaviour to the BLG simulations. The difference plots from globin have been graphed in a slightly different manner. In Figure 4.10, two profiles are shown, displaying the difference between the N= 40 (198 bp/nuc) simulation with either N= 38 (208 bp/nuc) or N= 42 (187 bp/nuc). The differences in positioning site occupancy are relatively small, especially in comparison to those observed in Figure 4.9 b). Another point of note is that the differences do not seem to be as localised as was the case for BLG. The feature of significant localised changes in occupancy that occurs in the BLG simulation in specific regions does not seem to occur: changes in nucleosome density appear to result in changes of positioning site occupancy across the maps.

As in the BLG difference plots, the changes in site occupancy are periodic, with the same characteristic 200 bp out of phase regular arrays. A further observation is that the variations in site occupancy between N= 40 and N=38 and between N= 40 and N=42 are often out of phase with each other. Examples of this behaviour can be found between 4000 and 5000 bp, where the positive red peaks are matched by negative black peaks and vice versa.

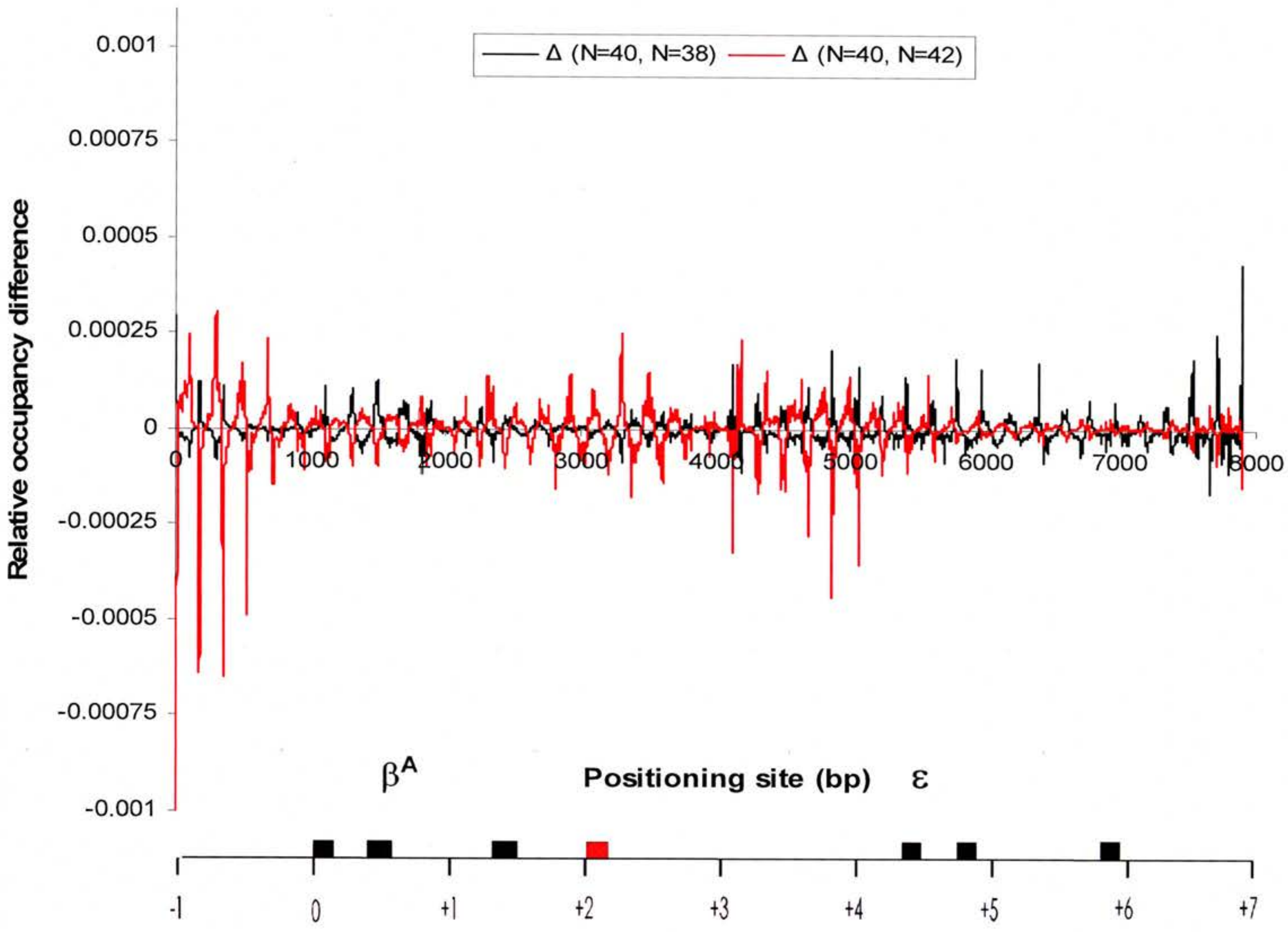


**Figure 4.10: Relative occupancy difference maps for the globin simulations. Red profile is the difference between N=40 and N=38 (198-208 bp/nuc), whilst the black profile is the difference between N=40 and 42 (198-188 bp/nuc).**

### 4.3.4 Occupancy Map Periodicity Analysis

#### 4.3.4.1 Motivation

The final method presented for extracting features to compare the occupancy maps is the scanning PSD analysis used previously. This is perhaps the most important of the analysis techniques, as the precise disposition of nucleosomes along the DNA, and particularly the regularity and periodic nature of strong positioning sites, may have some relevance to the structure and function of the higher order chromatin fibre. A study of the periodic arrangement of positioning site occupancy is therefore potentially indicative of functionally important features.

#### 4.3.4.2 Results and Discussion

Scanning PSD analyses from the BLG and globin simulations are presented in two forms. First, scanning PSDs for the N= 53 and 57 (201 and 187 bp/nuc) BLG simulations, for periods between 0 and 400, are presented in Figure 4.11 a) and b) respectively.

Given the design of the simulation model, coupled with the intrinsic periodic arrangement of strong positioning sites around both the physiological nucleosome density and the minimum nucleosome-nucleosome separation, it is expected that the occupancy maps would demonstrate enhanced periodic signals in comparison to those from the ME datasets. In the limit that nucleosome spacing is such that there is no linker DNA between the nucleosomes (nucleosome density ≈ 168 bp/nuc), the simulated nucleosomes will not be able to move and therefore the occupancy maps will be highly periodic. Therefore, locations of disruptions in the periodic arrangement of high occupancy sites are of particular interest.

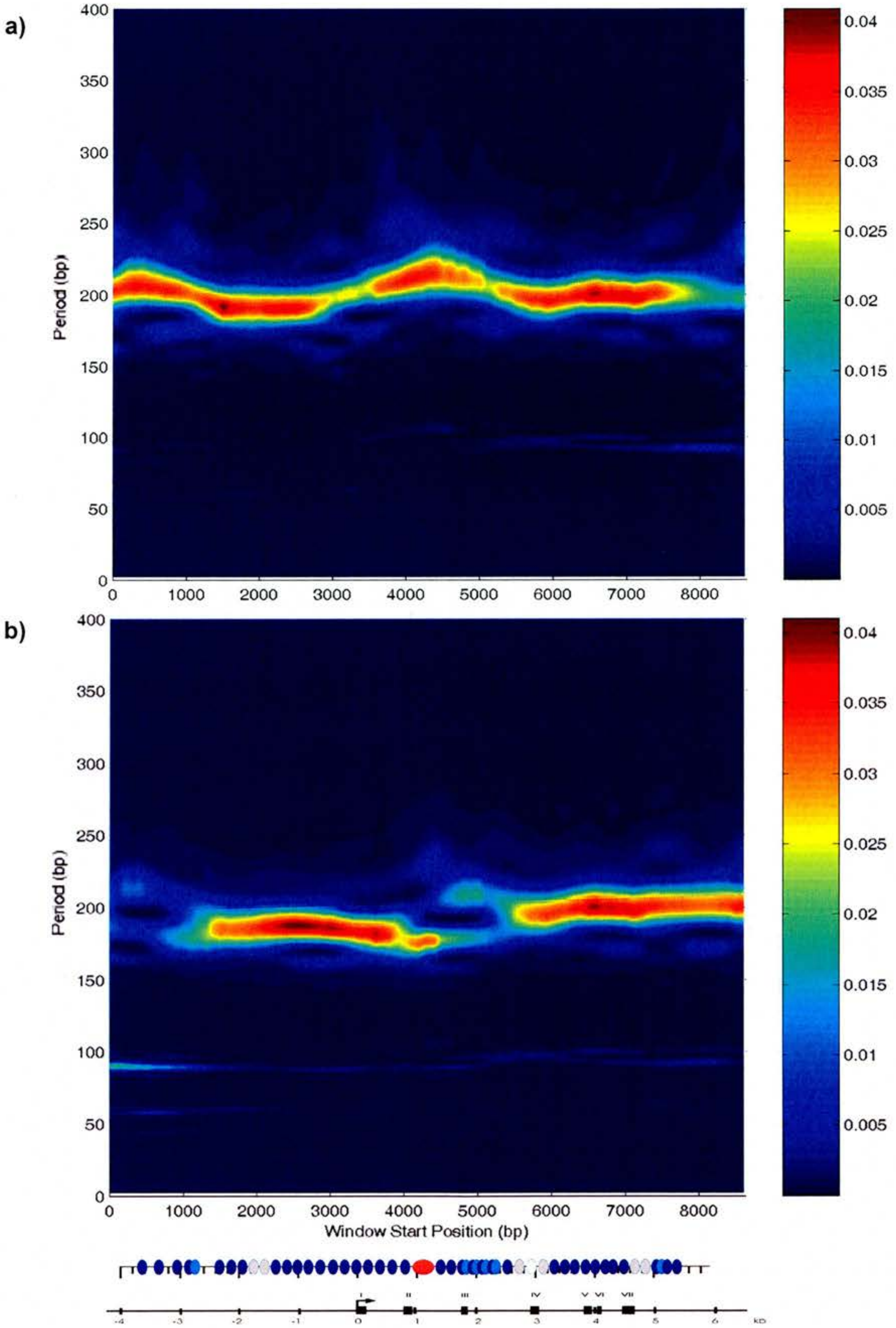


**Figure 4.11: PSD Analysis of BLG MC occupancy maps for a) N=53 (201 bp/nuc) and b) N=57 (187 bp/nuc) simulation.**

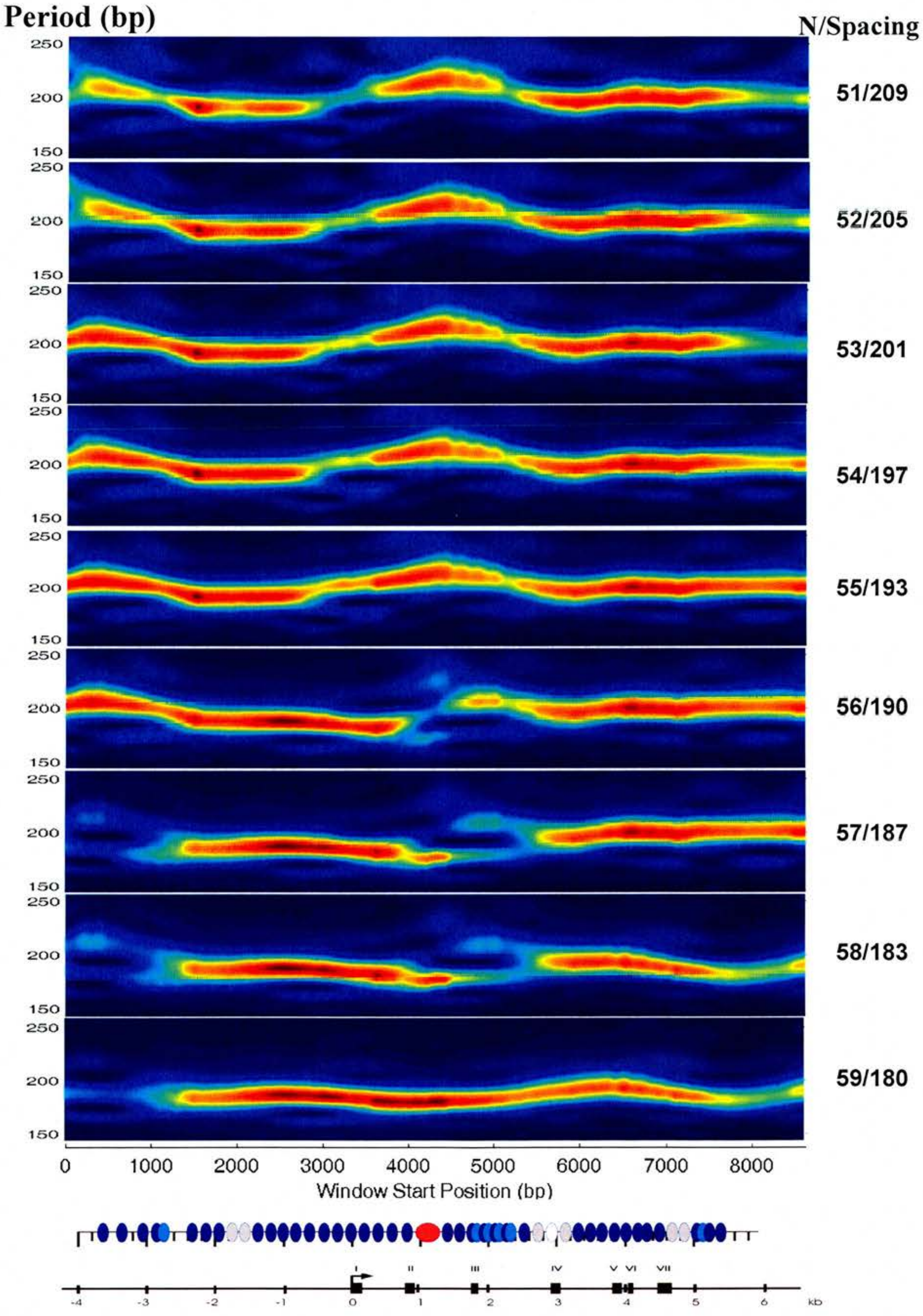


Figure 4.12: Multiple Scanning BLG PSD plots, from with N varying in integer steps from N=51 to 60. Only displayed is the 150-250 bp period range.

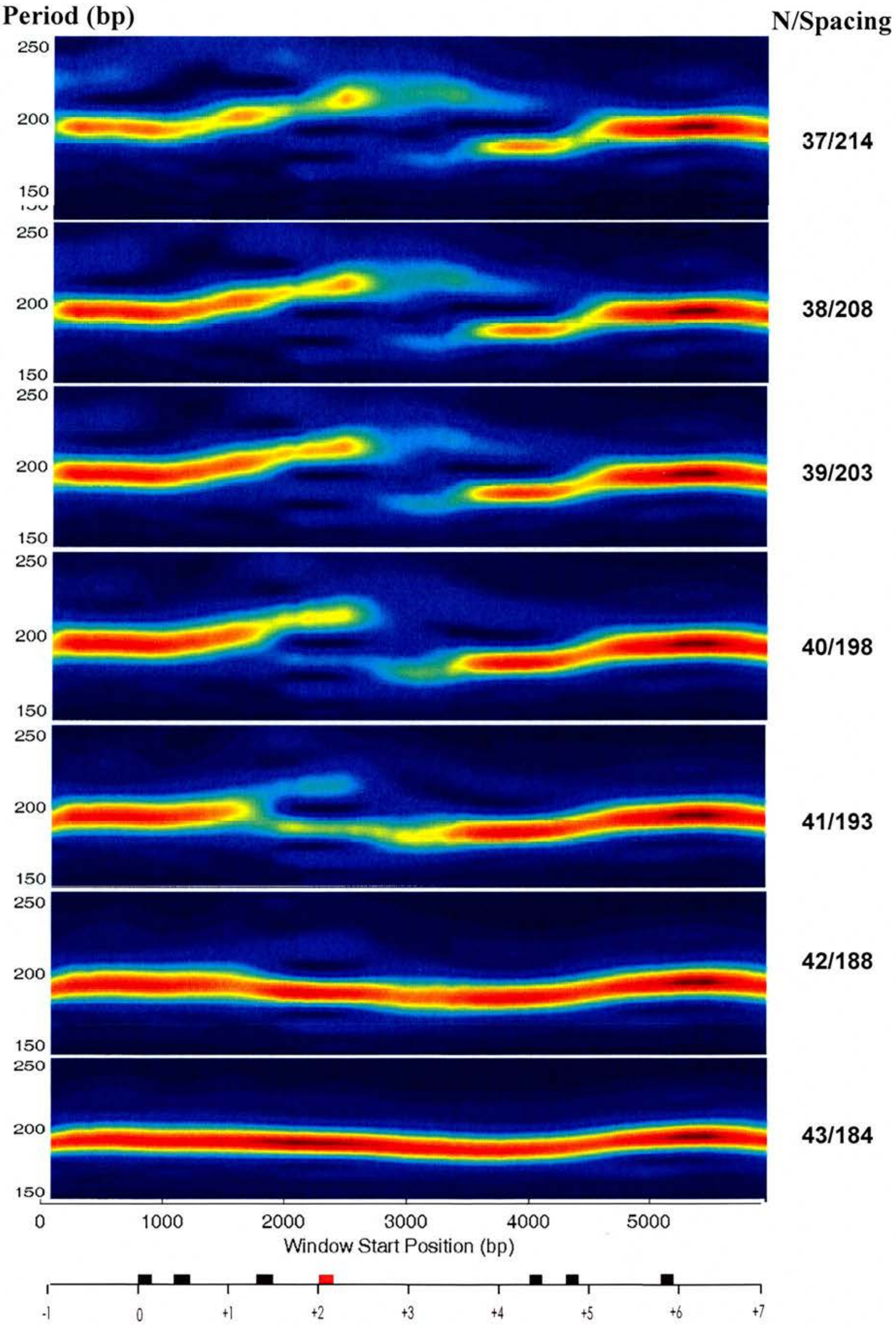


**Figure 4.13: Multiple Scanning globin PSD plots, from with N varying in integer steps from N=37 to 43. Only displayed is the 150-250 bp period range.**

The plots in Figure 4.11 illustrate two major features of the occupancy maps. Firstly, unlike the scanning PSDs for the ME affinity maps (Figure 2.2 a) and b)), all the periodic signals appear to be located in a tight periodic pattern, within a narrow range at around 200 ±50 bp, bar a small contribution at ~90 bp. Secondly, there is a marked change in the period behaviour within the dataset between the two simulations. The difference in the number of simulated nucleosomes makes a significant difference to the periodic behaviour evident in the maps.

In Figure 4.11 a), there is an approximately sinusoidal pattern to the periodicities. For windows starting within the first 1500 bp from the 5' end, the dominant periodicity spreads out to encompass a period of ~220 bp, before returning to narrow distribution around 1000 bp on the x-axis. There is then a shift towards a dominant periodicity of 195 bp for the next 2000 bp. The breadth of the distribution throughout this region appears to be stable. Towards the middle of the mapped region, from 3500 to 5000 bp, the dominant period once more rises above 200 bp, this time reaching a peak around a period of 225 bp at 4500 bp before once more retuning to 200 bp by 5000 bp. The dominant period then returns to just below 200 bp for the remainder of the map (excluding the extreme end of the map).

This behaviour is distinct from that seen in Figure 4.11 b). Here, there is no single dominant period for the first 1200 bp, with 4 significant contributions at 90, 160, 190 and 210 bp. After 1200 bp, the pattern of a single dominant period returns (~187 bp, the simulation density), and the width of the PSDs within the next 2800 bp appears constant. However, for windows starting at 4000 bp, and extending for a further 1200 bp, a disruption occurs in the dominant period: it splits into two components, one at ~180 bp, one at 210. Excepting a short 300 bp stretch to begin with, each of these components appears with equal power. These two components appear to merge by 5500 bp, and there is a gradual rise in the dominant period until 6200 bp, where the principal period is now around 200 bp where it remains throughout the rest of the map.

Most noticeable from these two is the behaviour between 4000 and 5500 bp in both graphs. In the low density simulation in 4.11 a), the dominant periodicity appears to climb to a global high for the dataset within this region. It is worth noting that the hypersensitive site (the red oval on the *in vivo map*) is located ~5000 bp from the 5' end of the map and the region of overlapping *in vivo* positions (the array of 4 dark and light blue ovals on the *in vivo* map) is located ~6000 – 6500 bp from the 5' end. Therefore, the hypersensitive site is in the middle of the window starting at 4000 bp, and the overlapping *in vivo* array is in the middle of windows starting at 5000 to 5500 bp. Interestingly, these line up quite accurately with the regions of unusual chromatin structure *in vivo.* In particular, it would appear that the overlapping array observed *in vivo* is very close to the region where the disruption in the dominant period occurs in the simulation at the physiological density (187 bp/nuc).

The second form of presentation is similar to the vertical alignment graphs used in section 4.3.2. Here, as the consequence of the behaviours observed in Figure 4.11, it seems logical to focus on periods between 150 and 250 bp. Therefore 9 BLG and 7 globin simulations are vertically aligned with respect to the underlying sequence and presented in Figures 4.12 and 4.13 respectively. These graphs allow for a direct visual comparison of the PSDs from multiple simulations at various nucleosome densities.

The most striking feature of the BLG scanning PSD analyses presented is the similarity between the periodicities noted from N=51 to 55 (209 to 193 bp nuc). Outwith the ends of the map, which should be regarded with some caution, the scanning PSDs appear to be effectively indistinguishable, with only minor variations. However, there is a dramatic difference when the nucleosome density is increased by just one nucleosome to 190 bp/nuc, which is entering the region of the nucleosome spacing *in vivo.* At 3000 bp, where in the previous graphs the dominant periodicities lengthen to peak above 200 bp, in the N= 56 simulation the dominant period continues to decrease for a further 1000 bp, reaching 170 bp. There is then a discontinuity, as the dominant period breaks up into 3 components ar 175, 200 and 220 bp. This disruption only lasts for approximately 200 bp, whereupon the dominant period appears to return and the pattern continues much as it does in the lower density simulations.

Turning to the N= 57 and N= 58 (187 and 183 bp/nuc respectively) simulations, one can see that the features are broadly similar outwith the end regions of the occupancy maps. It can also be seen that the behaviour is similar to the N= 56 simulation except in the region between 4000 and 6000 bp, where the behaviour diverges. The region in the N= 56 simulation, where the dominant period breaks down into 3 components, is not a feature of these two maps. Instead, there is a short stretch where the dominant period is ~175 bp (~10 bp lower than the average simulated repeat length). Conversely, in the region 4500–5200 bp, where the dominant period returns in the N= 56 simulation, for N= 57 and 58 there is a disruption evident: the periodicity is evenly split between contributions at approximately 210 and 180 bp.

By the time that nucleosome density has reached 180 bp, the simulation loses much of its variability seen at low densities. This process continues as the density is increased: the dominant period matches almost exactly the simulation density and the periodic arrangement is highly regular throughout the maps (data not shown).

Curiously, disruptions are also observable in the globin simulation scanning PSDs. Recall that the globin region contains two genes, β- and ε-globin, and that the adult gene is active at nucleosome densities of ~210 bp/nuc whereas the embryonic gene is active at ~185 bp/nuc. Similar to the pattern observable in the BLG simulations, the first 4 simulations, from N= 37 to 40, appear broadly similar in their periodic characteristics. However, quite unlike the BLG simulation, the disruption in the dominant periodicity appears at low nucleosome densities; it is visible at N=34 (214 bp/nuc). Throughout this disrupted region, there appears to be two contributing periods: one at ~210 bp, one at 175 bp.

The enhancer (the red rectangle on the gene structure map), which plays a crucial role in determining which gene is expressed, is located approximately 3000 bp from the 5' end of the map. Consequently, this region is in the middle of windows which start at 2000 bp. The disruption begins to occur between 1800 and 3200 bp, which is a good fit for the region surrounding the enhancer. By N= 42 (188 bp/nuc), the average spacing at which the embryonic gene is active, the discontinuity disappears and a regular array of nucleosomes, with a mean spacing of 188 bp, appears to have formed. The discontinuity has disappeared.

It is possible that these disruptive regions seen in both BLG and globin are caused by some feature of periodic energy landscapes common to the occupancy maps at higher nucleosome densities. To discount this possibility, control simulations were run with sine wave energy landscapes of various wavelengths from 150 to 250 bp. There were no similar disruptions noted in the PSD analyses of the simulations at physiological nucleosome densities: the occupancy maps and PSDs all show regular periodic behaviour with no disruptions (data not shown).

Discounting this possibility, one can speculate on what may cause these disrupted regions in the simulations at physiological nucleosomal densities. One plausible suggestion is that this behaviour is a result of a confluence of two nucleosomal arrays of different average repeat length.

### 4.3.5 Summary

The simulations tend to produce regular arrays of nucleosomes, the periodic nature of which is dependent on both the region of the dataset and the nucleosome density simulated. If this is a more general behaviour, found throughout the genome, it may explain why regular nucleosome repeat lengths are often found *in vivo* irrespective of the DNA sequence (Blank and Becker, 1995; Blank and Becker, 1996; Becker, 2002), even if the DNA has a wide range of affinities for the histone octamer.

The location of disruptions in the periodic pattern of occupancy sites occurs over functionally relevant regions in both BLG and globin, although only at nucleosome densities which have biological significance.

A natural conclusion to draw from section 4.3.3, and in particular Figure 4.9 b), is that even a small variation in nucleosome density may have dramatic effects on the positions of nucleosomes.

These results must be viewed in the context of the known limitations of the simulation. The simulation model takes into account the two most fundamental factors that are believed to affect nucleosome placement: the positioning affinity of the sequence for the core histone octamer and a fundamental structural constraint that nucleosomes cannot overlap. It follows that the model is justified to the first level of approximation.

Bearing in mind the model limitations raised in section 4.1, the simulations using the simplified model have produced a number of interesting results, which justifies this approach and suggests that further development of the model may prove fruitful. The most significant obstacle preventing the addition of these additional factors is not only determining the relative contribution from each of these additional parameters, but also how each is related to the free energies calculated from the ME maps as part of the MC simulation in chapter 3.

## 4.4 Comparison with BLG *in vivo* map

### 4.4.1 Design

One of the core aims of the simulation approach was to investigate potential links between nucleosome positions adopted *in vivo* and sequence-dependent nucleosome positioning observed *in vitro*. One method to accomplish this is to repeat the analysis performed in section 2.3, where a rectangular filter of various sizes was used to compare the *in vivo* and *in vitro* maps by summing the *in vitro* nucleosome positioning signals within ±75 bp[12] of the BLG *in vivo* positions, for MC occupancy maps of various nucleosome densities. The principal motivation behind this analysis is to determine if the simulation, which is attempting to simulate a more realistic nucleosome density than the conditions under which ME is run (limiting amounts of core histones) has a higher correlation to the *in vivo* data than the ME maps themselves. As previously stated, one can view the ME datasets as representing the probability of finding a single nucleosome positioning at a specific location on the DNA molecule. The equivalent interpretation for the MC occupancy maps is the probability to find a nucleosome at a certainly nucleosome density in the region of DNA simulated. One can therefore speculate that, since BLG has been observed to have a nucleosome repeat length of ~185 bp, that there might be an enhanced correlation between the *in vivo* map and the MC occupancy for this density. It is worthwhile to recall that the simulation only has two determining pieces of information. First, it is given the ME affinity (via the generation of the energy landscape) for a single nucleosome to position itself on each possible positioning site. Second, it knows that a nucleosome cannot be positioned within 168 bp of another nucleosome. It is given no information about the *in vivo* positions.

[12] Up to a window size of 150 bp

### 4.4.2 Results and Discussion

A plot of the Ω/E profiles for occupancy maps at selected nucleosome densities, along with the BLG Ω/E profile for comparative purposes, is presented in Figure 4.14.

What is perhaps most remarkable about these profiles is that the incidence of highest correlation between the *in vivo* map and the MC simulations (for window sizes > 5) occurs when the simulation is run at physiological nucleosomal densities (190 and 187 bp/nuc). Whilst it is true that this could be simple coincidence, it nonetheless remains highly suggestive of a significant correspondence between the *in vivo* and *in vitro* maps, from which the energy landscape used for the BLG simulation is produced.

Another striking feature is the significant increase in values of Ω/E for a window size of 1 bp (just the positioning site directly below the dyad). For all but the highest nucleosome densities, the value of Ω/E increases from ~1.2 to 1.3[13]. Quite why this should be the case is not presently known, although it does suggest that the *in vivo* nucleosomes are located at positioning sites which are significantly enhanced by MC simulations at physiological histone densities.

The overall shape of the profiles, except for very high nucleosome density (N>59, <177 bp/nuc), remains similar particularly in respect to the defining features of the BLG profile. However, two main dissimilarities are noticeable. Firstly, the breadth of the peak in the distribution from 20 to 60 bp identified in section 2.4.3 appears to be narrower for the MC simulation profiles; the hump centred around window size of 30 bp appears to be narrower for the MC occupancy maps than for the ME affinity map. Whilst for the N= 56 and 57 simulation, this increased breadth can to some extent be explained by their significantly higher peak in the range (which would naturally tend to increase the overall width), this is not so for the other 3 densities between 201 and 193 bp/nuc (N=53 to 55). The value of Ω/E profiles for these

[13] With a window size of 1, the value of Ω/E for N= 57, is higher in most of the other simulations plotted; the highest value of Ω/E is from the N= 56 simulation (190 nuc/bp), just one nucleosome less than the approximate physiological density.

simulations appears largely similar to that of the ME datasets from window sizes of 16 to 32.

The second major difference is the values of Ω/E beyond the broad peak. Whereas in the ME dataset, Ω/E is always greater than unity over the range of window sizes, all the MC profiles dip below one between window sizes of 54 for N=59 to 80 for N=57. The values of Ω/E for the N=57, although similar, are however the closest to unity of all the simulations. All values of Ω/E tend asymptotically to one as the window size increases (data not shown).

### 4.4.3 Summary

Enhanced correspondence between the MC simulated occupancies and the experimentally determined *in vivo* positions provides justification for the simulation model. It also provides further corroborating evidence of the relationship between *in vitro* and *in vivo* nucleosome positioning.

The primary conclusion from this analysis, therefore, is that the MC simulation appears, at physiological nucleosome densities, to enhance the relationship between the *in vivo* and *in silico* occupancy maps, which are simulated using the individual site occupancy data from the *in vitro* ME positioning maps. This finding demonstrates the utility of the Monte Carlo simulation approach.

Based on the analysis in this section, as well as section 2.3, it seems apparent that there is strong evidence that the *in vitro* positioning signals determined by ME are quantitatively related to the *in vivo* positions adopted by ME.

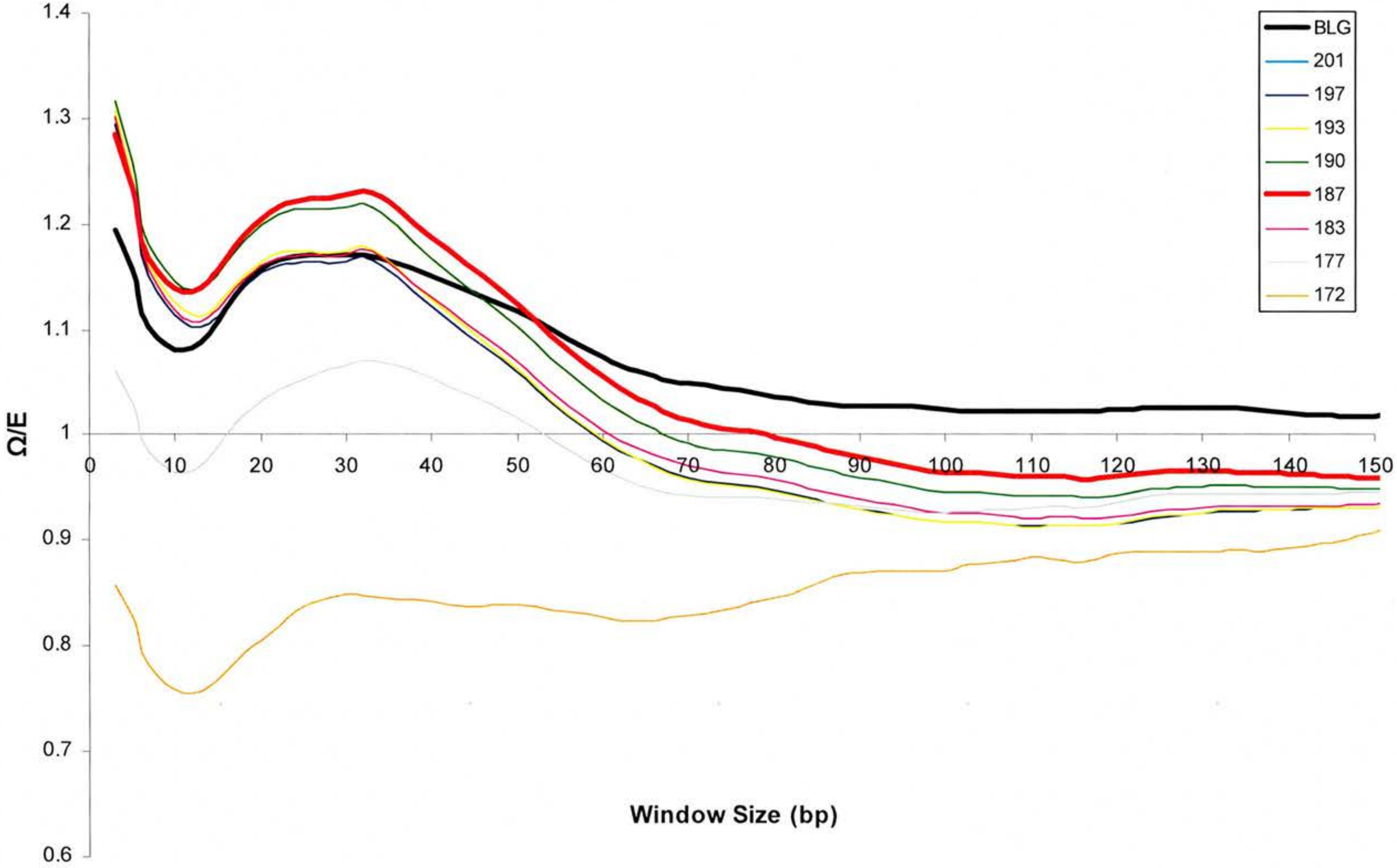


Figure 4.14: Relationship between the ratio of the summed observed (Ω) and expected (E) MC occupancy probabilities contained within windows centred on the dyads of the 50 in vivo nucleosome positioning sites for simulations at various nucleosome densities is presented as a function of window size. For comparison, the BLG Ω/E profile from Figure 2.4 is plotted in black. Nucleosome density in bp/nuc.

CHAPTER 5

# Conclusions and Future Directions

## 5.1 Thesis Synopsis

From the outset, the principal aims of this study were:

- to characterise the experimental nucleosome positioning datasets compiled prior to this research.
- to use this data to assess the influence of *in vitro* (sequence-dependent) nucleosome positioning signals (which involve only DNA-histone interactions) on the arrangement of nucleosomes *in vivo* (which involve DNA-histone interaction plus other factors).

By extending prior analyses, as well as designing novel analyses, a number of different results have been made in the furtherance of these aims. This final chapter brings together these threads and reviews the key results presented throughout the thesis.

In section 2.2, the inherent periodicities within the ME datasets were explored, with each dataset displaying a range of periodic components with biologically significant repeat length which varied between different sections of the maps. It was concluded, based on the work of Davey *et al*. (1995), that such regular nucleosome positioning signals may have an influence on higher order structures. A semi-quantitative comparison between two *in vitro* nucleosome positioning techniques, competitive reconstitution and ME, was presented in section 2.3. The range of affinities for the histone octamer found by both methods were broadly equivalent, leading to the tentative conclusion that the initial capture of the histone octamer (or tetramer) by the

DNA (assessed by competitive reconstitution) and histone octamer positioning (assessed by ME) may well be determined by the same factors of the DNA-histone interaction.

Section 2.4 demonstrated that there is a quantitative relationship between the *in vivo* and *in vitro* BLG nucleosome positioning maps. In and of itself, this analysis therefore lends some support to the proposition that *in vitro* positioning signals have at least some influence on the positions nucleosomes adopt *in vivo*.

An investigation of two readily available programs for predicting nucleosome positioning was presented in section 2.5. The prediction from neither program demonstrated any particular correlation with the ME affinities. Indeed, the predictions themselves showed no signs of any correlation with each other. The unfortunate, but inevitable, conclusion was that the current state of theoretical understanding of factors that influence nucleosome positioning is currently insufficient for reliable prediction of nucleosome positions *in vitro*.

This conclusion led to the development of a novel computational method for analysing the datasets, which takes advantage of the long range, contiguous mapping of nucleosome positioning affinities afforded by ME. The *in vitro* nucleosome positioning affinities are collected at unrealistically low nucleosome density, considerably lower than those found *in vivo*. *In silico*, however, one can combine the *in vitro* ME data with Monte Carlo modelling methods to simulate physiological nucleosome densities in order to extract features from the *in vitro* dataset more relevant to the higher nucleosome density situation *in vivo*. As such, the simulation allows for a more realistic comparison between the *in vivo* and *in vitro* experimental nucleosome positioning maps.

The results from the simulation method proposed in chapter 3 were presented in chapter 4. The straightforward examination of the generated occupancy maps in section 4.2 demonstrated that simulations, with biologically relevant parameters, tend to produce regular periodic arrays of nucleosomes, often enhancing the periodicities inherent to the ME datasets.

Three different types of analyses were presented in section 4.3, each looking for changes in the general features of the occupancy maps as the simulated nucleosome density is increased. The results raise interesting questions about the patterns in the context of both local and global site occupancy changes. The findings demonstrate that, for BLG, extra nucleosomes are generally accommodated within specific regions which respond accordingly, whereas in the globin simulations extra nucleosomes tend to have a more widespread influence upon positioning site occupancies. In particular, it seems that even very small changes in nucleosome density can have a wide ranging impact on positioning site occupancy. Perhaps the most interesting of these changes occurs at the points of conjunction between the two out of phase regular arrays of nucleosomes, where arrays with different repeat lengths meet in the middle of both the BLG and globin maps, at regions of functional significance. In the case of BLG, this occurs within a coding region which contains unusual chromatin structure, and for globin, it occurs around the enhancer.

Finally, in section 4.4, the occupancy maps from the BLG MC simulations were compared against the *in vivo* positions mapped by indirect end-labelling in a similar fashion to the analysis presented in section 2.4. The results demonstrated that the correlation between the *in vivo* and *in silico* maps is enhanced markedly at physiological nucleosome densities (187 – 190 bp/nuc). It was therefore concluded that *in vivo* nucleosome positioning seems to be strongly influenced by the same factors which determine nucleosome positions *in vitro*, and that therefore the ME datasets (and other techniques which assess positioning) have a significant role to play in predicting nucleosome positioning *in vivo*. The results also lend significant weight to the simplified model used by the simulation, and suggested that further development of the model could prove a fruitful avenue for future development.

## 5.2 Biological Implications

On a local level, the disruptions noted from the periodicity analyses presented in section 4.3.4 suggest that irregular higher order chromatin structures may well form at these regions. Such irregularities in the chromatin may well modulate access to the underlying sequence, which is normally inaccessible when packaged into higher order chromatin structures (see Figure 5.1 (b)). This could explain why these disruptions are seen within functionally important regions of both genes, in particular allowing more ready access to the first 4 exons in BLG and the enhancer in globin.

Taking a wider view, there is another potential repercussion of these irregular chromatin structures. The 30 nm fibre itself is known to be packaged into higher order structures (Figure 1.1). It is possible that these disruptions represent weak points within the 30 nm fibre that affords the bending potential required to allow the formation of the compact chromosome structure represented at the bottom of Figure 1.1. Figure 5.1 c) and d) are schematics of how this may work in practice, with these localised fragile points allowing the otherwise relatively inflexible fibre to bend. Higher order chromatin fibres require this level of flexibility to allow them to be efficiently packaged and to allow the interaction of spatially separated features within individual fibres, such as the interaction between promoters and enhancers. Sedimentation analyses of chromatin structure have demonstrated that chromatin displays a range of conformations: centromeric chromatin are typically more regularly packed into rigid fibres, whereas chromatin found in the bulk of the genome tends to display less regular folding consistent with occasional disruption (Gilbert and Allan, 2001). These folding irregularities were subsequently found to be more prevalent in regions of the genome which have an abundance of genes (Gilbert *et al.*, 2004). The analyses presented in section 4.3.4 are consistent with these observations and open up a new perspective on their interpretation.

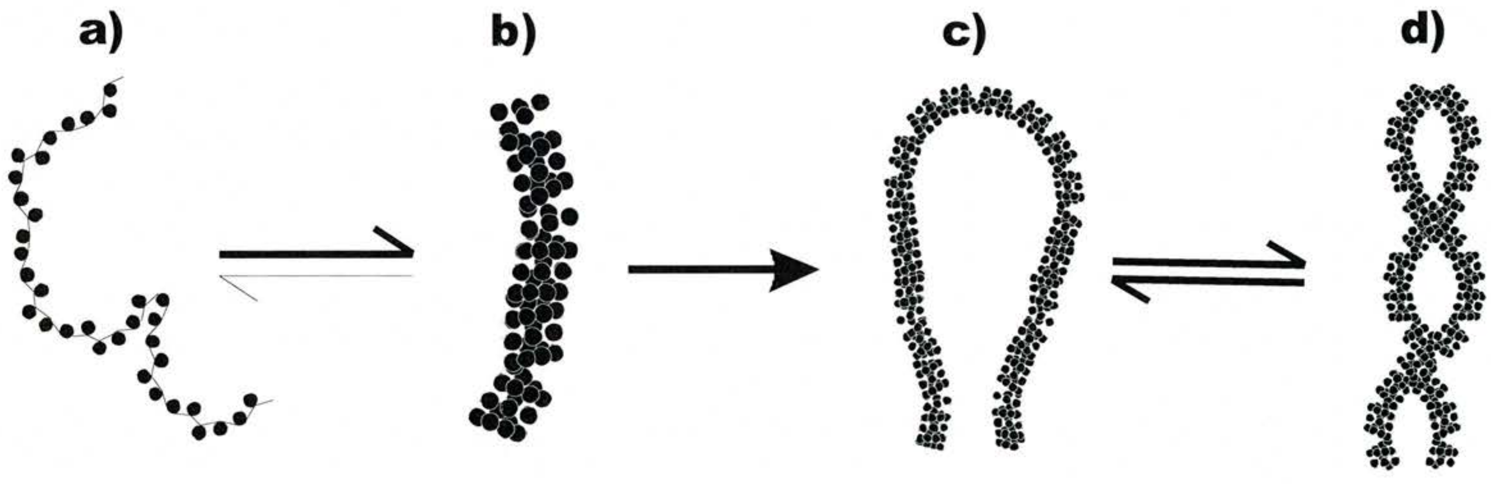


Figure 5.1: Models for various stages of chromatin packaging. a) 11-nm or "beads-on-a-string" level. b) A representation of a possible short stretch of the 30 nm fibre. Note the disruptions in the fibre highlighted by the red arrows. c) Representation of a looped chromatin domain, the bending of which is facilitated by disruptions in the structure of the fibre. d) Further compaction of c).

## 5.3 Future Directions

Achieving the ambitious goals in chapter 1, which set the scope for this thesis, remains an area of active research interest. Significant steps have been taken during the period of the work presented here, but there exists many areas for future work.

Improving the simplified simulation model is a prime candidate for future work. Including parameters accounting for steric and DNA orientation effects, based on recent work, could provide further insights. Some examples of factors which could be included: theoretical work on nucleosome-nucleosome interactions (Mergell *et al.*, 2004), ongoing efforts to finally resolve the structure of the 30 nm fibre (Schalch *et al.*, 2005), work developing the understanding of the underlying physics (reviewed in Schiessel (2003)), as well as recent molecular dynamics simulations of DNA and nucleosome (Bishop, 2005). In addition, there is scope to link the MC methods based simulation used here with more direct simulations of the 30 nm fibre (Katritch *et al.*, 2000; Wedemann and Langowski, 2002). Such studies will inevitably rely, at least in part, on the determination of structure of the 30 nm fibre; this is of critical importance as the diverse models suggest different opportunities for nucleosome-nucleosome interactions (Figure 1.4).

With the MC data already produced, there exists the possibility of refining and expanding the analyses presented. In particular, in individual site occupancy configurations, stored at every step of the MC simulations, could potentially prove a productive avenue for additional effort. Improvements in computational resources and storage (even over the last few years) may well prove sufficient to allow at least limited analysis of the ~1 terabyte of data generated thus far. An example of potential uses for this data include studying the possible correlations between individual binding site occupancies (as opposed to studying the configuration-averaged occupancy maps).

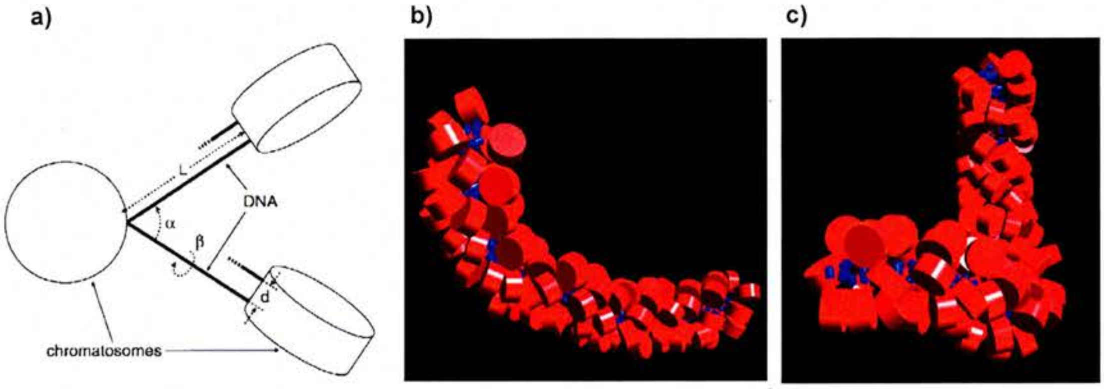


**Figure 5.2:** ***In silico*** **simulation of a 30 nm fibre. a) Model of fibre geometry used to simulate the 30 nm fibre.** ***L*** **represents the length of linker DNA, α is the angle between the linker DNAs, β is twist angle between neighbouring nucleosomes, and** ***d*** **is the distance between the incoming and outgoing linker DNA. b) and c) Two conformations of a 100 nucleosome simulated 30 nm fibre, with α=26°, β=110°,** ***d*****= 3.1 nm and** ***L*****= 11 bp. Figure adapted from (Wedemann and Langowski, 2002)**

There is also the potential to use the simulation data to produce realistic models of 30 nm chromatin fibres. It is possible to view the chromatin fibre, from the point of view of studying its physical properties, as having many of the same properties as DNA, except that instead of the base composition determining the physical characteristics of the fibre, the properties are determined by the specific distribution and geometrical properties of neighbouring nucleosomes (see Figure 5.2 a)). Figure 5.2 b) and c) represent two such simulated fibres, with biologically relevant parameters, but with constant linker length. There are two possible methods for generating such fibres from the MC simulation method presented in chapter 3. The first would involve using optimisation techniques on the ME affinity and MC occupancy maps to find the energetically most favourable combination of positioning sites and produce models using parameters derived from such combinations. The second method would use the MC nucleosome configuration data stored every MCS to generate the appropriate model parameters. This work would, to some extent, be reliant on the resolution of the structure of the 30 nm fibre, although previous simulations of this type have provided insights that could assist in the determination of the most appropriate general structure for the 30 nm fibre (Wedemann and Langowski, 2002).

In particular, a more thorough analysis of the sequence information, making use of the quantitative nature of the nucleosome positioning signals available from ME, may well suggest further avenues in which to expand current sequence analyses and develop novel techniques. The more recent ME datasets, at very high resolution, for human and mouse H19 (Davey *et al.*, 2003) and mouse Igf2r (Davey and Allan, 2003) presents an excellent opportunity for in depth analyses. Repeating and developing existing algorithms, including the two explored in section 2.5, using these datasets is particularly appealing, especially as the potential error associated with these smaller datasets reduces the problem of reliably identifying the sequence responsible for the given nucleosome positioning signal. Extension of preliminary work using machine learning pattern recognition technique may also prove fruitful, especially with more accurate nucleosome positioning data.

It would be unfair, given current understanding, and the lack of experimental data, to be too dismissive of the current crop of computational algorithms for nucleosome positioning. Levitsky *et al.*'s RECON program in particular seems based on sound principles. It is likely that, whilst the method is sound, the reason for the failure to predict ME positioning signals lies with the quality of the training data. Hence, this method could be adapted to make use of the high resolution, quantitative ME nucleosome positioning data.

Undoubtedly the biggest single obstacle to solving the issues surrounding nucleosome positioning is the lack of high quality, quantitative nucleosome positioning data, both *in vivo* and *in vitro*. However, there is potential scope for automation of the ME technique. Such automation could potentially allow for the rapid collection of significant quantities of very high resolution, quantitative nucleosome positioning data and such data may well prove the catalyst for making significant headway towards answering many of currently unresolved issues faced by those working in the chromatin field. And chief amongst those problems is what factors determine nucleosome positioning. It promises to be a stimulating (and hopefully rewarding) period for nucleosome positioning research.

# References

Adroer, R., and Oliva, R. (1998). Nucleosome positioning in the rat protamine 1 gene in vivo and in vitro. *Biochim Biophys Acta* **1442**, 252-260.

Alder, B. J., and Wainwright, T. E. (1957). Phase Transition for a Hard Sphere System. *Journal of Chemical Physics* **27**, 1208-1209.

Allan, J., Rau, D. C., Harborne, N., and Gould, H. (1984). Higher order structure in a short repeat length chromatin. *J Cell Biol* **98**, 1320-1327.

Anderson, J. D., Thastrom, A., and Widom, J. (2002). Spontaneous access of proteins to buried nucleosomal DNA target sites occurs via a mechanism that is distinct from nucleosome translocation. *Mol Cell Biol* **22**, 7147-7157.

Baldi, P., Brunak, S., Chauvin, Y., and Krogh, A. (1996). Naturally occurring nucleosome positioning signals in human exons and introns. *J Mol Biol* **263**, 503-510.

Bash, R. C., Vargason, J. M., Cornejo, S., Ho, P. S., and Lohr, D. (2001). Intrinsically bent DNA in the promoter regions of the yeast GAAL1-10 and GAL80 genes. *J Biol Chem* **276**, 861-866.

Beard, P. (1978). Mobility of histones on the chromosome of simian virus 40. *Cell* **15**, 955-967.

Becker, P. B. (2002). Nucleosome sliding: facts and fiction. *Embo J* **21**, 4749-4753.

Becker, P. B., and Horz, W. (2002). ATP-dependent nucleosome remodeling. *Annu Rev Biochem* **71**, 247-273.

Bennink, M. L., Leuba, S. H., Leno, G. H., Zlatanova, J., de Grooth, B. G., and Greve, J. (2001). Unfolding individual nucleosomes by stretching single chromatin fibers with optical tweezers. *Nat Struct Biol* **8**, 606-610.

Binder, K. (1986). Monte Carlo methods in statistical physics, 2nd edn (Berlin: Springer-Verlag).

Bishop, T. C. (2005). Molecular dynamics simulations of a nucleosome and free DNA. *Journal of Biomolecular Structure & Dynamics* **22**, 673-685.

Blank, T. A., and Becker, P. B. (1995). Electrostatic mechanism of nucleosome spacing. *J Mol Biol* **252**, 305-313.

Blank, T. A., and Becker, P. B. (1996). The effect of nucleosome phasing sequences and DNA topology on nucleosome spacing. *J Mol Biol* **260**, 1-8.

Blomquist, P., Belikov, S., and Wrange, O. (1999). Increased nuclear factor 1 binding to its nucleosomal site mediated by sequence-dependent DNA structure. *Nucleic Acids Res* **27**, 517-525.

Boa, S. A. (1999) Nucleosomal organisation over the ovine Beta-Lactoglobulin gene. Ph.D Thesis University of Edinburgh.

Bracewell, R. N. (1986). The Fourier transform and its applications, 2nd edn (New York ; London: McGraw-Hill).

Brower-Toland, B. D., Smith, C. L., Yeh, R. C., Lis, J. T., Peterson, C. L., and Wang, M. D. (2002). Mechanical disruption of individual nucleosomes reveals a reversible multistage release of DNA. *Proc Natl Acad Sci U S A* **99**, 1960-1965.

Bucher, P., and Trifonov, E. N. (1986). Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res* **14**, 10009-10026.

Buttinelli, M., Di Mauro, E., and Negri, R. (1993). Multiple nucleosome positioning with unique rotational setting for the Saccharomyces cerevisiae 5S rRNA gene in vitro and in vivo. *Proc Natl Acad Sci U S A* **90**, 9315-9319.

Calladine, C. R., and Drew, H. R. (1986). Principles of sequence-dependent flexure of DNA. *J Mol Biol* **192**, 907-918.

Calladine, C. R., and Drew, H. R. (1997). Understanding DNA : the molecule & how it works, 2nd edn (San Diego, Calif. ; London: Academic).

Cao, H., Widlund, H. R., Simonsson, T., and Kubista, M. (1998). TGGA repeats impair nucleosome formation. *J Mol Biol* **281**, 253-260.

Choi, O. R., and Engel, J. D. (1988). Developmental regulation of beta-globin gene switching. *Cell* **55**, 17-26.

Clark, D. J., and Kimura, T. (1990). Electrostatic mechanism of chromatin folding. *J Mol Biol* **211**, 883-896.

Cohanim, A. B., Kashi, Y., and Trifonov, E. N. (2005). Yeast nucleosome DNA pattern: deconvolution from genome sequences of S. cerevisiae. *J Biomol Struct Dyn* **22**, 687-694.

Cohanim, A. B., Kashi, Y., and Trifonov, E. N. (2006). Three sequence rules for chromatin. *J Biomol Struct Dyn* **23**, 559-566.

Cui, Y., and Bustamante, C. (2000). Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure. *Proc Natl Acad Sci U S A* **97**, 127-132.

Daban, J. R., and Bermudez, A. (1998). Interdigitated solenoid model for compact chromatin fibers. *Biochemistry* **37**, 4299-4304.

Davey, C., and Allan, J. (2003). Nucleosome positioning signals and potential H-DNA within the DNA sequence of the imprinting control region of the mouse Igf2r gene. *Biochim Biophys Acta* **1630**, 103-116.

Davey, C., Fraser, R., Smolle, M., Simmen, M. W., and Allan, J. (2003). Nucleosome positioning signals in the DNA sequence of the human and mouse H19 imprinting control regions. *J Mol Biol* **325**, 873-887.

Davey, C., Pennings, S., and Allan, J. (1997). CpG methylation remodels chromatin structure in vitro. *J Mol Biol* **267**, 276-288.

Davey, C., Pennings, S., Meersseman, G., Wess, T. J., and Allan, J. (1995). Periodicity of strong nucleosome positioning sites around the chicken adult beta-globin gene may encode regularly spaced chromatin. *Proc Natl Acad Sci U S A* **92**, 11210-11214.

Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W., and Richmond, T. J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution. *J Mol Biol* **319**, 1097-1113.

Drew, H. R. (1991). Can one measure the free energy of binding of the histone octamer to different DNA sequences by salt-dependent reconstitution? *J Mol Biol* **219**, 391-392.

Drew, H. R., and Calladine, C. R. (1987). Sequence-specific positioning of core histones on an 860 base-pair DNA. Experiment and theory. *J Mol Biol* **195**, 143-173.

Drew, H. R., and Travers, A. A. (1985). DNA bending and its relation to nucleosome positioning. *J Mol Biol* **186**, 773-790.

Edayathumangalam, R. S., Weyermann, P., Gottesfeld, J. M., Dervan, P. B., and Luger, K. (2004). Molecular recognition of the nucleosomal "supergroove". *Proc Natl Acad Sci U S A* **101**, 6864-6869.

Edgar, M. (1999) Nucleosome positioning on the chicken beta-globin genes. PhD Thesis, The University of Edinburgh.

Ellington, A. D., and Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818-822.

Evans, T., Felsenfeld, G., and Reitman, M. (1990). Control of globin gene transcription. *Annu Rev Cell Biol* **6**, 95-124.

Felsenfeld, G., and Groudine, M. (2003). Controlling the double helix. *Nature* **421**, 448-453.

Finch, J. T., and Klug, A. (1976). Solenoidal model for superstructure in chromatin. *Proc Natl Acad Sci U S A* **73**, 1897-1901.

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

Fitzgerald, D. J., Dryden, G. L., Bronson, E. C., Williams, J. S., and Anderson, J. N. (1994). Conserved patterns of bending in satellite and nucleosome positioning DNA. *J Biol Chem* **269**, 21303-21314.

Foley, K. P., and Engel, J. D. (1992). Individual stage selector element mutations lead to reciprocal changes in beta- vs. epsilon-globin gene transcription: genetic confirmation of promoter competition during globin gene switching. *Genes Dev* **6**, 730-744.

Gardiner, E. J., Hunter, C. A., Packer, M. J., Palmer, D. S., and Willett, P. (2003). Sequence-dependent DNA structure: a database of octamer structural parameters. *J Mol Biol* **332**, 1025-1035.

Gencheva, M., Boa, S., Fraser, R. M., Simmen, M. W., Whitelaw, B., and Allan, J. (2006). In vitro and in vivo nucleosome positioning on the ovine beta-lactoglobulin gene are related. *Submitted.*

Gilbert, N., and Allan, J. (2001). Distinctive higher-order chromatin structure at mammalian centromeres. *Proc Natl Acad Sci U S A* **98**, 11949-11954.

Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N. P., and Bickmore, W. A. (2004). Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555-566.

Goodsell, D. S., and Dickerson, R. E. (1994). Bending and curvature calculations in B-DNA. *Nucleic Acids Res* **22**, 5497-5503.

Gould, H. (1998). Chromatin : a practical approach (Oxford: Oxford University Press).

Graziano, V., Gerchman, S. E., Schneider, D. K., and Ramakrishnan, V. (1994). Histone H1 is located in the interior of the chromatin 30-nm filament. *Nature* **368**, 351-354.

Hagerman, P. J. (1988). Flexibility of DNA. *Annual Review of Biophysics and Biophysical Chemistry* **17**, 265-286.

Hammersley, J. M., and Handscomb, D. C. (1965). Monte Carlo methods, Repr. with minor corr. edn (London
New York: Methuen
Wiley).

Herzel, H., Weiss, O., and Trifonov, E. N. (1998). Sequence periodicity in complete genomes of archaea suggests positive supercoiling. *J Biomol Struct Dyn* **16**, 341-345.

Herzel, H., Weiss, O., and Trifonov, E. N. (1999). 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics* **15**, 187-193.

Hozier, J., Renz, M., and Nehls, P. (1977). The chromosome fiber: evidence for an ordered superstructure of nucleosomes. *Chromosoma* **62**, 301-317.

Ioshikhes, I., Bolshoy, A., Derenshteyn, K., Borodovsky, M., and Trifonov, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J Mol Biol* **262**, 129-139.

Ioshikhes, I., Bolshoy, A., and Trifonov, E. N. (1992). Preferred positions of AA and TT dinucleotides in aligned nucleosomal DNA sequences. *J Biomol Struct Dyn* **9**, 1111-1117.

Jackson, J. R., and Benyajati, C. (1993). DNA-histone interactions are sufficient to position a single nucleosome juxtaposing Drosophila Adh adult enhancer and distal promoter. *Nucleic Acids Res* **21**, 957-967.

Karplus, M., and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature Structural Biology* **9**, 646-652.

Katritch, V., Bustamante, C., and Olson, W. K. (2000). Pulling chromatin fibers: computer simulations of direct physical micromanipulations. *J Mol Biol* **295**, 29-40.

Kefalas, P., Gray, F. C., and Allan, J. (1988). Precise nucleosome positioning in the promoter of the chicken beta A globin gene. *Nucleic Acids Res* **16**, 501-517.

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

Klug, A., Rhodes, D., Smith, J., Finch, J. T., and Thomas, J. O. (1980). A low resolution structure for the histone core of the nucleosome. *Nature* **287**, 509-516.

Kornberg, R. D., and Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285-294.

Kulic, I. M., and Schiessel, H. (2003a). Chromatin dynamics: nucleosomes go mobile through twist defects. *Phys Rev Lett* **91**, 148103.

Kulic, I. M., and Schiessel, H. (2003b). Nucleosome repositioning via loop formation. *Biophys J* **84**, 3197-3211.

Langmore, J. P., and Schutt, C. (1980). The higher order structure of chicken erythrocyte chromosomes in vivo. *Nature* **288**, 620-622.

Levitsky, V. G. (2004). RECON: a program for prediction of nucleosome formation potential. *Nucleic Acids Res* **32**, W346-349.

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

Levitsky, V. G., Ponomarenko, M. P., Ponomarenko, J. V., Frolov, A. S., and Kolchanov, N. A. (1999). Nucleosomal DNA property database. *Bioinformatics* **15**, 582-592.

Li, G., Levitus, M., Bustamante, C., and Widom, J. (2005). Rapid spontaneous accessibility of nucleosomal DNA. *Nat Struct Mol Biol* **12**, 46-53.

Li, G., and Widom, J. (2004). Nucleosomes facilitate their own invasion. *Nat Struct Mol Biol* **11**, 763-769.

Lowary, P. T., and Widom, J. (1997). Nucleosome packaging and nucleosome positioning of genomic DNA. *Proc Natl Acad Sci U S A* **94**, 1183-1188.

Lowary, P. T., and Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol* **276**, 19-42.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260.

Martinez-Campa, C., Politis, P., Moreau, J. L., Kent, N., Goodall, J., Mellor, J., and Goding, C. R. (2004). Precise nucleosome positioning and the TATA box dictate requirements for the histone H4 tail and the bromodomain factor Bdf1. *Mol Cell* **15**, 69-81.

Meersseman, G., Pennings, S., and Bradbury, E. M. (1992). Mobile nucleosomes--a general behavior. *Embo J* **11**, 2951-2959.

Mengeritsky, G., and Trifonov, E. N. (1983). Nucleotide sequence-directed mapping of the nucleosomes. *Nucleic Acids Res* **11**, 3833-3851.

Mergell, B., Everaers, R., and Schiessel, H. (2004). Nucleosome interactions in chromatin: fiber stiffening and hairpin formation. *Phys Rev E Stat Nonlin Soft Matter Phys* **70**, 011915.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics* **21**, 1087-1092.

Metropolis, N., and Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association* **44**, 335-341.

Muyldermans, S., and Travers, A. A. (1994). DNA sequence organization in chromatosomes. *J Mol Biol* **235**, 855-870.

Negri, R., Buttinelli, M., Panetta, G., De Arcangelis, V., Di Mauro, E., and Travers, A. (2001). Sequence dependence of translational positioning of core nucleosomes. *J Mol Biol* **307**, 987-999.

Ner, S. S., Blank, T., Perez-Paralle, M. L., Grigliatti, T. A., Becker, P. B., and Travers, A. A. (2001). HMG-D and histone H1 interplay during chromatin assembly and early embryogenesis. *J Biol Chem* **276**, 37569-37576.

Newman, M. E. J., and Barkema, G. T. (1999). Monte Carlo methods in statistical physics (Oxford: Clarendon Press).

Patterton, H. G., and Graves, S. (2000). DNAssist: the integrated editing and analysis of molecular biology sequences in Windows. *Bioinformatics* **16**, 652-653.

Pennings, S., Meersseman, G., and Bradbury, E. M. (1991). Mobility of positioned nucleosomes on 5 S rDNA. *J Mol Biol* **220**, 101-110.

Renz, M., Nehls, P., and Hozier, J. (1977). Involvement of histone H1 in the organization of the chromosome fiber. *Proc Natl Acad Sci U S A* **74**, 1879-1883.

Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D., and Klug, A. (1984). Structure of the nucleosome core particle at 7 A resolution. *Nature* **311**, 532-537.

Rubinstein, R. Y. (1981). Simulation and the Monte Carlo method (New York ; Chichester: Wiley).

Satchwell, S. C., Drew, H. R., and Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* **191**, 659-675.

Satchwell, S. C., and Travers, A. A. (1989). Asymmetry and polarity of nucleosomes in chicken erythrocyte chromatin. *Embo J* **8**, 229-238.

Schalch, T., Duda, S., Sargent, D. F., and Richmond, T. J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature* **436**, 138-141.

Schieg, P., and Herzel, H. (2004). Periodicities of 10-11bp as indicators of the supercoiled state of genomic DNA. *J Mol Biol* **343**, 891-901.

Schiessel, H. (2003). The physics of chromatin. *Journal of Physics-Condensed Matter* **15**, R699-R774.

Schmid, C. D., Perier, R., Praz, V., and Bucher, P. (2006). EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res* **34**, D82-85.

Shen, C. H., Leblanc, B. P., Alfieri, J. A., and Clark, D. J. (2001). Remodeling of yeast CUP1 chromatin involves activator-dependent repositioning of nucleosomes over the entire gene and flanking sequences. *Mol Cell Biol* **21**, 534-547.

Shrader, T. E., and Crothers, D. M. (1989). Artificial nucleosome positioning sequences. *Proc Natl Acad Sci U S A* **86**, 7418-7422.

Simpson, R. T. (1991). Nucleosome positioning: occurrence, mechanisms, and functional consequences. *Prog Nucleic Acid Res Mol Biol* **40**, 143-184.

Staffelbach, H., Koller, T., and Burks, C. (1994). DNA structural patterns and nucleosome positioning. *J Biomol Struct Dyn* **12**, 301-325.

Staynov, D. Z. (2000). DNase I digestion reveals alternating asymmetrical protection of the nucleosome by the higher order chromatin structure. *Nucleic Acids Res* **28**, 3092-3099.

Stein, A., and Bina, M. (1999). A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucleic Acids Res* **27**, 848-853.

Tanaka, S., Zatchej, M., and Thoma, F. (1992). Artificial nucleosome positioning sequences tested in yeast minichromosomes: a strong rotational setting is not sufficient to position nucleosomes in vivo. *Embo J* **11**, 1187-1193.

Thastrom, A., Bingham, L. M., and Widom, J. (2004a). Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J Mol Biol* **338**, 695-709.

Thastrom, A., Lowary, P. T., Widlund, H. R., Cao, H., Kubista, M., and Widom, J. (1999). Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J Mol Biol* **288**, 213-229.

Thastrom, A., Lowary, P. T., and Widom, J. (2004b). Measurement of histone-DNA interaction free energy in nucleosomes. *Methods* **33**, 33-44.

Thoma, F. (1992). Nucleosome positioning. *Biochim Biophys Acta* **1130**, 1-19.

Thoma, F., Koller, T., and Klug, A. (1979). Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *J Cell Biol* **83**, 403-427.

Tomita, M., Wada, M., and Kawashima, Y. (1999). ApA dinucleotide periodicity in prokaryote, eukaryote, and organelle genomes. *J Mol Evol* **49**, 182-192.

Travers, A., and Drew, H. (1997). DNA recognition and nucleosome organization. *Biopolymers* **44**, 423-433.

Travers, A. A., and Muyldermans, S. V. (1996). A DNA sequence for positioning chromatosomes. *J Mol Biol* **257**, 486-491.

Trifonov, E. N., and Sussman, J. L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc Natl Acad Sci U S A* **77**, 3816-3820.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-510.

Uberbacher, E. C., Harp, J. M., and Bunick, G. J. (1988). DNA sequence patterns in precisely positioned nucleosomes. *J Biomol Struct Dyn* **6**, 105-120.

Ulyanov, A. V., and Stormo, G. D. (1995). Multi-alphabet consensus algorithm for identification of low specificity protein-DNA interactions. *Nucleic Acids Res* **23**, 1434-1440.

van Holde, K., and Zlatanova, J. (1996). What determines the folding of the chromatin fiber? *Proc Natl Acad Sci U S A* **93**, 10548-10555.

Van Holde, K. E. (1989). Chromatin (New York: Springer-Verlag).

Wada-Kiyama, Y., Kuwabara, K., Sakuma, Y., Onishi, Y., Trifonov, E. N., and Kiyama, R. (1999a). Localization of curved DNA and its association with nucleosome phasing in the promoter region of the human estrogen receptor alpha gene. *FEBS Lett* **444**, 117-124.

Wada-Kiyama, Y., Suzuki, K., and Kiyama, R. (1999b). DNA bend sites in the human beta-globin locus: evidence for a basic and universal structural component of genomic DNA. *Mol Biol Evol* **16**, 922-930.

Wang, J. P., and Widom, J. (2005). Improved alignment of nucleosome DNA sequences using a mixture model. *Nucleic Acids Res* **33**, 6743-6755.

Wasserman, W. W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276-287.

Wedemann, G., and Langowski, J. (2002). Computer simulation of the 30-nanometer chromatin fiber. *Biophys J* **82**, 2847-2859.

Weintraub, H. (1978). The nucleosome repeat length increases during erythropoiesis in the chick. *Nucleic Acids Res* **5**, 1179-1188.

Widlund, H. R., Cao, H., Simonsson, S., Magnusson, E., Simonsson, T., Nielsen, P. E., Kahn, J. D., Crothers, D. M., and Kubista, M. (1997). Identification and characterization of genomic nucleosome-positioning sequences. *J Mol Biol* **267**, 807-817.

Widom, J. (1992). A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells. *Proc Natl Acad Sci U S A* **89**, 1095-1099.

Widom, J. (1996). Short-range order in two eukaryotic genomes: Relation to chromosome structure. *Journal of Molecular Biology* **259**, 579-588.

Widom, J. (1998). Structure, dynamics, and function of chromatin in vitro. *Annu Rev Biophys Biomol Struct* **27**, 285-327.

Widom, J. (2001). Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys* **34**, 269-324.

Widom, J., Finch, J. T., and Thomas, J. O. (1985). Higher-order structure of long repeat chromatin. *Embo J* **4**, 3189-3194.

Williams, S. P., Athey, B. D., Muglia, L. J., Schappe, R. S., Gough, A. H., and Langmore, J. P. (1986). Chromatin fibers are left-handed double helices with diameter and mass per unit length that depend on linker length. *Biophys J* **49**, 233-248.

Wolffe, A. (1998). Chromatin : structure and function, 3rd edn (San Diego, Calif. ; London: Academic Press).

Wong, M., Allan, J., and Smulson, M. (1984). The mechanism of histone H1 cross-linking by poly(ADP-ribosylation). Reconstitution with peptide domains. *J Biol Chem* **259**, 7963-7969.

Woodcock, C. L., Frado, L. L., and Rattner, J. B. (1984). The higher-order structure of chromatin: evidence for a helical ribbon arrangement. *J Cell Biol* **99**, 42-52.

Worcel, A., Strogatz, S., and Riley, D. (1981). Structure of chromatin and the linking number of DNA. *Proc Natl Acad Sci U S A* **78**, 1461-1465.

Wu, C. (1980). The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**, 854-860.

Wu, C., and Travers, A. (2005). Relative affinities of DNA sequences for the histone octamer depend strongly upon both the temperature and octamer concentration. *Biochemistry* **44**, 14329-14334.

Yenidunya, A., Davey, C., Clark, D., Felsenfeld, G., and Allan, J. (1994). Nucleosome positioning on chicken and human globin gene promoters in vitro. Novel mapping techniques. *J Mol Biol* **237**, 401-414.

Zentgraf, H., and Franke, W. W. (1984). Differences of supranucleosomal organization in different kinds of chromatin: cell type-specific globular subunits containing different numbers of nucleosomes. *J Cell Biol* **99**, 272-286.