A computational approach to characterizing nucleosome dynamics

By Ngoc Tu Le

A thesis submitted to
School of Knowledge Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Knowledge Science
Graduate Program in Knowledge Science

Written under the direction of
Professor Tu Bao Ho

March, 2010

A computational approach to characterizing nucleosome dynamics

By Ngoc Tu Le (850803)

A thesis submitted to

School of Knowledge Science,

Japan Advanced Institute of Science and Technology,

in partial fulfillment of the requirements

for the degree of

Master of Knowledge Science

Graduate Program in Knowledge Science

Written under the direction of

Professor Tu Bao Ho

and approved by

Professor Tu Bao Ho

Professor Mitsuru Ikeda

Professor Takuya Honda

Associate Professor Takaya Yuizono

February, 2010 (Submitted)

Abstract

The ability of the cells in living organisms to grow, replicate, repair themselves and even evolve under the stimuli of environment depends much on how genetic information kept inside the cell nucleus is expressed through diverse biological processes, such as protein synthesis, DNA replication, DNA repair and genetic recombination. By controlling these processes, the cell can decide most of its functions.

Genomes of eukaryotic organisms are packaged into chromatin, a compact structure containing fundamental units of nucleosomes. The mobility of nucleosomes is known to play important roles in many DNA-related processes by regulating the accessibility of regulatory elements to biological machineries. Although it has been known that various factors, such as DNA sequences, histone modifications and histone variants, chromatin remodeling complexes, could affect nucleosome stability, the mechanisms of how they regulate this stability are still unclear.

The work here proposed a computational method based on rule induction learning for characterizing nucleosome dynamics using both genomic and histone modification information. Our results on *S.cerevisiae* showed that, some DNA motifs and post-translational modifications of histone proteins play significant roles in regulating nucleosome stability. Interestingly, these DNA motifs are strong determinants for nucleosome forming and inhibiting, and these histone modifications have strong relation with transcriptional activation and repression. We also found some new patterns which may reflect the cooperation between these two factors in regulating the stability of nucleosomes. These results led to the conclusion that DNA motifs and histone modifications can independently and, in some cases, cooperatively regulate nucleosome stability. This suggests additional insights into mechanisms by which cells control important biological processes, such as transcription, replication and DNA repair.

Acknowledgement

I would like to express my sincere thank to the Ministry of Education, Culture, Sports, Science and Technology of Japan for its financial support for my study in Japan. Without such support I would not be able to complete this work.

I am indebted to Professor Tu Bao Ho, my supervisor, for offering me an opportunity to study in his laboratory and for his tireless guidance and constant encouragement during my MSc period.

I am also grateful to the members of Ho Laboratory for their kindly helps to me not only in research work but also in my daily life. It would be a mistake if I forget to thank all Vietnamese members at JAIST for their kindly treat to me.

I would like to show my sincere appreciation to Japan Advanced Institute of Science and Technology for providing me uncountable support and fantastic working environment during my memorable time here.

Finally, I dedicate this work to my family and my special friend for their endless loves, everlasting encouragement and tolerance during my hard time pursuing research at JAIST.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem

The cell's ability to maintain a high degree of order under various stimuli depends on how genetic information is expressed, maintained, replicated, and occasionally evolved through basic cellular processes such as RNA and protein synthesis, DNA repair, DNA replication, and genetic recombination. In these processes, which produce and maintain the proteins and nucleic acids of a cell, the information in a sequence of nucleotides is used to specify either another chain of nucleotides (a DNA or an RNA molecule) or a chain of amino acids (a protein molecule). By controlling these processes, the cell can decide all of its functions throughout its life cycle.

We have known for a long time that genetic materials of eukaryotic organisms are packaged into chromatin inside cell nucleus. Since it was first recognized [2], there have been increasing evidences showing that chromatin plays a much more important role far beyond DNA compaction. By burying *cis*-regulatory elements under histone proteins and/or modifying related epigenetic information, chromatin imposes ubiquitous and profound effects on many DNA-based processes, including transcription, DNA repair and replication. To ensure faithfully copy both genetic and epigenetic information during replication or to facilitate the binding of Transcription Factors (TFs) to regulatory elements during transcription in the context of chromatin, cells have developed complicated

biological pathways [35, 3, 5, 6]. In these pathways, by regulating nucleosome stability cells can control the accessibility of underlying DNA sequences to biological machineries. For example, in replication, during the process known as parental histone segregation, pre-existing nucleosomes located ahead of replication forks are transiently disrupted from parental DNA strands and later transferred onto nascent DNA [35, 5]. In transcription, moving nucleosomes to different translational positions is known as one way to change the accessibility of nucleosomal DNA to TFs [3]. Also, promoter regions of actively transcribed genes are usually free of nucleosomes [7, 8]. So, understanding how cells regulate nucleosome stability will bring us additional insights into mechanisms of many important biological processes.

1.2 Related works

Nucleosome stability can be regulated by many factors, such as DNA sequences, histone modifications and histone variants, and chromatin remodelling complexes [9]. For example, DNA sequence is known as a reliable determinant for nucleosome preference, which can be used to predict nearly 50% of nucleosome positions [10], so it is likely to be an important factor in favouring or disfavouring nucleosome eviction. Histone variant H2A.Z (Htz1) is found to be preferentially enriched at promoters where some nucleosomes have to be quickly removed upon transcriptional activation [3]. Also, acetylated histones are shown to be easily dissociated from DNA [11, 12]. Chromatin remodelling complexes, such as Swi/Snf, act in concert with histone chaperones (e.g Asf1, Nap1) to displace histones from their original positions [3]. Although the complete list of factors has been fairly known, the mechanisms of how they act to mobilize nucleosome are still unclear.
Owing to recent advanced profiling techniques, such as ChIP-Chip and ChIP-Seq, we now have increasing amount of information about how nucleosomes and various kinds of histone modifications are distributed over the genomes of many organisms, including yeast, drosophila, and human [13, 14, 15, 16, 17]. This opens up a chance for thorough

investigation of nucleosome organization, its regulatory mechanisms and functions. Until now, there have been many works, both experimental and computational, concentrating on revealing the effects of factors stated above on nuclesome distribution [13, 10, 18, 19] but most of them have some common drawbacks. First, they mainly considered the effect of each factor separately while bypassing their combinatorial effects on nucleosome distribution. Second, although the distribution of destabilized nucleosomes is usually inhomogeneous throughout the genome and is known to have strong relation with transcriptional activities [13], it is still not well-characterized compared with that of stable nucleosomes.

There are several efforts trying to overcome these limitations. For example, Rippe et al. [20] and Schnitzler [21] investigated co-effects of DNA sequences and chromatin remodelling complexes; Widlund et al. [22] and Yang et al. [23] investigated co-effects of histone tails and DNA sequences on nucleosome distribution. Most of them, however, were based on experimental methods. More recently, Dai et al. [24] used both transcriptional interaction and genomic sequence information to computationally identify dynamic nucleosome distribution, but the number of works like this is still limit.

1.3 Objective and results

Motivated by aforementioned reasons, we propose here a novel method for computationally characterizing nucleosome dynamics from both genomic sequences and histone modification profiles. Our method is based on induction rule learning adapted for subgroup discovery, which can discover sufficiently large and statistically meaningful subsets of population as shown in [25], so it is well suited for characterizing inhomogeneous distribution of destabilized nucleosomes. Moreover, by combining both genetic sequence and histone modification information, our method can discover the combinatorial nature of these two factors in regulating nucleosome stability. Our results on *S.cerevisiae* show that, some DNA motifs, which are reliable determinants for nucleosome forming/inhibiting poten-

tial, and post-translational modifications of histone proteins, which have strong relation with transcriptional activities, are likely to be more significant to nucleosome dynamics. We also found some patterns of cooperation between these DNA motifs and histone modifications in regulating nucleosome stability. Our results give additional insights into mechanisms of how cells regulate important biological processes, such as transcription, DNA repair and replication.

The thesis is organized as following:

- Chapter II provides fundamental knowledge about molecular biology. It presents basic concepts such as DNA, protein, gene transcription, etc. in detail.
- Chapter III introduces current view on chromatin role in gene regulation. Experimental techniques for measuring chromatin state are also presented here.
- Chapter IV presents a novel computational approach to characterizing nucleosome dynamics from both genomic and epigenetic information, as well as some important results while applying this method on *C.cerevisiae* data.
- Chapter V shows concluding remarks and extended direction for the work.

Chapter 2

Fundamental of molecular biology

2.1 Basic concepts

The basic unit of all living organisms is the cell. A *cell* is basically a watery solution of certain molecules, surrounded by a lipid membrane. Typical sizes of cells range from 1μm (bacteria) to 100μm (plant cells). Figure 2.1 illustrates the basic structures of the plant and animal cells. The most important properties of a living cell are the following:

- It consists of a set of molecules that is separated from the exterior (as a human being is separated from his or her surroundings).
- It has a metabolism, that is, it can take up nutrients and convert them into other molecules and usable energy. The cell uses nutrients to renew its constituents, to grow, and to drive its actions.
- It is able to replicate, that is, produce offspring that resemble itself.
- It can react to its environment in a way that tends to prolong its own existence and the existence of a number of offspring.

There are two types of living organisms: *prokarya*, which are always single cells, and *eukarya* (which include all animals, plants, and fungi). Eukaryotic cells are more complex than prokarya in that their interior is more organized: the eukaryote is divided into

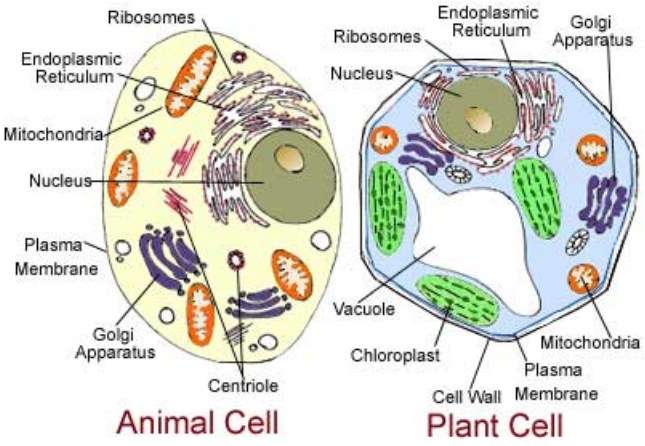

Figure 2.1: The basic structures of animal and plant cells

so-called compartments. For instance, the *nucleus* contains hereditary information, and a number of *mitochondria* serve to supply the cell with certain energy-rich molecules. Especially important for the integrity of cells are three kinds of macromolecules - DNA, RNA and proteins - which will be introduced in detail below. These molecules are *polymers*, which means that they are composed of a large number of covalently linked *monomers*, small molecular building blocks. The set of different monomers and the way they are linked determine the type of polymer.

2.1.1 DNA

The major part of the heritable information of a cell is stored in the form of DNA molecules. They are called the cell's *genome*. *DNA* (*deoxyribonucleic acid*) is a chain molecule that is composed of linearly linked nucleotides. *Nucleotides* are small chemical compounds. There are essentially four different nucleotides that occur in cellular DNA, which are usually called A (adenine), C (cytosine), G (guanine), and T (thymine). The chain of nucleotides has a direction, because its two ends are chemically different. Consequently, each DNA molecule can be described by a text over a four-letter alphabet. Chemists denote its beginning as the *5'-end* and its end as the *3'-end*. The two directions are denoted by *upstream*, for "towards" the beginning, and *downstream*, for "towards"

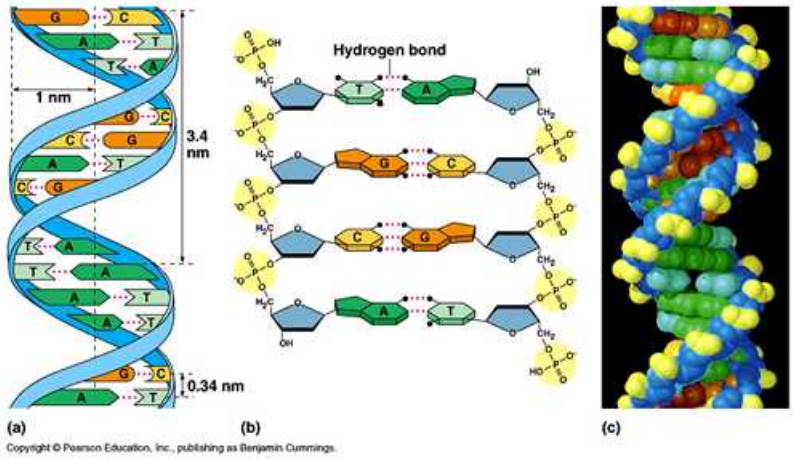

Figure 2.2: DNA double helix structure

the end. Molecular chains of only a few nucleotides are called *oligonucleotides*. DNA is a good carrier of information that is supposed to be retained for a long time. DNA can form very stable structures due to the following properties. The nucleotides A and T can bind to each other by forming two hydrogen bonds; therefore, A and T are said to be *complementary*. G and C are also complementary: they form three hydrogen hybridization bonds. Importantly, the ability to bind in this way holds for chains of nucleotides, that is, for DNA molecules. The *complement* of a DNA sequence is the sequence of the complements of its bases, but read in the reverse direction; complements are often called *complementary DNA (cDNA)*. Complementary strands can bind to each other tightly by forming a double helix structure, which enables all the hydrogen bonds between the pairs of complementary bases. The binding of two complementary DNA molecules is often referred to as *hybridization*. In cells, the genomic DNA is indeed present in the form of a double helix of two complementary strands, as illustrated in Figure 2.2.

Apart from the increased stability, this provides redundancy, which serves the cell in two ways. First, erroneous changes from one nucleotide to another, termed *point mutations*, can thereby be detected and corrected. Second, there is a natural way to duplicate the genome, which is necessary when the cell divides to produce two daughter cells. The double helix is separated into two single strands of DNA, each of which then serves as a

template for synthesizing its complement. Since the complement of a complement of a DNA sequence is again the primary sequence, the above procedure results in two faithful copies of the original double-stranded DNA. The size of genomes can be enormous; for instance, the human genome consists of more than 3 billion nucleotides. Although the human genome is separated into 23 separate DNA molecules, each part still has an average length of about 5 cm - about 5000 times longer than the diameter of a human cell. Consequently, the DNA in cells chromosomes is kept in a highly compact form. In regular intervals, assemblies of proteins (called *histones*) bind to the DNA. The DNA double helix winds about one and a half times around each histone complex to form a *nucleosome*; the nucleosomes resemble beads on a string (of DNA). The nucleosomes themselves are usually packed on top of one another to form a more compact fibrous form called *chromatin*. An even higher level of packing is achieved by introducing loops into the chromatin fiber. The resulting structures, one for each genomic DNA molecule, are known as *chromosomes*. They do not flow around freely in the nucleus, but are anchored to nuclear structures at sites called *matrix attachment regions (MARs)*. In many organisms, two or more versions of the genome may be present in a cell. This is called a diploid or polyploid genome. In contrast, a single set of chromosomes is said to be haploid. In sexual organisms, most cells contain a diploid genome, where one version is inherited from each parent. The germ cells giving rise to offspring contain a haploid genome: for each chromosome, they randomly contain either the maternal or the paternal version (or a mixture thereof).

2.1.2 RNA

RNA (ribonucleic acid) is very similar to DNA: again, it consists of nucleotides linked in a chain. In contrast to DNA, the nucleotide U (for uracil) is used instead of T, and the chemical details of the nucleotides differ slightly. Due to these difference RNA molecules are usually single-stranded, which allows them to form a variety of structures in three-dimensional (3D) space that can perform complex tasks (such RNAs are called *ribozymes*). The importance of the genome is that it typically contains many genes. Although there is still debate about the exact definition, a *gene* can be thought of as

Figure 2.3: Flow of genetic information

a substring of the genome that is responsible for the production of one or a couple of types of RNA molecules. In the process of *gene expression*, the RNA is synthesized to be complementary to a part of the DNA template. As a result, each gene can control one or more properties of the organism, although often quite indirectly, as will become apparent below. Note that genes also include parts of DNA that are not copied into RNA. Most important, each gene contains a sequence called a promoter, which specifies the conditions under which RNA copies of certain parts of the gene are produced. Although ribozymes are responsible for a few very important tasks in cells, the purpose of the vast majority of genes in a cell is to encode building instructions mRNA for proteins. The RNA molecules involved in this process are called *messenger RNAs*, or *mRNAs*. The flow of information from the DNA to the proteins is illustrated in Figure 2.3.

2.1.3 Proteins

Proteins are polymers composed of amino acids. Cells use 20 different types of amino acids for protein synthesis. Common to each amino acid are two chemical groups (an amino [N] group and a carboxyl [C] group) which form *peptide bonds* (a special kind of covalent bond) to link two amino acids. Since a water molecule is split off during the formation of such a bond, a protein is actually composed of *amino acid residues* (often, just *residues*). Proteins are also sometimes called *polypeptides* (most commonly in contexts where their 3D structures are not important); molecules consisting of only a few amino acids are called oligopeptides, or simply peptides. Due to their chemistry, the beginning and the end of a protein are called its *N-terminus* and its *C-terminus*, respectively. The chain of peptide links forms the backbone of a protein. Importantly, each amino acid also has a third group, the *side chain*. The side chains of the 20 natural amino acids show very different

chemical properties. The functions of proteins in cells are as diverse as the tasks that cells have to perform. Functional categories include (but are not limited to) the following:

- *Metabolism*: Proteins called *enzymes* bind small molecules called *metabolites* to catalyze reactions yielding other small molecules. In this way, nucleotides for DNA and RNA, amino acids for proteins, lipids for membranes, and many other essential compounds are produced. Cells may be viewed as tiny but highly complex and competent chemical factories.

- *Energy*: This can be seen as a special case of metabolism, because cells produce a few types of small molecules as energy carriers.

- *Transcription, protein synthesis, and protein processing.* The huge machinery required to produce proper proteins from DNA is, to a great extent, run by proteins (although ribozymes play a crucial role, too).

- *Transport and motor proteins*: Cells can be more efficient due to a nonrandom spatial distribution of molecules. In particular, compartmentalized cells contain elaborate transport mechanisms to achieve and maintain appropriate local concentrations. Molecular motion can even become visible on a macroscopic scale: muscle contractions are driven by the motion of myosin proteins on actin filaments (longish intracellular structures built from actin proteins).

- *Communication (intra- or intercellular)*: Communication is most important for multicellular organisms. While signaling molecules are usually much smaller than proteins, they are received and recognized by proteins. The processing of signals allows computations to be performed; this may be most obvious for the human brain (involving nearly 1011 cells), but also underlies the directed motion of unicellular organisms.

- *Cell cycle*: Most cells (be they alone or part of a multicellular organism) recurrently divide into two daughter cells to reproduce. This complex process is orchestrated and carried out by proteins.

In summary, proteins are major building blocks of the cell and, above all, the machines that keep cells running.

2.1.4 Macromolecules

We have now met the three most important types of macromolecules in the cell (DNA, RNA, and protein) and their relation (the genetic flow of information). A fourth type of macromolecule which also occurs in cells shall Saccharides only briefly be mentioned here: the polysaccharide. *Polysaccharides* are polymers composed of covalently linked *monosaccharides* (sugars, such as glucose, fructose,galactose). In contrast to the macromolecules discussed earlier, their bonding pattern is not necessarily linear, but often rather treelike. Proteins, RNA, and DNA can be parts of even more intricate *assemblies* or, synonymously, *complexes*. For example, as described above, histone proteins are used to pack DNA into chromatin. The ribosome, which performs the translation of mRNAs to proteins, is a huge assembly of several proteins and ribosomal RNA (rRNA). The individual molecules in an assembly (which are not connected by covalent bonds) are referred to as *subunits*. Just to make things more confusing, (stable) complexes of proteins (in the sense of individual translation products, as introduced above) are sometimes also called *proteins*; the subunits are then also called (protein) *chains*.

2.1.5 Metabolites

Of course, small molecules are vital for cells, too. Here we give just a few selected examples:

- Adenosine triphosphate (ATP) and NADPH (both derived from the nucleotide A) serve as ubiquitous ready-to-use sources of energy.
- Monosaccharides (sugars) and lipids (fats) can be converted into ATP, and therefore serve as a long-term source of energy. Saccharides are also often attached to proteins to modify their properties.

- *Signaling molecules* convey information by docking to their respective receptor proteins and triggering their action. For example, steroids (which include many sex hormones) can diffuse into a cell's nucleus and induce the activation of some genes.

Small molecules are more generally called *compounds*.

2.2 Expression of Genes and Proteins

One of the most fundamental processes in the cell is the production (and disposal) of proteins. Below, the life cycle of proteins is outlined for eukaryotic cells.

1. *Transcription*: *Messenger RNA (mRNA)* copies of a gene are produced. The products, called *pre-mRNAs*, are complementary to the DNA sequence.

   (a) Initiation: Certain proteins, called *transcription factors (TFs)*, bind to *TF binding sites* in the gene promoters in the DNA.

   (b) Elongation: The mRNA copy of the gene is synthesized by a special protein (RNA polymerase II). It moves along the DNA and thereby sequentially extends the pre-mRNA by linking a nucleotide complementary to that found in the DNA.

   (c) Termination: A signal in the DNA causes the transcription to end and the mRNA to be released.

2. *Splicing*: Parts of the pre-mRNA, which are called *introns*, are removed. The remaining parts, called *exons*, are reconnected to form the mature mRNA. The spliced mRNAs travel from the nucleus (through huge, selective pores in its double membrane) into the cytosol. To increase the chemical stability of the mRNA, a chemical cap is formed at the 5'-end and a *poly(A)* sequence (built from many A nucleotides) is appended to the 3'-end.

3. *Translation*: In the cytosol, ribosomes await the mRNAs. Ribosomes synthesize proteins as specified by *codons* - triplets of consecutive nucleotides - in the mRNA.

(a) Initiation: The ribosome finds a *start codon* (usually, the first AUG subsequence that has favorable neighboring nucleotides) in the mRNA.

(b) Elongation: One by one, the ribosome attaches amino acids to the growing polypeptide (protein) chain. In each step, the ribosome translates the current codon into an amino acid according to the *genetic code*. The ribosome then moves to the next codon in the same *reading frame*, that is, to the next adjacent nonoverlapping codon.

(c) Termination: Translation is stopped by any of three different *stop codons* encountered in the current reading frame.

4. *(Posttranslational) modification* (PTM). The protein may be chemically modified, if it contains the relevant signals and if it resides in a compartment where these signals are recognized.

   (a) Additional chemical groups can be covalently attached to proteins (glycosylation (sugars), phosphorylation, methylation, etc).

   (b) Covalent bonds can be formed between amino acids.

   (c) Proteins can be covalently bound to each other.

   (d) Proteins can be cleaved, that is, cut into parts.

5. *Translocation*: Proteins are delivered to the appropriate compartment, which is specified by signals in the amino acid sequence. The signal can either be a typical short segment of a sequence, or a structural motif on the surface of the protein (which may be composed of amino acids that are not neighbors in the sequence). In the absence of signals, the protein stays in the cytosol.

6. *Degradation*: Almost all proteins are eventually destroyed by digestion into their individual amino acids.

In prokaryotes, the entire process is a bit less complex because splicing is uncommon and the translocation has only three different targets (cytosol, membrane, exterior) due

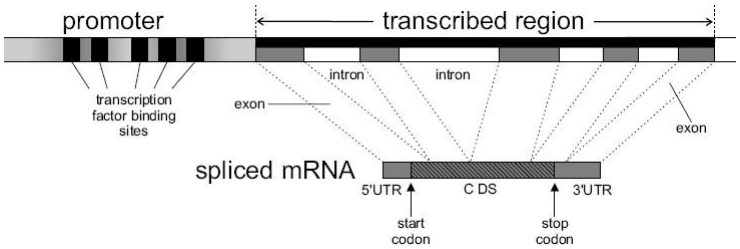

Figure 2.4: Eukaryotic gene structure

to the lack of compartments.

The process of splicing implies complex *gene structures* composed of alternating introns and exons; an illustration is given in Figure 2.4. However, it allows for splicing increased flexibility by a mechanism known as *alternative splicing*: certain proteins can cause certain exons to be lengthened, shortened, or even skipped completely. Thus, the same gene can give rise to the production of different proteins. This is an important way for cells to adapt to the circumstances, including their cell type and extracellular signals. It is estimated that a human gene on average encodes for eight or nine different proteins.

The process described above is called *gene expression* (illustrated in Figure 2.5). The term *expression level* of a molecule type is used to refer to either its current abundance in the cell, or to the rate of synthesis of new molecules. This difference is often neglected for gene expression, which may or may not be justified by the fact that mRNAs are degraded relatively quickly after having been translated several times. However, for proteins the distinction is crucial, because their lifetimes may be very long and differ vastly.

The cellular concentration of any type of protein can be influenced by changing the efficiencies of the above steps. This is called *regulation* of expression. While cells in fact regulate each of the above steps, the main point for the quantitative control of protein expression is certainly transcription initiation. In addition to the general TFs, which are always required for initiation, there are additional TFs which modify the probability or speed of transcription. They bind to short DNA motifs, for obvious reasons called *enhancers* and *silencers*. The effects of TF binding sites can extend over huge distances in the DNA sequence; therefore *insulators* (certain DNA signals) may be required to sepa-

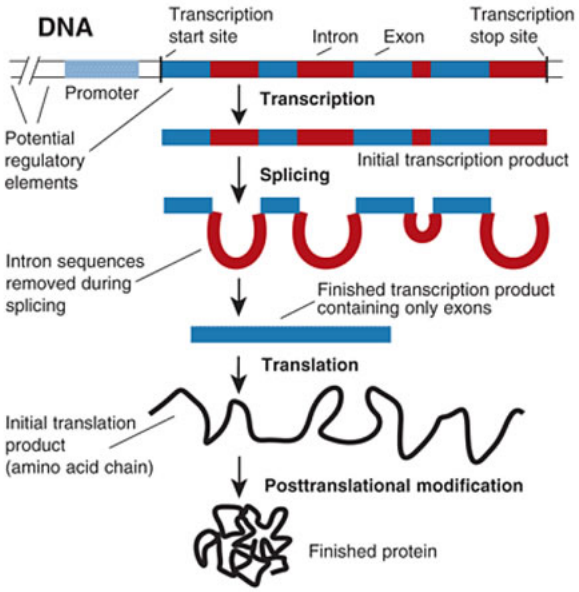

Figure 2.5: Gene expression process

rate genes from each other and prevent mutual regulatory interference.

The steps of protein expression have a natural temporal ordering, where each step operates on the result of the preceding step. However, there are at least three types of deviation from a clear, serial manufacturing process: (1) Some of the steps may occur concurrently, or can be performed before the preceding step is finished. For example, much of the splicing is carried out while the gene is still being transcribed. Also, the translocation from the cytosol into the *endoplasmatic reticulum* (ER) and some modifications take place during translation. (2) There is no compulsory ordering of translocation and modification. In fact, many proteins are modified in the ER and the Golgi apparatus, which are intermediate stations on the journey to their destination compartment. (3) Degradation may occur even before the protein is finished and delivered.

Chapter 3

Chromatin and gene regulation

3.1 Basic unit of chromatin: the nucleosome

3.1.1 How DNA is packaged in the nucleus

All organisms, prokaryotic or eukaryotic, must deal with the problem of packaging a relatively long piece of DNA into a small space within the cell. In eukaryotes, DNA is sequestered in a particular subcellular organelle, the nucleus. The size of the problem was illustrated by scaling things up so that the nucleus was the size of a large grapefruit (ie. about 10 cm in diameter). On this scale, the DNA molecules to be packaged in a typical eukaryotic cell would, in total, be about 20 km long. Admittedly the DNA is thin, even on this scale (about 0.02 mm), and the volume of the container is more than sufficient to accommodate it, but the problem is not a simple packaging one. The overriding requirement is that the DNA, once in place, must be able to function. It must be replicated, with complete accuracy just once (and no more than once) each cell cycle, the two copies must be separated into the two daughter cells each time the cell divides and the genetic information encoded by the DNA must be expressed in a way that is appropriate to the particular cell at the particular stage of development that it has reached.

It seems inevitable that the mechanisms by which DNA is packaged into the nucleus will have a major effect on its function. It is also worth remembering that solutions to the

structural and functional aspects of the DNA packaging problem must have coevolved so as to accommodate the differing requirements of each. Such coevolution will lead to mechanistic links between the two processes. Moreover, the two major components of DNA function, namely replication and transcription, are also likely to have become increasingly interlinked as a result of adjustments, refinements and compromises during evolution.
In 1974, Roger Kornberg proposed an elegantly simple model for the structure of chromatin (depicted in Figure 3.1). This model has provided the basis for our understandings on chromatin structure and function ever since. He suggested that:

- Chromatin consists of a fundamental repeating unit made up of 200 base pairs of DNA and two each of the four histones H2A, H2B, H3 and H4; i.e. the histones formed an eight-subunit (octameric) structure.

- A chromatin fibre consists of many such nucleosomes forming a flexibly jointed chain. This corresponds to the beads-on-a-string fibres.

3.1.2 The structure of the nucleosome

Analysis of histone-histone interactions within the nucleosome showed that the eight histones were all connected in a single particle (the histone octamer) that could be dissociated into an $(H3-H4)_2$ tetramer and two H2A-H2B dimers. However, it was not immediately obvious exactly how the DNA was organized in relation to the histones.
The first experimental clues came from nuclease digestion studies when a careful analysis of how nucleosomal DNA was digested by the endonuclease DNaseI led to the proposal that the DNA was coiled around the outside of the histone octamer. The conclusions drawn from the nuclease digestion studies were confirmed by the completely independent technique of neutron scattering. This procedure can distinguish signals generated by DNA and protein, and studies of nucleosomes in solution confirmed that nucleosomal DNA is indeed wrapped around the outside of a histone core.

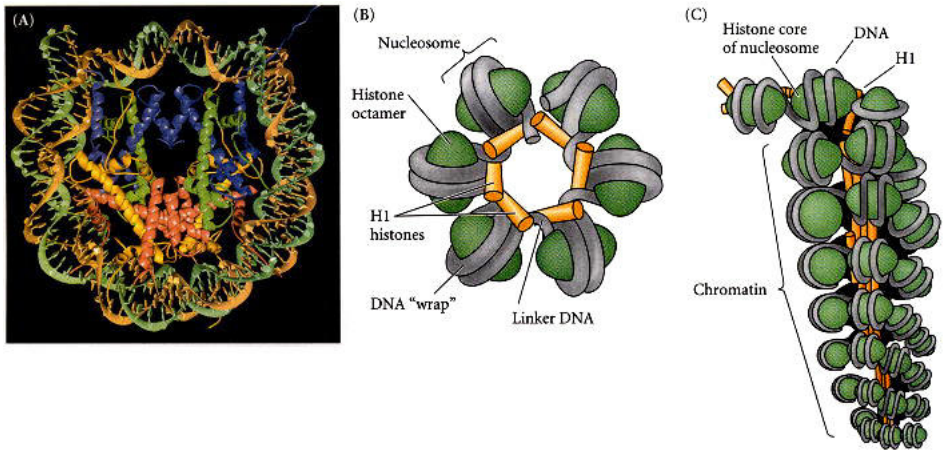

Figure 3.1: Nucleosome and chromatin structure

3.2 Histone: modifications and epigenetic information

In previous part, we considered nucleosome core particle as the basic structural unit of chromatin and saw its role in DNA packaging. However, it is not the only role of the nucleosome. The particle has a second function that gives it an importance far beyond its initial packaging role. This is its ability to carry epigenetic information, the information that is itself not encoded by DNA sequence.

3.2.1 Histones and modifications of their tails

Each histone core has the regions of up to 25 or so amino acids at the amino-terminal (N-terminal) ends, so-called histone tails. These tails are exposed on the surface of the nucleosome and do not adopt fixed structures in core particle crystals. As with the core histones in general, the amino acid sequences of the tail domains have been highly conserved through evolution. However, the histone tails, unlike the globular, core domains, are subject to a wide variety of enzyme-catalysed, post-translational modifications. Specific amino acid side-chains can be modified by attachment of specific chemical groups (e.g. phosphate or acetate), changing both their charge and conformation. So, despite

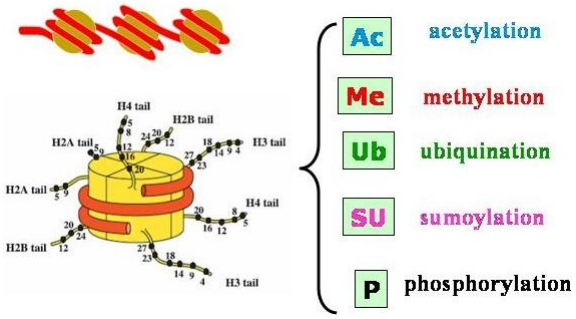

Figure 3.2: Histone tails and modifications

their evolutionary conservation, the tail domains show enormous variability from one core particle to another in the spectrum of post-translational modifications that they carry. These modifications constitute a major source of epigenetic information (as illustrated in Figure 3.2).

A second clue about the possible role of the tails comes from the fact that many of the amino acids within the N-terminal histone tails can be modified by specific enzyme activities in vivo. This became clear as a result of some of the very first sequencing studies, which showed that particular lysine residues in the tail domains of H3 and H4 were often modified by attachment of an acetate group. Subsequent studies have shown that the tail domains of the core histones can also be modified by phosphorylation, methylation and ADP-ribosylation. The shorter, exposed C-terminal domains of H3 and H2A are not modified, with one exception.the attachment of a small peptide, ubiquitin, to lysine residue 119 of H2A. Interestingly, this residue falls just within the trypsin-sensitive (i.e. exposed) C-terminal region of H2A. Each of these modifications is carried out, and reversed, by specific enzymes or families of enzymes. These modifications, like the histones themselves, have been conserved through evolution. For example, all species that have been tested so far have the ability to acetylate histone H4. From an experimental point of view, these modifications make a significant contribution to the heterogeneity of nucleosomes.

3.2.2 Histone variants

In higher eukaryotes, histone genes are present in multiple copies. Usually, the genes for the five different histones are present in a cluster, which is then repeated many times. Only in this way can the cell provide the vast numbers of histones needed to package newly replicated DNA during S-phase. Expression of histone genes is at its highest during S-phase, with only residual expression during the rest of the cell cycle, providing sufficient histone to cope with the demands of DNA repair and associated chromatin remodelling activities. Most copies of the genes encoding any given histone are the same. However, for all the histones apart from H4, there are variant genes encoding histones with differences in amino acid sequence.

In some cases these differences are relatively subtle, comprising amino acid substitutions that have little or no apparent effect on the properties or function of the histone. For example, there are three H3 variants in mammals, H3.1, H3.2 and H3.3. H3.2 differs from H3.1 by a single amino acid substitution (serine for cysteine at position 96), while H3.3 has two additional substitutions. These substitutions are enough to cause major changes in the mobility of the three variants, but they seem to behave equivalently in most experimental systems (e.g. in the frequency with which they are acetylated or methylated). However, whereas H3.1 and H3.2, like most histones, are synthesized only during S-phase, H3.3 is synthesized throughout the cell cycle, so the variants are not equivalent in all respects.

3.3 Chromatin role in gene regulation

The intimate association of histones with DNA and the various levels of higher-order DNA packaging all influence the binding to DNA of transcription factors and other components of the transcription complex. Chromatin is something that the transcription machinery must have learned to get along with from the very earliest stages of eukaryotic evolution. These simple considerations bring us to a more difficult problem, namely, the need to establish the extent to which chromatin is an integral and necessary component of tran-

scriptional control mechanisms. The significance of this problem may become clearer if two extreme views of the role of chromatin are considered. One holds that chromatin is primarily a DNA packaging device; it constitutes an obstacle that the transcription machinery must overcome and its effects on gene expression, although unavoidable, are essentially passive. The other holds that chromatin is an integral component of mechanisms of transcriptional control and that this role has evolved in parallel with its packaging function; it plays an active role in control of gene expression. Specific effects of chromatin can result from either active or passive mechanisms. Unravelling a mechanism that involves chromatin as an active participant will lead to useful insights into how genes are regulated.

It is also worth mentioning that, in some cases, chromatin can exert a positive effect on gene expression. It can do this by folding DNA in such a way as to bring together protein binding sites and thereby facilitate useful protein-protein interactions. A typical mammal has tens of thousands of genes in each of its cells, and each one of these genes must be regulated in a way appropriate to the needs of the cell. This is not to say that there are likely to be tens of thousands of different mechanisms of gene regulation, but nor will there be just one. Even in a single-celled eukaryote such as yeast, some genes change their levels of activity through differentiation, the cell cycle or in response to environmental changes while others are expressed in all cells for most or all of the time. In multicellular organisms even more regulatory demands must be met. Chromatin is an essential element in the mechanisms used to address these demands, but the ways in which it is used will differ depending on the nature of the problem that has to be solved.

3.3.1 Transcription on "naked" DNA

The principles and mechanisms underlying transcription on naked DNA are remarkably similar between eukaryotes and prokaryotes despite the increased complexity of eukaryotic transcription machinery. The typical RNA polymerase II (Pol II) transcription cycle begins with the binding of activators upstream of the core promoter (including the TATA box and transcription start site). This event leads to the recruitment of the adaptor

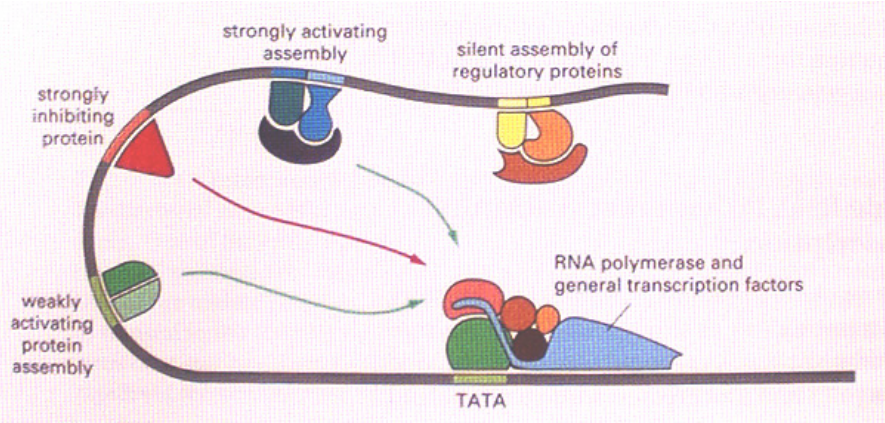

Figure 3.3: Transcription on "naked" DNA

complexes such as SAGA or mediator, both of which in turn facilitate binding of general transcription factors (GTFs). Pol II is positioned at the core promoter by a combination of TFIID, TFIIA, and TFIIB to form the closed form of the preinitiation complex (PIC). TFIIH then melts 11.15 bp of DNA in order to position the single-strand template in the Pol II cleft (open complex) to initiate RNA synthesis. This process is illustrated in Figure 3.3. The carboxy-terminal domain (CTD) of Pol II is phosphorylated by the TFIIH subunit during the first 30 bp of transcription and loses its contacts with GTFs before it proceeds onto the elongation stage. Meanwhile, the phosphorylated CTD begins to recruit the factors that are important for productive elongation and mRNA processing.

3.3.2 Transcription in chromatin environment

- **Transcription Factor Recruitment**

  Eukaryotic and prokaryotic TFs share universal properties in targeting and binding to sequence-specific binding sites in the context of free DNA. However, when recognition sites are buried in chromatin, eukaryotic TFs have to exploit various strategies to achieve proper binding. Early biochemical experiments suggested that TFs can bind to nucleosomal DNA in a cooperative manner. This has been confirmed by *in vivo* studies showing that activator Pho4 can bind to the PHO5 promoter before

nucleosome disassembly. Numerous examples have made it apparent that chromatin remodeling complexes can stimulate binding of TFs to nucleosomal sites.

In different studies TF-binding sites have been mapped either to the nucleosome-free region or within a nucleosome. Recent genome-wide studies found that nucleosome density at promoter regions is typically lower than that in the coding region. The earlier analytical studies and the recent rigorous mathematic modeling led to the hypothesis that organizational information for positioning nucleosomes is embedded within the sequence of the genome. Remarkably, the models predict that there is low-level nucleosome occupancy at functional TF-binding sites and that there are more stable nucleosomes at the nonfunctional sites. Therefore, it seems that eukaryotic cells tend to position sequence-specific TF-binding sites within accessible regions. Thus, the first step of gene activation (activator binding) could be more responsive to signaling pathways than it would be if the binding sites were sequestered within nucleosomes. However, this oversimplified view apparently cannot account for all activator binding in vastly diverse genomes. In a large-scale screen of the human genome, high levels of histone H3K4/79 methylation and H3 acetylation were found to be strict prerequisites for binding of the Myc transcription activator, which implies that chromatin modifications can actually regulate TF binding.

- **Transcription Initiation**

  Once activators bind to the promoter, they trigger a cascade of recruitment of coactivator complexes (Figure 3.4). Coactivators (such as chromatin-remodeling complexes, histone-modification enzymes, and mediator) not only facilitate stronger binding of activators to DNA but also make nucleosomal DNA elements more accessible to GTFs. How do cells adjust chromatin structure to accommodate the proper docking of the massive PIC and its ancillary factors?

  Historically, increased histone acetylation at the promoter region has been linked to active transcription. Recently, a research using high-resolution tiling microarray demonstrated that acetylation of H3 and H4 peaks sharply at active yeast promoters and that, when normalized to nucleosome density, the level of acetylation is pro-

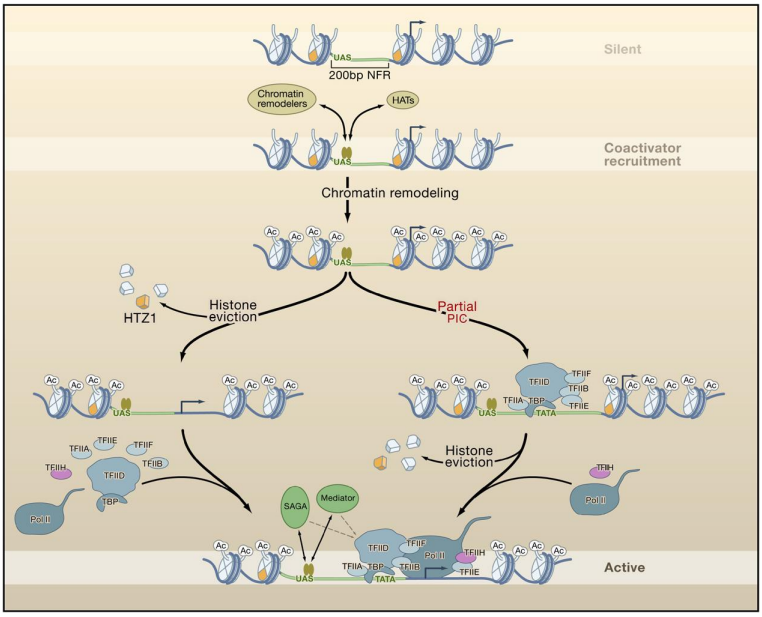

Figure 3.4: Models of Chromatin Regulation during Transcription Initiation

portional to the transcription rate. In addition, elegant biochemical and genetics studies provide further mechanistic support for such a notion. SAGA is recruited to the promoter through direct interaction between its Tra1 subunit and a bound activator. SAGA recruitment and histone acetylation occur prior to PIC formation at the GAL1 promoter. Moreover, to make DNA more accessible, promoter-bound activators also target chromatin remodeling complexes such as Swi/Snf. Interestingly, although the sequence of events leading to recruitment of HATs and chromatin remodelers by the same activators is dependent on their promoter context, their recruitment occurs in a coordinated manner.

Considering the amount of DNA directly contacted by Pol II/GTFs, the structure of the nucleosome seems to pose a significant obstacle to PIC formation. Indeed, it is clear from both ChIP and topological studies that histones are lost at the yeast PHO5 and HSP82 promoters upon gene activation and that nucleosomes are reassembled as a gene turns off. A genome-wide survey has found that a large number of promoters partial PICs, including TFIIA, TFIID (and/or SAGA), TFIIB,

TFIIE, and TFIIF, were assembled, whereas in these cassRNAPol II and TFIIH are generally not present (Figure 3.4, right). Remarkably, in this case, nucleosomes are not displaced, thus implying that engaging template DNA into the Pol II active site might create a reasonable point where DNA-histone contacts must be broken. This is reminiscent of a previous observation where Pol II itself was found to be required for chromatin remodeling at the RNR3 promoter.

The histone variant H2A.Z (Htz1) is preferentially enriched at promoters that are poised for transcription activation. High-resolution mapping reveals that two well-positioned Htz1-containing nucleosomes flank a 200 bp nucleosome-free region (NFR). Htz1-containing nucleosomes are resistant to transcription elongation-related modifications and to chromatin remodeling. In addition, Htz1 is easily dissociated from nucleosomes, presumably as a dimer with H2B. Upon transcription activation, however, Htz1 is rapidly evicted from the promoter, and its loss is required for full transcription. Therefore, Htz1 is specifically positioned at the promoter, where some nucleosomes have to be removed to accommodate PIC formation. However, it should be noted that although there is solid evidence for histone loss, the promoter is not completely nucleosome free. Acetylated histones H3 and H4 continue to accumulate during gene activation, and Htz1 K14 is acetylated at active promoters. Hence, the reason for Htz1 removal might be to make room for the mobilization of residual nucleosomes. For example, at the IFN-b promoter, sliding of a nucleosome upon TBP binding is indeed beneficial to transcription. A second reason would be to make the underlying DNA completely accessible.

- **Transcription Elongation**

  Transcription elongation begins when Pol II releases from GTFs and travels into the coding region. This event signals the recruitment of the elongation machinery, which includes the factors involved in polymerization, mRNA processing, mRNA export, and chromatin function. At this point, one might expect that Pol II would deal with the downstream nucleosomes in a similar manner. However, the opposite is true. Cells exploit a very sophisticated array of factors to control chromatin architecture

during elongation, and the events and factors required at the beginning of the gene differ significantly from those required at the end. This is done not only to promote efficient RNA synthesis but also to ensure the integrity of the chromatin structure while Pol II travels through the body of the gene.

3.4 Biological measurement of chromatin state

The combination of *chromatin immunoprecipitation* (ChIP) and DNA microarrays, a technique that is known as genome-wide location analysis or ChIP-Chip, marked the beginning of an era of rapid progress in highthroughput studies, with studies of chromatin modifications being no exception. Although ChIP-Chip was first used to map DNA-binding proteins on a genomewide scale, it did not take long before it was applied to map other phenomena globally, such as histone modifications and nucleosome distribution (or *nucleosome positioning*).

In the past few years, various sequencing-based protocols have been developed to analyse ChIP samples. Most of them combine ChIP with *serial analysis of gene expression* (SAGE). The recent combination of ChIP with massively parallel sequencing (ChIP-Seq) allows researchers to survey more of the genome in less time and promises to unveil new aspects of biology in the coming years.

Application of these techniques has led to great advances in our understanding of how epigenetic phenomena are regulated and how they affect gene expression. This part focuses on the technical aspects of genome-scale approaches to study epigenomes and their application to profiling histone modifications, nucleosome positioning.

3.4.1 Genome-scale approaches to studying histone modifications

The most prevalent technique used to map histone modifications at a genomic scale has been the combination of ChIP with DNA microarrays (ChIP-Chip). The ChIP-Chip method can be used to study many of the epigenomic phenomena. The example pre-

sented in Figure 3.6 shows how ChIP-Chip can be used to study histone modifications. Modified chromatin is first purified by immunoprecipitating crosslinked chromatin using an antibody that is specific to a particular histone modification (shown in green). DNA is then amplified to obtain sufficient DNA. The colour-labelled ChIP DNA, together with the control DNA prepared from input chromatin and labelled with a different colour, is hybridized to a DNA microarray. The microarray probes can then be mapped to the genome to yield genomic coordinates. Briefly, chromatin fragments are isolated using antibodies that are specific to a feature of interest and the isolated fragments are amplified to generate micrograms of fluorescently labelled DNA; this is followed by hybridization to DNA microarrays. The first ChIP-Chip studies of histone modifications in *S.cerevisiae* and *Drosophila melanogaster* suggested that histone modifications are associated with distinct genomic regions and with distinct transcription states. These studies were followed by other ChIP-Chip studies with higher resolution tiling arrays in yeast that further reinforced the concept of redundancy in histone-modification maps. ChIP-Chip has also been used to profile histone modifications in mammalian genomes.

Another high-throughput technique that combines ChIP with SAGE is GMAT, which is also known as ChIP-SAGE (Figure 3.7). Here, ChIP is carried out and is followed by SAGE. Short sequence tags of 21 bp are extracted from the sequencing library and mapped to a reference genome. The number of tags that are detected at a genomic region directly correlates with the modification level of the region. Since there is no probe hybridization involved in the process, the results obtained from GMAT might be more quantitative than ChIP-Chip, though these two techniques have not been directly compared.

ChIP-Seq is a recently developed technique for analysing ChIP DNA using a high-throughput massively parallel signature sequencing-like technique developed by Solexa (Figure 3.8). Briefly, the ChIP DNA is ligated to a pair of adaptors and subjected to very limited amplification to generate 200 ng of DNA. It is then bound by hybridization on a solid surface to covalently bonded oligos that are complementary to the adaptor sequences. A

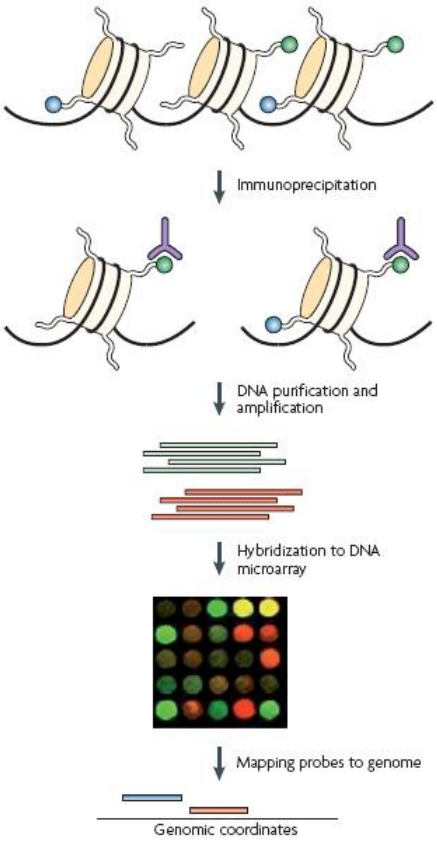

Figure 3.5: Chromatin immunoprecipitation combined with DNA microarrays (ChIP-Chip)

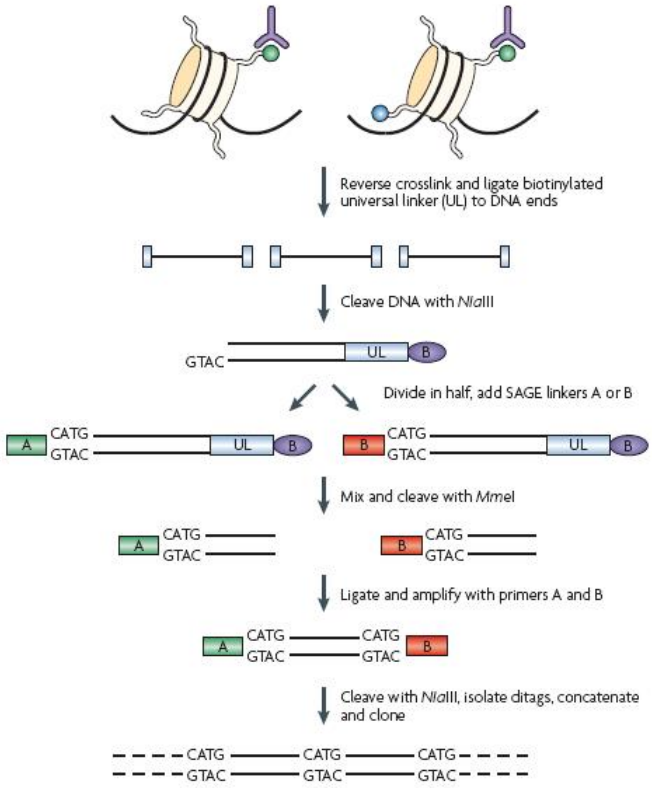

Figure 3.6: Chromatin immunoprecipitation combined with serial analysis of gene expression (ChIP-SAGE)

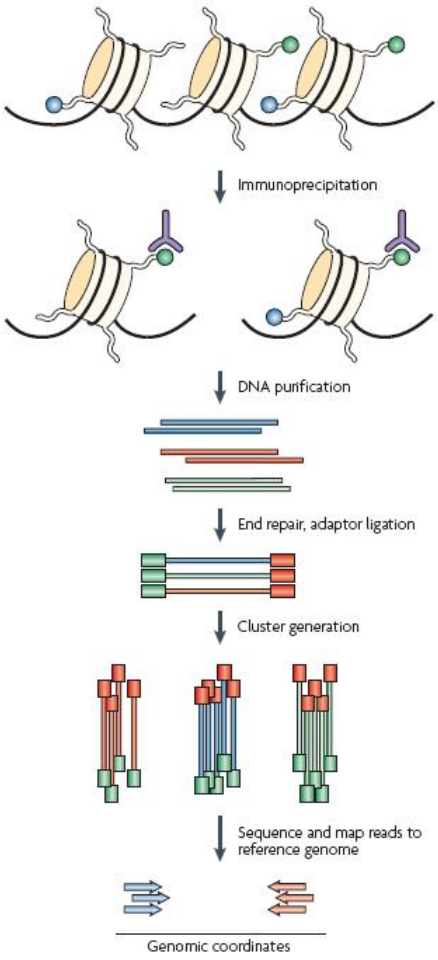

Figure 3.7: Chromatin immunoprecipitation combined with high-throughput sequencing techniques (ChIP-Seq)

short sequence (25-50 bp) for each of the 30-60 million DNA templates is then determined from its end by 'sequencing-by-synthesis', which is a modified Sanger sequencing procedure. The first applications of ChIP.Seq to profile histone modifications were done in CD4+ T cells and mouse embryonic stem (ES) cells. The number of sequenced reads that are mapped to a genomic locus is directly proportional to its modification level. Because ChIP-Seq requires less PCR amplification and does not depend on the efficiency of probe hybridization, in contrast to ChIP-Chip, it is probably more quantitative and the modification levels that are obtained in ChIP-Seq experiments at different genomic regions can be directly compared.

3.4.2 Nucleosome positioning

The positioning of nucleosomes with respect to DNA can directly influence gene regulation. In recent years, several genome-wide maps of nucleosome positions in yeast, worm and across all human promoters have emerged. Most of these studies have taken advantage of the preferential cleavage of linker DNA over nucleosomal DNA by MNase. The mononucleosome-sized DNA that is isolated from MNase-digestion is analysed by either tiling microarrays containing overlapping probes or high-throughput sequencing. ChIP-Seq data for certain histone modifications can also be used to map nucleosomes in certain regions of the genome.

Chapter 4

A computational approach to characterizing nucleosome dynamics

4.1 Methods

4.1.1 Data preparation

We used experimental data from Yuan et al. [13] and Liu et al. [16], which covered nearly 4% of yeast genome including chromosome III and 223 additional promoter regions, for our experiments. Data from Yuan contained 50-base DNA fragments tiled every 20 base pairs, and for each fragment we extracted its genomic sequence and HMM inferred state showing that it is nucleosomal sequence or not. Data extracted from Liu contained 12 different histone modification levels corresponding to DNA fragments above, including acetylations of H3K9, H3K14, H3K18, H4K5, H4K8, H4K12, H4K16, H2AK7, H2BK16 and mono-, di- and tri-methylations of H3K4. To investigate whether there exists any difference in characteristics of nucleosome dynamics between regulatory regions and genomic regions, we separated the data above into two datasets, corresponding to chromosome III and promoter regions. For each dataset, we filtered out data of linker regions to keep only nucleosomal data. Each nucleosome was assigned either as *Well-positioned* if it stretched from 6 to 8 fragments or as *Delocalized* if it stretched more than 9 fragments.

Nucleosomes which had no histone modification values or delocalized nucleosomes whose lengths were longer than 350 base pairs were also treated as noise and removed. After these preprocessing steps, the dataset of chromosome III contained 997 well-positioned nucleosomes and 154 delocalized nucleosomes, the dataset of promoter regions contained 995 well-positioned nucleosomes and 69 delocalized nucleosomes. These two datasets were used for further analysis.

4.1.2 Method overview

In this work we aim at characterizing how DNA sequences and histone modifications affect nucleosome dynamics. To this end, we propose a novel method that takes significant DNA motifs and histone modifications along with nucleosome states as the input for the rule induction system to infer patterns which may represent the dependence of nucleosome stability on these two factors. Figure 4.1 depicts the overview of our method. At first, DNA motifs, which might be significantly related to nucleosome stability, were extracted from nucleosomal sequences by applying two different approaches. The first one was to find potentially conserved motifs related to nucleosome states using WordSpy, the software that has been shown to outperform other competing motif finding methods on benchmark datasets. The second one was to find motifs which could serve as discriminative information for two states of nucleosomes using feature selection function of Gist software package [28]. Motifs were ranked based on their important levels identified by Fisher criterion. Significant histone modifications were also extracted by applying the same feature selection procedure as the second approach above. We then constructed a decision table from these significant DNA motifs and histone modifications (see Figure 4.1) and used it as the input for CN2-SD rule induction system (Section *Rule learning*) to produce a set of rules. Some filtering procedures were applied to remove uninteresting rules and keep rules which may meaningfully characterize nucleosome dynamics.

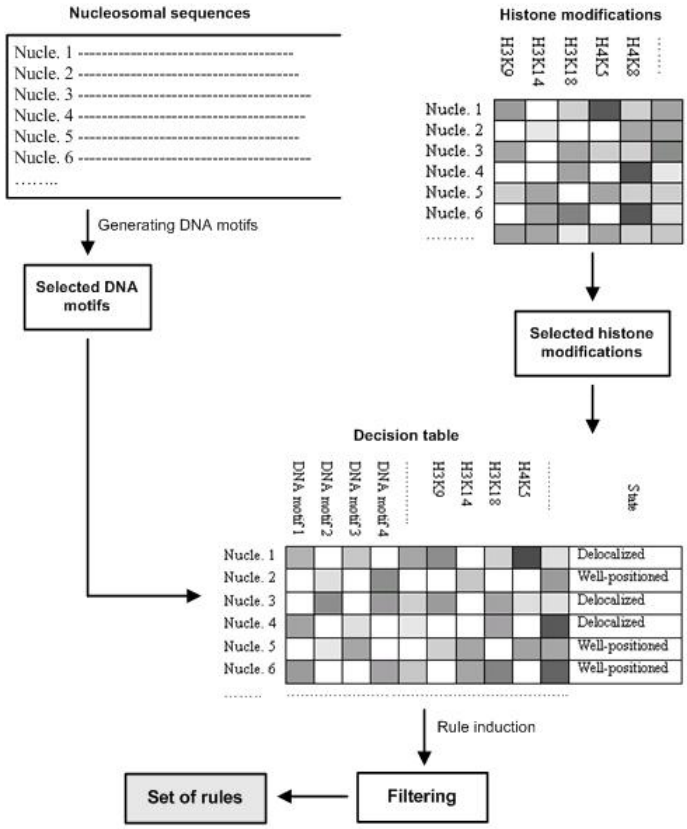

Figure 4.1: Method overview

4.1.3 Feature selection with Fisher criterion

Feature selection is a process of selecting a subset of relevant features available from the data that most contribute to distinguishing instances from different classes. In our method, significant sequence and histone modification features related to two states of nucleosomes, *Well-positioned* and *Delocalized*, were identified and ranked by their Fisher scores (or F-score in short). This is one of statistical criteria that is simple, effective and independent of the choice of classification method. Because our method only concentrated on identifying features with highly discriminative strength instead of building any concrete classifiers so we chose F-score as the selection criterion. The discriminative strength of each feature is defined as following:

Given a dataset X with two classes, denote instances in class 1 as X^1, and those in class 2 as X^2. Assume $\overline{x}_j^k$ is the average of the jth feature in X^k, the F-score of the jth feature is:

$$F(j) = \frac{\left(\overline{x}_j^1 - \overline{x}_j^2\right)^2}{\left(s_j^1\right)^2 + \left(s_j^2\right)^2} \tag{4.1}$$

where

$$\left(s_j^k\right)^2 = \sum_{x \in X^k} \left(x_j - \overline{x}_j^k\right)^2 \tag{4.2}$$

The numerator indicates the discrimination between two classes, and the denominator indicates the scatter within each class. The larger the F-score is, the more likely this feature is more discriminative.

4.1.4 Rule learning

We consider this problem as a subgroup discovery problem and use a rule-based learning method for inducing rules. The problem of subgroup discovery can be defined as follows: given a population of individuals and a property of them, we are interested in finding population subgroups that are interesting with respect to the property of interest [25]. The induced rules usually have the form $Cond \rightarrow Class$, where *Class* is a value of the property of interest, and *Cond* is a conjunction of attribute-value pairs selected from the

features describing the training instances. In our work, *Class* has two values, *Delocalized* and *Well-positioned*. Attributes are significant histone modifications and DNA motifs as described above (Section *Method overview*).

Among several available rule induction systems, CN2 is a rule induction system implementing the separate-and-conquer strategy [27]. It learns a rule set by iteratively adding rules one at a time. Examples covered by a rule are removed from the search space before learning the next rule to add to the rule set. This is repeated until all examples are covered by at least one rule in the rule set or some stopping criteria is satisfied. Finally, CN2 can induce a set of independent rules, where each rule describes a specific subgroup of instances. However, CN2 only induces the first few rules discovered are usually interesting. Subsequently induced rules are obtained from biased example subsets, i.e., subsets including only positive examples that are not covered by previously induced rules. In 2004, Lavrac and her colleagues developed an improvement of CN2 for subgroup discovery, so-called CN2-SD [25]. The CN2-SD generalizes the covering algorithm by introducing example weights. Initially, all examples have a weight of 1.0. However, the weights of examples covered by a rule will not be set to 0 (they are not removed as in CN2), but instead will be reduced by a certain factor. The resulting number of rules is typically higher than with CN2, since most examples will be covered by more than one rule. CN2-SD is, therefore, better in learning local patterns, since the influence of previously covered patterns is reduced, but not completely ignored. In order to evaluate the rules with higher generality, CN2-SD also uses a weighted relative accuracy heuristic as presented in Equation 4.3. The weighted covering strategy tends to find rules that explain overlapped subgroups of instances in the search space, so the weighted relative accuracy heuristic produces highly general rules that express the knowledge contained in one specific subgroup. For these reasons, we utilize the CN2-SD in the rest of this paper for finding rules.

$$h_{WRA}(Cond \rightarrow Class) = \frac{p(Cond)}{p(Class|Cond) - p(Class)} \tag{4.3}$$

4.1.5 Rule filtering

Though the CN2-SD rule induction system uses a weighted covering strategy to restrict the redundancy of learned rules and guarantee the scanning of the whole search space, uninteresting rules are still produced [25]. Let us assume that our rule r has a form: IF [$Cond$] $THEN$ [$ClassDistribution$]. Where

$$\begin{aligned} Cond &= motif_1 = motifVal_1 \wedge \ldots \wedge motif_m = motifVal_m \wedge \\ & \quad histoneMod_1 = hisVal_1 \wedge \ldots \wedge histoneMod_n = hisVal_n \end{aligned}$$

with $motif_i$ is a DNA motif, $motifVal_i$ is *enriched* or *low*, $histoneMod_j$ is one kind of histone modification and $hisVal_j$ is *hyper* or *neutral* or *hypo*; $ClassDistribution = [p, q]$ with p and q are the number of *Well-positioned* and *Delocalized* nucleosomes covered by r, respectively. We used several heuristics to filter out unexpected rules: rules that cover less than 2 positive examples or $p/(p+q) < 0.8$ if positive class is *Delocalized* and rules that cover less than 10 positive examples or $q/(p+q) < 0.8$ if positive class is *Well-positioned* (Positive class is the class characterized by the rule).

4.2 Results and discussion

4.2.1 Potentially significant motifs to nucleosome dynamics

DNA sequence has long been known to be a strong determinant for nucleosome formation potential, which can be used to identify nearly 50% of positioned nucleosomes *in vivo*, so it is likey to be an important factor affecting nuclesome stability. To determine DNA motifs which may be importantly related to nucleosome stability, two different approaches were applied (Section *Method overview*). In the first one, we used WordSpy [29] with the word length set to 6 to identify statistically significant motifs related to nucleosome states. The length of 6 was chosen because, as shown in some previous research [10, 18], nucleosome forming ability of DNA sequences may be decided mostly by short motifs, with length from 2 to 6. WordSpy uses dictionary-based approach so it is suitable to find short motifs among

a group of DNA sequences [26]. Tables 4.1 and 4.2 show the 15 most significant motifs related to 2 states of nucleosomes found by WordSpy when run on chromosome III and promoter region data, respectively. The results show no big difference between important motifs of genetic regions and those of promoter regions. For example, both of them are enriched of dinucleotides TG/CA and this coincides with previous research [18], showing that TG/CA are highly flexible dinucleotides so they have large impact in imparting nucleosome forming ability. From the results given by WordSpy, it is difficult to identify motifs that may be important in discriminating nucleosome states. So, we used the second approach based on feature selection with Fisher criterion (Section *Feature selection with Fisher criterion*) to overcome this limitation. Table 4.3 shows 20 strongest discriminative motifs corresponding to chromosome III and promoter regions ranked by their F-score values. Among them, dinucleotides are likely the most important motifs compared with the others in deciding nucleosome stability: 14 and 15 over 20 in chromosome III and promoter sequences, respectively. Moreover, among 10 strongest discriminative signals are AA/TT/AT/TA/CA/TG (for chromosome III) and AT/TT/CA/TG (for promoter regions), which are related with nucleosome forming (e.g. CA/TG) and inhibiting (e.g. AA/TT/AT/TA) potential of DNA sequences.

4.2.2 Significant histone modifications to nucleosome dynamics

Histone modification is one of the most important non-sequence regulatory factors of many chromatin-based processes and has also been known to affect nucleosome stability. To identify histone modifications potentially significant to nucleosome stability, we applied feature selection procedure, the same as what was done with DNA sequences, on the data of 12 different histone modifications corresponding to chromosome III and promoter regions (Section *Data preparation*). The result was ranked by F-score and given in Table 4.4. This result shows that, the first 9 modifications of chromosome III, including H3K14Ac/H4K5Ac/H3K4Me3/H4K12Ac/H3K4Me1/H3K9Ac/H2AK7Ac/H4K16Ac/ H2BK16Ac, and the first 6 ones of promoter regions, including H3K4Me3/H3K9Ac/H3K18Ac/ H4K16Ac/H4K12Ac/ H4K8Ac, seem to be more important to nucleosome stability. No-

Table 4.1: Significant DNA motifs on chromosome III given by WordSpy

	Chromosome III		*Promoter Regions*	
Order	Motifs	F-score	Motifs	F-score
1	AT	1.37683	AG	0.69706
2	CA	1.12833	CT	0.623328
3	GA	0.913882	TG	0.577693
4	TG	0.894409	GA	0.575111
5	AA	0.882082	AT	0.572648
6	TA	0.813029	GC	0.537435
7	AG	0.811749	TC	0.517756
8	AC	0.803107	CA	0.507869
9	AAT	0.741735	GT	0.483424
10	TT	0.736747	TT	0.455674
11	CT	0.68323	CTT	0.452965
12	TC	0.64163	TA	0.446487
13	GT	0.615279	AA	0.41366
14	CAA	0.574223	AC	0.381596
15	GAA	0.523384	GAG	0.367994
16	GC	0.501134	GG	0.363897
17	ATT	0.499311	CC	0.362195
18	TAA	0.477322	TTC	0.330391
19	CC	0.455241	TAG	0.329403
20	TGA	0.453114	ATT	0.32476

Table 4.2: Significant DNA motifs on promoter regions given by WordSpy

	Well-positioned				*Delocalized*			
Order	Motifs	ZScore	Occur#	Seq#	Motifs	ZScore	Occur#	Seq#
1	TG	11.4	10865	995	TG	3.7	1164	69
2	CA	10.4	10913	995	TTG	5.7	406	66
3	GC	4.7	7254	992	TTC	5.3	400	67
4	GA	4.6	10360	993	TGG	4.7	286	61
5	CAA	14.9	3707	949	AGA	4.6	377	67
6	GAA	14.8	3696	948	CAA	4.5	371	69
7	TTC	13.6	3576	954	TTTC	5.3	141	52
8	TGG	12.6	2552	897	GGAA	5.1	101	48
9	CCA	10.5	2493	909	TTCTT	9.9	79	38
10	CTG	8.6	2384	897	TCTTC	7.5	52	34
11	TCT	8.2	3323	926	TTTCT	7.4	65	36
12	TTTG	14.1	1239	720	CTTCT	7.1	50	35
13	TTTC	14	1237	692	TCTTT	6.1	58	35
14	CTTC	13.2	910	553	AGGAA	5.8	42	31
15	CTTT	13.2	1216	668	AAGAA	5.6	53	39

tably, all significant modifications in promoter regions are strongly related to transcriptional activation (e.g. H3K4Me3/H3K9Ac/H3K18Ac) and repression (e.g. H4K16Ac/H4K12Ac/H4K8Ac) [30, 16, 17]. That is also true with some significant modifications in chromosome III, where H3K4Me3/H3K9Ac and H4K12Ac/H4K16Ac/H2BK16Ac are known to have strong relation with transcriptional activation and repression, correspondingly.

4.2.3 Effects of DNA sequences and histone modifications on nucleosome dynamics

In order to see how DNA sequences and histone modifications affect nucleosome stability, we applied our method to the data containing significant DNA motifs and histone modifications identified above (Section *Method overview*). After filtering out uninteresting rules (Section *Rule filtering*), we received two sets of 60 and 38 rules characterizing nucleosome dynamics on chromosome III and promoter regions, correspondingly. Table 4.5 shows some selected rules from these rule sets. Analyzing these rules, we discovered that the enrichness of some specific DNA motifs has special impact on nucleosome

Table 4.3: Discriminative motifs ranked by F-scores

	Chromosome III		*Promoter Regions*	
Order	Motifs	F-score	Motifs	F-score
1	AT	1.37683	AG	0.69706
2	CA	1.12833	CT	0.623328
3	GA	0.913882	TG	0.577693
4	TG	0.894409	GA	0.575111
5	AA	0.882082	AT	0.572648
6	TA	0.813029	GC	0.537435
7	AG	0.811749	TC	0.517756
8	AC	0.803107	CA	0.507869
9	AAT	0.741735	GT	0.483424
10	TT	0.736747	TT	0.455674
11	CT	0.68323	CTT	0.452965
12	TC	0.64163	TA	0.446487
13	GT	0.615279	AA	0.41366
14	CAA	0.574223	AC	0.381596
15	GAA	0.523384	GAG	0.367994
16	GC	0.501134	GG	0.363897
17	ATT	0.499311	CC	0.362195
18	TAA	0.477322	TTC	0.330391
19	CC	0.455241	TAG	0.329403
20	TGA	0.453114	ATT	0.32476

Table 4.4: Histone modifications ranked by F-scores

	Chromosome III		*Promoter Regions*	
Order	Modifications	F-score	Modifications	F-score
1	H3K14Ac	0.102054	H3K4Me3	0.0328115
2	H4K5Ac	0.0863558	H3K9Ac	0.0322587
3	H3K4Me3	0.0754543	H3K18Ac	0.0315715
4	H4K12Ac	0.0660357	H4K16Ac	0.0253305
5	H3K4Me1	0.0586061	H4K12Ac	0.0230635
6	H3K9Ac	0.0398707	H4K8Ac	0.0229266
7	H2AK7Ac	0.0309521	H3K4Me1	0.00913233
8	H4K16Ac	0.0219245	H2AK7Ac	0.00767291
9	H2BK16Ac	0.019511	H4K5Ac	0.00318472
10	H3K18Ac	0.00603551	H3K4Me2	0.00283706
11	H3K4Me2	0.004844	H2BK16Ac	0.00022866
12	H4K8Ac	9.68E-06	H3K14Ac	9.89E-06

stability. For example, nucleosomes bound by sequences enriched with AT/ATT/CTT are more stable (rules 1, 2, 6, 9, 10). This agrees with the result from [18], which said that sequences enriched with dinucleotides AT/TT have potential to inhibit nucleosome forming and deforming them on nucleosomes is more costly, so nucleosomes bound by these sequences may be more stable. Also, H3K9Ac/H3K18Ac/ H3K4Me3 are known to have positive relation with transcriptional activation [30, 16, 17], so nucleosomes which are hyper-acetylated at H3K9/H3K18 and hyper-trimethylated at H3K4 seem to be more dynamic (rules 7, 8). In contrast, H4K12Ac is known to have positive relation with transcriptional repression [30], so H4K12 hyper-acetylated nucleosomes are more stable (rule 5) while H4K12 hypo-acetylated nucleosomes are more dynamic (rules 11, 12). However, there is no DNA pattern or post-translational modification showing dominant effect on nucleosome stability. Instead, there exist combinatorial effects, by DNA motifs themselves (rules 3, 4, 9) or by both DNA motifs and histone modifications (rules 2, 5, 7, 8, 10, 11, 12), on nucleosome stability. For example, if H3K4Me3 or H3K9Ac nucleosomes are located in regions enriched with ATT tri-nucleotide, they will become more stable (rules 2, 10); and even being located in regions enriched with AT dinucleotide, H4K12 hypo-acetylated nucleosomes still have potential of becoming unstable (rule 12). This agrees with the results from previous and recent works showing that the effects of histone acetylations depend on which lysines are acetylated and the locations of modified nucleosomes [31, 32, 33]; and nucleosome positioning effect of DNA sequences is decided by the combination of nucleosome favouring and disfavouring motifs [18, 34].

Table 4.5: Selected rules characterizing nucleosome dynamics

No.	Rules	Class dist.
1	$AA, ATT = enr \wedge H3K9Ac = neutral \rightarrow State = Well$	[300 0]
2	$ATT = enr \wedge H3K4Me3 = hyper \rightarrow State = Well$	[156 0]
3	$AT, GC = enr \wedge CC = low \rightarrow State = Well$	[159 0]
4	$AT, CC = enr \wedge GC = low \rightarrow State = Well$	[56 0]
5	$AT = low \wedge H3K9Ac = neutral \wedge H4K12Ac = hyper \rightarrow State = Well$	[10 0]
6	$AT, TC = low \wedge ATT = enr \rightarrow State = Well$	[13 0]
7	$CT, TG, GA, AT, CTT, GAG, ATT = low \wedge H3K18Ac, H3K4Me3 = hyper \rightarrow State = Del$	[0 6]
8	$GA, TT, GG = low \wedge H3K9Ac = hyper \wedge H3K4Me3 = hypo \rightarrow State = Del$	[0 3]
9	$AA = low \wedge GT, ATT = enr \rightarrow State = Well$	[77 0]
10	$ATT = enr \wedge H3K9Ac = hyper \rightarrow State = Well$	[66 0]
11	$GA, AG, ATT = low \wedge H2BK16Ac = neutral \wedge H4K12Ac = hypo \rightarrow State = Del$	[0 15]
12	$AT = enr \wedge TA, TAA = low \wedge H3K9Ac = neutral \wedge H4K12Ac = hypo \rightarrow State = Del$	[0 4]

Chapter 5

Conclusion and future directions

Nucleosome dynamics plays important roles in many DNA-based processes and is regulated by many factors, such as DNA sequences, post-translational modifications of histone proteins, and chromatin remodelling complexes. However, most of the previous works only investigated the effect of individual factor while bypassing their combinatorial effects on the distribution of stable nucleosomes. In this paper, we proposed a novel method based on induction rule learning to computationally characterize nucleosome dynamics from both genomic and histone modification information. Our method is shown to be suitable for characterizing inhomogeneous distributions like that of destabilized nucleosomes; and by combining both genomic and histone modification information, it can discover potential co-effects of these two factors on nucleosome dynamics.

Our results on *S.cerevisiae* show that, some DNA motifs and histone modifications are more important in stabilizing and destabilizing nucleosomes. These DNA motifs and histone modifications are known to have strong relations with nucleosome forming/inhibiting potential and transcriptional activities, correspondingly. They not only act individually but also cooperate with each other by some specific patterns to combinatorially affect nucleosome stability.

Although our method is efficient in characterizing nucleosome dynamics, it produces a larger number of rules, of which many may be irrelevant. One way for achieving better results is to improve rule filtering procedure. Another issue is that, the method proposed

here hasn't considered the effects of other factors such as chromatin remodeling complexes as well as histone variants on nucleosome dynamics. To capture the effects of all these factors we need a more powerful computational model. These issues are now under consideration.

Publications

[1] Le NT, Ho TB, Tran DH: "Characterizing nucleosome dynamics from genomic and epigenetic information using rule induction learning", BMC Genomics 2009, 10(Suppl.3): S27

[2] Le NT, Tran DH, Ho TB: "Combinatorial effects of histone modifications and DNA sequences on nucleosome dynamics", Workshop on Knowledge, Language, and Learning in Bioinformatics (KLLBI 2008), in conjunction with Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI-08), December 16-19, 2008

References

[1] Luger K, Mader AW, Richmond AK, Sargent DF, Richmond TJ: "Crystal structure of the nucleosome core particle at 2.8 A resolution", Nature 1997, 389:251-260

[2] Kornberg RD, Thomas JO: "Chromatin structure; oligomers of the histones", Science 1974, 184(139):865-868

[3] Li B, Carey M, Workman JL: "The Role of Chromatin during Transcription", Cell 2007, 128(4):707-719

[4] Groth A, Rocha W, Verreault A, Almouzni G: "Challenges during DNA Replication and Repair", Cell 2007, 128(4):721-733

[5] Corpet A, Almouzni G: "Making copies of chromatin: the challenge of nucleosomal organization and epigenetic information", llenge of nucleosomal organization and epigenetic information. Trends in Cell Biology 2008, 19:29-41

[6] Probst AV, Dunleavy E, Almouzni G: "Epigenetic inheritance during the cell cycle", Nature Reviews Molecular Cell Biology 2009, 10:192-206

[7] Lee CK, Shibata Y, Rao B, Strah BD, Lieb JD: "Evidence for nucleosome depletion at active regulatory regions genome-wide", Nature Genetics 2004, 36:900-905

[8] Steven Henikoff: "Nucleosomes at active promoters: unforgettable loss", Cancer cell 2007, 12(5):407-409

[9] Steven Henikoff: "Nucleosome destabilization in the epigenetic regulation of gene expression", Nature Reviews Genetics 2008, 9(1):15-26

[10] Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JPZ, Widom J: "A genomic code for nucleosome positioning", Nature 2006, 442(7104):772-778

[11] Reinke H, Horz W: "Histones Are First Hyperacetylated and Then Lose Contact with the Activated PHO5 Promoter", Molecular Cell 2003, 11(6):1599-1607

[12] Zhao J, Diaz JH, Gross DS: "Domain-Wide Displacement of Histones by Activated Heat Shock Factor Occurs Independently of Swi/Snf and Is Not Correlated with RNA Polymerase II Density", Molecular and Cellular Biology 2005, 25(20):8985-8999

[13] Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: "Genome-Scale Identification of Nucleosome Positions in S. cerevisiae", Science 2005, 309(5734):626-630

[14] Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C: "A high-resolution atlas of nucleosome occupancy in yeast", Nature Genetics 2007, 39(10):1235-1244

[15] Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, Tomsho LP, Qi J, Glaser RL, Schuster SC, Gilmour DS, IstvanAlbert , Pugh BF: " Nucleosome organization in the Drosophila genome", Nature 2008, 453:358-362

[16] Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ: "Single-nucleosome mapping of histone modifications in S. cerevisiae", PLoS Biology 2005., 3(10)

[17] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: "High-Resolution Profiling of Histone Methylations in the Human Genome", Cell 2007, 129(4):823-837

[18] Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, Struhl K, Weng Z: "Nucleosome positioning signals in genomic DNA", Genome Research 2007, 17(8):1170-1177

[19] Zhang Y, Shin H, Song JS, Lei Y, Liu XS: "Identifying Positioned Nucleosomes with Epigenetic Marks in Human from ChIP-Seq", BMC Genomics 2008, 9:537

[20] Rippe K, Schrader A, Riede P, Strohner R, Lehmann E, Langst G: "DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes", Proceedings of the National Academy of Sciences of the United States of America 2007, 104(40):15635-15640

[21] Schnitzler GR: "Control of Nucleosome Positions by DNA Sequence and Remodeling Machines", Cell Biochemistry and Biophysics 2008, 51(2-3):67-80

[22] Widlund HR, Vitolo M, Thiriet C, Hayes JJ: "DNA sequence-dependent contributions of core histone tails to nucleosome stability: differential effects of acetylation and proteolytic tail removal", Biochemistry 2000, 39(13):3835-3841

[23] Yang Z, Zheng C, Hayes JJ: "The core histone tail domains contribute to sequence-dependent nucleosome positioning", Journal of Biological Chemistry 2007, 282(11):7930-7938

[24] Dai Z, Dai X, Xiang Q, Feng J, Deng Y, Wang J, He C: "Transcriptional interaction-assisted identification of dynamic nucleosome positioning", BMC Bioinformatics 2009, 10(Suppl l):S31

[25] Lavrac N, Kavsek B, Flach P, Todorovski L: " Subgroup discovery with CN2-SD", Journal of Machine Learning Research 2004, 5:153-188

[26] Das MK, Dai HK: " A survey of DNA motif finding algorithms", BMC Bioinformatics 2007, 8(Suppl 7):S21

[27] Clark P, Nibblet T: " The CN2 induction algorithm", Machine Learning 1989, 3:261-283

[28] Pavlidis P, Wapinski I, Noble WS: " Support vector machine classification on the web", Bioinformatics 2004, 20:586-587

[29] Wang G, Zhang W: " A steganalysis-based approach to comprehensive identification and characterization of functional regulatory elements", Genome Biology 2006., 7(6)

[30] Kurdistani SK, Tavazoie S, Grunstein M: " Mapping global histone acetylation patterns to gene expression", Cell 2004, 117(6):721-733

[31] Hebbes TR, Thorne AW, Crane-Robinson C: " A direct link between core histone acetylation and transcriptionally active chromatin", The EMBO Journal 1988, 7(5):1395-1402

[32] Wang A, Kurdistani SK, Grunstein M: " Requirement of Hos2 histone deacetylase for gene activity in yeast", Science 2002, 298(5597):1412-1414

[33] de Nadal E, Zapater M, Alepuz PM, Sumoy L, Mas G, Posas F: " The MAPK Hog1 recruits Rpd3 histone deacetylase to activate osmoresponsive genes", Nature 2004, 427(6972):370-374

[34] Jiang C, Pugh BF: "Nucleosome positioning and gene regulation: advances through genomics", Nature Reviews Genetics 2009, 10:161-172

[35] Groth A, Rocha W, Verreault A, Almouzni G: "Chromatin challenges during DNA replication and repair", Cell 2007, 128(4):721-33

[36] Turner BM: "Chromatin and Gene Regulation: Molecular Mechanisms in Epigenetics", Blackwell Science Ltd. 2001

[37] Scholkopf B, Tsuda K, Vert JP: "Kernel Methods in Computational Biology", MIT Press 2004