1

Modelling DNA Methylation Dynamics

Karthika Raghavan and Heather J. Ruskin
Centre for Scientific Computing and Complex Systems Modeling (SCI SYM), School of Computing, Dublin City University Ireland

1. Introduction

"Epigenetics" as introduced by Conrad Waddington in 1946, is defined as a set of interactions between genes and the surrounding environment, which determines the phenotype or physical traits in an organism, (Murrell et al., 2005; Waddington, 1942). Initial research focused on genomic regions such as *heterochromatin* and *euchromatin* based on dense and relatively loose DNA packing, since these were known to contain inactive and active genes respectively, (Yasuhara et al., 2005). Subsequently, key roles of DNA methylation, Histone Modifications and other assistive proteins such as Methyl Binding Proteins (MBP) during gene expression and suppression were identified, (Baylin & Ohm, 2006; Jenuwein & Allis, 2001). An emergent and persistent view that every epigenetic event affects another, to strengthen or suppress gene expression has made this an active field of research. DNA methylation refers to the modification of DNA by addition of a methyl group to the cytosine base, and is the most stable, heritable and well conserved epigenetic change. It is introduced and maintained, (Riggs & Xiong, 2004; Ushijima et al., 2003) by an enzyme family called DNA Methyl Transferases (DNMT), (Doerfler et al., 1990). Methyl-Cytosine or "mC", often referred to as the fifth type of nucleotide plays an extremely important role in gene expression and other cellular activities. Although DM is defined a simple molecular modification, its effect, can range from altering the state of a single gene to controlling a whole section of chromosome in the human genome.

The human genome is largely made of complex sequences evolved over time due to replication, mutations and insertion of foreign DNA. Based on the nucleotide distribution and functional significance, the genome has been categorized into different block of sequences, namely genes or coding and non-coding regions. A special type of sequence located near genes, in relation to spread of DNA methylation and dinucleotide frequencies are the CpG islands[1]. These islands are mostly found near the promoters, (5'end), of genes and their methylation levels are closely monitored to investigate the spread of Cancer. Useful insight on epigenetic mechanisms may be found from analysing the DNA sequence patterns or the genotype of the organism, (Gertz et al., 2011; Glass et al., 2004; Segal & Widom, 2009). Since more than 90% of DM occurs in CG dinucleotides, (Raghavan et al., 2011), knowledge of the distribution and location of CG can be utilized to understand the biological

[1] DNA sequences are defined and classified as CpG islands if , (a) length of that DNA sequence >200 bp, (b) Total amount of Guanine and Cytosine nucleotides >50%, and, (c) the observed/expected ratio of CG dinucleotides for that given length of sequence, >60%, (Takai & Jones, 2002)

significance associated with determining the level of DM. A general overview of pattern analysis techniques is given and application of time series analyses in understanding "CG" dinucleotide occurrences in specific human sequences are discussed in detail in the following sections.

Histones are proteins that protect DNA from restriction enzymes and also act as bolsters in chromosome condensation, (Ito, 2007). A "Histone Core", made of nine types of histone proteins, is attached to DNA molecules whose length varies from 146bp to 148bp. In the histone core, a combination of modifications, within specific amino acids in each histone subtype leads to gene expression or inactivation, (Kouzarides, 2007). These modification patterns, unlike stable DNA methylation, are dynamic and activation of one change leads to successive modifications of other amino acids during cellular events, (Allis et al., 2007; Jung & Kim, 2009). Even though new findings with regard to the impact of several modifications have been recently reported, information is inconsistent and less precise with regard to how a network of histone modifications communicates and is influenced by DM. Despite this insufficiency, the interactions between histones and DNA methylation are known to be disrupted at some stage, during the onset of cancer, (Esteller, 2007). Hence, a novel stochastic model, based on Markov Chain, Monte Carlo class of algorithms, (MCMC), was recently developed to mimic the epigenetic system and predict the effects of dynamic histone modifications over DNA methylation and gene expression levels, (Raghavan et al., 2010), (Details are discussed in Background section).

In this chapter, the focus on modelling the feedback dynamics of DNA methylation is dealt with in four parts, consisting of: (1) DNA Methylation mechanisms, controlling factors – DNA sequence pattern analyses and Histone modifications and their association with disease initiation, (2) A background on the recent data explosion, multiple methods and modelling approaches developed so far to investigate DM mechanisms and associated factors, (3a) Description of methods to investigate CG distribution in human DNA sequences – Results obtained and their association with DM spread, (3b) Developments on a novel micromodel framework, (based on MCMC) used to investigate Histone modifications for different DM levels and, (4) Results obtained for DM and HM feedback influence. Finally, conclusions and future directions for continuing investigation are considered.

2. Background

DNA Methylation was initially addressed as one of the most primitive mechanisms that organisms utilize to (a) protect genomic DNA and initiate the host resistance mechanism towards foreign DNA insertion and subsequently, (b) control gene expression, (Doerfler & Böhm, 2006). From an evolutionary point of view as well, the catalytic domain in the structure of the methylation enzymes across all organisms has been preserved to perform methyl group addition. A major change however, in the level and functional utility of DNA methylation was noted in higher organisms such as eukaryotes, when DM mechanism evolved from protecting the genomic contents to controlling their level of gene expression. In humans, there are two ways by which DNA Methylation is established – (a)*De novo* methylation that establishes new DM patterns, (b) Maintenance methylation responsible for inheriting existing DM patterns. Within the family of methylating enzymes (DNMT), two types namely DNMT3a/b/L and DNMT1 establish DM patterns in these two ways, (Doerfler

& Böhm, 2006). The *De novo* methylation process carried out by DNMT3a/b/L, is responsible for methylating embryonic cells which are totally erased of any previous DM patterns and methylated based on the DNA sequence contents. These mechanisms are also responsible for establishing parental imprinting and X-chromosome inactivation that is set permanently within the organism enabling it to exhibit unique phenotypes from birth. On the other hand, DNMT1 distribution is dynamic across a cell during its lifetime. This enzyme type is highly biased towards hemi-methylated[2] DNA sequences, making it responsible for propagating methylation patterns after each cell cycle. DNMT1 is also known to interact with histone deacetylases enzyme and some methyl adding proteins, (e.g. HP1), to remove acetyl and add methyl groups in histones, (Allis et al., 2007; Turner, 2001).

Associated aberrations in DNA methylation

As elaborately discussed by Chahwan et al, "the significant role played by DM in epigenetic regulation is quite apparent when the cell is affected due to impaired methylation marks during establishment, maintenance or recognition". Such changes in the "methylation marks" are mainly attributed to the abnormal function of DNMT enzyme complex which leads to failure of DM mechanisms. This abnormality results in gene imprinting disorders and malignancy formation due to *hyper/hypo methylation* of specific sections in the chromosomes, (Chahwan et al., 2011). Among the most studied abnormalities recorded in connection to failure of DNMT enzyme complex, is Immunodeficiency–Centromere instability–Facial anomalies (ICF) syndrome. This is caused due to mutations associated with coding for DNMT3B enzymes leading to global hypomethylation of repeat regions located in the pericentromere of human chromosomes, (Ehrlich. et al., 2008). Prader-Willi syndrome, Angelman syndromes and specific type of cancers such as Wilm's tumour have also been associated with imprinting disorders characterized by growth abnormalities, (Chahwan et al., 2011). In these diseases, genetic mutations or altered DNA methylation cause improper imprinting patterns and lead to aberrant expression of the normally suppressed genes, (Chamberlain & Lalandea, 2010). Based on accumulative information in literature, (Chahwan et al., 2011), Cancer initiation is mainly attributed to the imbalanced connectivity between oncogenes and tumor suppressor genes. Hence a combination of genetic abnormalities such as mutations and aberrant DM spread trigger cancerous conditions leading to malignancies that spread across different systems in the human body, (Allis et al., 2007). For example, in Wilm's tumour, the loss of imprinting of *IGF2* gene is associated with spread cancer to lung, ovaries and colon area. In general the DNA methylation pattern when disrupted can lead to, (i) gene activation, promoting the over-expression of oncogenes, (b) chromosomal instability, due to demethylation and movement of retrotransposons and consequently acquire resistance to drugs, toxins or virus, (Chahwan et al., 2011). Apart from failure in the control exercised by DM, there are certain protein "Onco-modifications" recently categorized as definitive signatures during occurrence of malignancies. Some of the most frequently studied histone modifications, associated with DNA methylation and tumor progress are – acetylation of H3K18, H4K16 and H4K12, trimethylation of H3K4 and H4K20, acetylation/trimethylation of H3K9, trimethylation of H3K27, occurrence of histone variants and also other external proteins such as MBP, HP1 and Polycomb that play role in chromosome rearrangement, (Chi et al., 2010; Fullgrabe et al., 2011).

[2] DNA sequences which have one of its double strands methylated

The above considerations make a compelling case to model and understand the DNA methylation mechanisms. In the following subsections, analyses of DNA methylation frequency and influence of genotype or DNA sequence patterns in humans are discussed, followed by elaborations on the control by DNA methylation mechanisms over Histone modifications.

2.1 DNA sequences and patterns analysis – Dimension 1

The human genome, consisting of more than three billion base pairs, is very complex and efforts to comprehend its organization and contents are still ongoing, (Collins et al., 1998; Strachan & Read, 1999). The spread of DNA methylation in the genome is not randomly determined. Emerging evidence indicates that, although chromatin modeling factors, iRNA, histone modifications and even parental imprinting memory can influence methylation, the underlying genotype or DNA sequence has a stronger key role in enabling and propagating a spectrum of methylation patterns, (Doerfler & Böhm, 2006; Gertz et al., 2011). The nature of every biological cell is characterized by its preservation of the genetic and epigenetic contents also known as "dual inheritance" and in consequence it is of utmost importance to look at the underlying genetic pattern maps for further comprehension of the epigenetic phenomenon.

When it comes to studying the epigenome or methylation landscape in connection to the initiation of Cancer, the focus is on genes and their alleles, non coding regions, and also *CpG Islands*, (Takai & Jones, 2002). The islands are one of the main locations for studying DM patterns in association with cell adaptability to environmental stress, epigenetic control and disease onset, (Allis et al., 2007). Furthermore, repetitive sequences or "Retrotransposon" which mostly belong to the non-coding regions, contain highly methylated CG dinucleotides in the human genome. These regions are silenced and kept under control due to the fact that they can replicate quickly and place themselves in different locations within the genome. They are also the favoured loci of "foreign" DNA insertions, which tend to disturb the existing DNA methylation patterns, (Collins et al., 1998).

Information from literature indicates that a majority of DNA methylation occurs in nucleotides, specifically located in these repeat regions (non coding) and in CpG islands, (Raghavan et al., 2011). The CG dinucleotides are usually under-represented across the human genome as a whole but are densely located in certain repeat regions and islands which may be differentially methylated during cancer initiation, (Esteller, 2007). CG dinucleotides in these regions follow a specific pattern and thus are easy targets for enzyme recognition and consequently, for methylation. The indications are also that certain patterns of CG base pairs, that are accessible by the DNMTs enzyme complexes, appear near promoters and islands of non-expressed genes in the human genome. Emerging evidence from genome analyses for example, reveals that the *De novo* methylating enzymes such as DNMT3a/L, are biased toward CG dinucleotides, appearing after every 8-10bp near promoters of methylated genes, (Glass et al., 2004). Hence it is vital to perform a complete distribution or pattern analysis of nucleotides in human sequences, in particular of CG to understand how methylation is established and maintained based on the sequence patterns within the genome. Although there is no complete evidence about the nature of DNMT mechanisms in setting new methylation patterns, analysing the global periodicities or distributions of CG dinucleotides will help to reveal a part of the hidden picture.

2.1.1 Methods to analyse DNA patterns

Since the advent of DNA sequencing technologies, (França et al., 2002), deciphering the significance of sequence blocks has been an important focus for geneticists. Apart from encoding for proteins, the human genome is a reservoir of information that has inherent patterns, corresponding to chromosomal condensation and evidence of evolution through common patterns among organisms. Several pattern recognition/analysis techniques or time series analysis methods[3] have been explored starting from simple statistical measures to complicated transformation and decomposition methods such as the Discrete Wavelet Transformation(DWT). A well-known approach in sequence analysis is to calculate "Expected Frequency" based on the empirical probabilities of the occurrence of nucleotides. This method was proposed by Whittle, and further developed to apply on DNA sequences by Cowan, (Cowan, 1991; Whittle, 1955). In the latter, transition probabilities (for all 16 types of dinucleotides) in the form of a matrix were constructed from known DNA sequences, to predict patterns along a new sequence. This particular analysis was performed on specific sequences containing the same starting and ending nucleotides. Another tool developed to visualize sequences, was "GC-Profile" which was based on, calculating nucleotide frequencies from the total amount of G and C nucleotides, and use of quadratic equations to check for purine levels in small genomes, (Gao & Zhang, 2006).

A standard pattern analysis can be conducted using the Fourier Transformation (FT), which allows decomposition of the time/spatial components in the data and construction of a frequency map, (Morrison, 1994). Fields of application are wide in range with examples from – Physics (optics, acoustics and diffraction), Signal Processing and Communication Systems, Image Processing, Astronomy, and DNA sequence analysis, amongst others, (A'Hearn et al., 1974; Goodman, 2005; Salz & Weinstein, 1969). Early work using Fourier technique in DNA pattern recognition was carried out by Tiwari et al. In this method, small sequences from bacteria were first converted into four distinct sets of binary sequences, (each corresponding to location of a nucleotide), then analysed by applying Fourier. This was followed by a comparison between genes and non-coding, and identification of characteristic features/patterns such as 3bp periodicity in genes. This type of application gave rise to the phrase "Periodicity" of nucleotides i.e. count of appearance of specific patterns that appear in sequences. Subsequent research focused on these periodicities of small patterns (length upto 10 bp) in blocks of sequences. Thus the Fourier transformation was used to study frequency components of the sequences along a spatial axis where each nucleotide was represented by a directional vector. Periodicities in virus strains (SV40) were also studied to check for patterns of dinucleotides and their corresponding role in genome condensation, (Silverman & Linskera, 1986). The most prominent periodical pattern of 10-11bp, portrayed by pyridines (AA/TT/AT), which are involved in long range interactions of upto 147 bp and aid in nucleosome alignment, was confirmed through these attempts. Refinement of this method through introduction of new parameters included calculation of *autocorrelation*[4] for specific patterns from DNA sequences. More recently, further improvements have been employed and tested on example sequences, (Epps, 2009). Complete and significant analyses of patterns or

[3] Applied to study patterns along the spatial-varying data in DNA sequences.

[4] Autocorrelation of patterns is an extension for periodicity, i.e. appearance of a pattern after a lag or distance of "k" base pairs.

biological markers on sequences were identified by, (Herzel et al., 1999) and (Hosid et al., 2004) from *E.coli* genome. In the latter paper, authors discuss landmark periodicities in detail, along with supportive evidence of their biological significance inside the genome. This includes – 3bp spacing followed by all 16 dinucleotides in genes, 10-11bp spacing by pyridines, and some organism specific distributions. The corresponding power spectrum, that provide information on global periodicities, was calculated, (Hosid et al., 2004) using:

$$\mathrm{f}_p = \frac{\sqrt{\sum_{i=1}^{m} sin\left((2\pi * \frac{i}{p}) * (X - X')\right)^2 + \sum_{i=1}^{m} cos\left((2\pi * \frac{i}{p}) * (X - X')\right)^2}}{\left(2\pi * \sum_{i=1}^{m}(X - X')^2\right)} \qquad (-1)$$

f_p = Normalized wave function amplitude at period - p
X = Auto correlation profile of the dinucleotide
X′ = Mean Auto Correlation
m = Maximum autocorrelation distance
p = Periodicity or in this case distance between identical patterns or nucleotides.

A Fourier analysis in our case involves calculating the auto correlation profile for desired dinucleotide/ nucleotide followed applying the formula shown above. More details on this approach and its application to study nucleotide distribution in genes, non-coding regions and CpG islands are discussed in the Methods section. The aim of this initiative was to understand the distribution of CG dinucleotides, similiar to the work of (Clay et al., 1995), and on different datasets containing genes, CpG islands and non-coding regions[5].

2.1.2 Note on Discrete Wavelet Transformation

An extension to the Fourier analysis, Discrete Wavelet Transformation, is the application of a set of orthonormal vectors in space to localize and study both frequency and time/spatial components for a given dataset, (Kaiser, 1994). The resulting coefficient matrix, a product of this family of vectors and input data helps to indicate regions of high and low frequencies along the spatial, (or sequential) axis based on an initial resolution factor, (e.g. Haar and Mortlet, (Kaiser, 1994)). Wavelets or specifically the method of DWT addressed here, have been quite extensively used to study financial markets, experimental data from Protein Mass Spectrometry and DNA sequence patterns amongst others, (Kwon et al., 2008). Although DWT is not quite often used as fourier, it has also been applied to visualise both frequency and location specific information of the DNA sequence patterns, (Tsonis et al., 1996; Zhao et al., 2001). Elaboration on this family of approaches, is not explicitly dealt in this chapter, hence more details on the method of Maximal Overlap Discrete Wavelet Transformation, (MODWT - extension to DWT), (Conlon et al., 2009), application to study patterns in DNA sequence and results thus obtained, are reported in (Raghavan et al., 2011).

So far we have discussed various methods and algorithms, used to detect nucleotide patterns in human DNA sequences and have considered in more detail the role of Fourier

[5] The non coding regions referred here in this analysis are the segments in-between exons/coding regions and are removed during translation or protein production phase

Transformation technique in investigating these patterns. In the next subsection, attempts to investigate the occurrence of histone modifications are reviewed. We describe ways to explore the relationship between these and DNA sequences. To test these approaches, we combine the results from Fourier analysis, or dinucleotide patterns with information on specific histone modification effects at fixed DNA methylation levels, using our recently developed, EpiGMP prediction tool.

2.2 Histone modifications – Dimension 2

Histones are closely linked to DNA molecules and play a vital part in encoding information from them. Over time, histone proteins have diversified from a few ancestors into five distinct types of subunits (2 copies of H2A, H2B, H3 and H4 each and a H1 subunit) in eukaryotes thus forming the octomeric structure of a nucleosome, (Allis et al., 2007). This nucleosome comprising of histone complex and 146 to148bp bp of DNA molecules on average, forms a “bead on string” structure. The histone octomer or core plays the most important role in condensing billions of DNA base pairs compactly within 23 pairs of chromosomes in the human genome. Covalent posttranslational histone modifications are mainly held responsible for chromatin architecture and propagation of many cellular events from simple gene expression to cell fate determination, differentiation, and, sometimes, disease onset. Thus, with more than one type of histone containing multiple types of modification (acetylation, methylation, phosphorylation, ubiquitination and sumoylation) in their tails present a potentially complex scenario, (Cedar & Bergman, 2009; Jenuwein & Allis, 2001; Kouzarides, 2007; Zheng & Hayes, 2003). DM and HM most often have a mutual feedback influence hence maintaining a strong dependency over one another. A very interesting fact about histone modifications is that though the exact mechanisms are unknown, they are memorized by the cells “post replication”, especially those that aid in gene expression, methylation maintenance and chromosome structure stability. Among all the histone modifications, methylation (mono/di/tri) and acetylation have been most studied in regard to their influence over gene expression. These modifications are quite often noted to compete for the same type of residues and are also known to recruit antagonistic regulatory complexes such as trithorax and polycomb proteins, (Allis et al., 2007). For example, histone methylation was found to be important for DNA methylation maintenance at imprinted loci, which could lead to disorders such as the Prader-Willi syndrome, (Chahwan et al., 2011). Such individual experiments have helped unravel the connection step by step between levels of DM and specific histone modifications including special histone variants, (Barber et al., 2004; Ito, 2007; Meng et al., 2009; Sun et al., 2007; Taplick, 1998; Wyrick & Parra, 2008). Hence a complete picture of the molecular communications that control the cellular events is lacking. Consequently, attempts have been made to accumulate the *cross-talk information* from laboratory experiments and decipher the modification patterns in the human genome during different cellular events, (Bock et al., 2007; Yu et al., 2008).

2.2.1 Modeling DNA methylation and histone modification interactions

Epigenetics, as a field, is relatively new and models to study the associated phenomena are limited to date. The advent of favourable experimental techniques such as Protein Mass

Spectroscopy, (Sundararajan et al., 2006), ChIP-Seq and ChIP-on-Chip[6], (Collas, 2010), have led to new data and confirmed facts with regard to DNA-protein interactions and their role in cancer onset. Such experiments usually generate a large amount of data including measures such as direct count of modification detected along the genome after specific intervals of DNA sequences, (standard intervals are 200 or 400 base pairs for histone modifications detection). As discussed in detail, by Bock et al, extracting comprehensible epigenetic information is a three-stage process. First, the biochemical interactions are stored as genetic information in DNA libraries, followed by applying DNA experimental protocols such as tiling microarray, (special type of microarray experiment) along with ChIP-on-ChIP, and lastly applying computational algorithms to infer error free epigenetic information from these experiments. These algorithms are mainly quantitative and help to establish a pipeline for prediction of probable epigenetic events. An initial coarse attempt to define the epigenetic, genetic and environmental interdependencies paved the way for an in depth study of the molecular factors that trigger these effects, (Cowley & Atchley, 1992).

Among the many computational attempts to model and analyse epigenetic mechanisms some have successively identified correlated histone signatures during gene expression using data from ChIP-on-ChIP experiments and microarray based gene expression measurements, (Karlić et al., 2010; Yu et al., 2008). A Bayesian network model was constructed using the high-resolution maps from laboratory experiments to establish casual and combinatorial relationships among histone modifications and gene expression, (Yu et al., 2008). Quantitative measure of other proteins such as Polycomb, CTCF (insulating proteins) and Transcription factors were also included to build these models. Based on Bayesian networks, conditional probabilities and joint probability distribution measures of datasets were calculated and a finely clustered molecular modification network was obtained.

Repeated bootstrapping or random sampling verified the robustness of this Bayesian Network. For initial analysis, datasets containing information from ChIP-on-ChIP experiments ((Cuddapah et al., 2009) and (Boyer et al., 2006)) for histone protein modifications in human CD4+ (immunity), cells and gene expression measurements from microarray experiments (obtained from (Su et al., 2004)), were extracted for clustering (using k-means), followed by construction of the bayesian network.

Another quantitative model based on the same type of information such as data from ChIP-on-ChIP experiments, obtained from literature, (Cuddapah et al., 2009), was developed using Linear Regression (Karlić et al., 2010). In this case, a regression expression was used to build the model: ($N_{i,j}'$=$N_{i,j}$+constant), where, $N_{i,j}$ = count of j_{th} modification in i_{th} gene in template samples. This equation was modified by inclusion of more variables, to study multiple histone modifications, thus giving rise to more than one model type. Secondary information was also extracted and included in the model, namely, microarray expression data from literature, (Schones et al., 2008) and promoter blocks information from Unigene databases, (`http://www.ncbi.nlm.nih.gov/unigene`). Here, loci of new sets of ChIP-on-ChIP experimental results for histone modifications, were mapped on human genome using annotation track information obtained from University of California Santa Cruz genome browser, (`http://genome.ucsc.edu`). These multivariable models were

[6] Experiments conducted to check for protein-DNA interactions combining chromatin immuno precipitation and massively parallel DNA sequencing techniques or microarray (chip) experiments

applied on different sequence datasets which were based on Low CG or High CG dinucleotide concentration. The whole dataset thus obtained was divided into training and test sets namely – D1 and D2, where Pearson correlation coefficient values were used to confirm the accuracy of prediction(D1) over the test set, (D2). This model was also extended over different cells, (with initial trials being conducted on CD4+ human cells), for nine histone modifications and for confirmation on CD36+ and CD133+ human immune cells respectively.

Other model types based on Bayesian networks, have focused on developing tools to study DNA methylation and protein modifications, (Bock et al., 2007; Das et al., 2006; Jung & Kim, 2009; Su et al., 2010). Among those, two models by Jianzhang et al and Bock et al have mainly focused on identifying the function of CpG islands using information on Histone Modifications. These type of "reverse" models explain the feedback connectivity between the two epigenetic events (HM and DM). Bock's model was an important initiative in computational epigenetics, since a clear pipeline for analysis of epigenetic data was proposed. The training model used several inputs from the experimental datasets to identify *bonafide* CpG islands. Inputs included – CpG islands that qualified based on criteria defined, (Takai & Jones, 2002) and epigenetic datasets from experiments (such as lysine modifications in histones, transcription binding factors, MBP, and SP1 proteins). This work consisted of three main steps, the first of which involved identification of predictive parameters from the datasets, followed by cross validation and training of data using a linear support vector machine, and lastly comparison of CpG islands previously identified in chromosome 21. These elaborate measures took into account the level of histone modifications affecting the methylation status hence emphasizing on the strong connectivity between methylation levels and their corresponding epigenetic states. Similar to the model described, (Yu et al., 2008), another complementary attempt was made to construct regulatory patterns that appear in histone during high DNA methylation. A Bayesian network once again was used to predict a list of methylation modifications that leveraged the occurrence of DNA methylation (using the same datasets obtained from CD+4 cells in humans), (Jung & Kim, 2009). These independent and repeated attempts, on accumulation, helped to identify and confirm a definitive pattern and characteristic modifications that exist in epigenetic events in the human cells: for example, more acetylation modification appear during gene expression and more methylation modifications are preferred during gene suppression.

A major disadvantage in the development of these quantitative models was the restriction of obtaining results from a single source or studies performed to investigate a single disease onset. Such a scenario cannot account for the epigenetic events for all conditions due to absence of a general model framework that could definitively link different epigenetic events. This has ultimately indicated a need to develop a general predictive model that can report modifications occurring in genes associated with any type of cell or cancer (provided there is evidence on the role of genes in diseases). As a consequence, we recently developed a theoretical model based on cumulative information of the nature of epigenetic events and tested it on synthetic data, (Raghavan et al., 2010). The novelty of this micromodel lies in accounting for the dynamics in the epigenetic mechanisms based on a stored library of possible histone modifications as well as DM associated patterns in the DNA sequences. The model, which is based on MCMC algorithm, allows sampling of possible solutions of histone modifications, using probabilities of transition. Based on the accumulative knowledge on the nature of modifications as mentioned above, probabilistic cost functions are used to set the interdependencies between variables (HM and DM based patterns) in this model. This dependency, influences the random sampling and calculates the final output or rate of transcription (T) using exponential equations ($T= e^{x} * e^{y} * k$, "x" and "y" being histone modifications and DNA methylation respectively and "k" a constant value of transition probability – Figure 4). As a part of the validation, the initial probabilities of transition set have been assigned random values so as to investigate results, (Monte Carlo or boot strapping). Ultimately, our micromodel, in a simple and consistent manner can predict or forecast a possible network of molecular events that occur during specific cellular events such as gene expression and suppression.

3. Methods and modelling approaches

In this section, we discuss the current approaches and algorithms that were applied to study each epigenetic component influencing DNA methylation mechanisms. The use of Fourier Transformation to detect patterns in specific genes extracted from human genome databases is elaborated. This is followed by a detailed explanation of a stochastic algorithm recently developed, and its application on the gene datasets, to predict histone modifications corresponding to changes in DNA methylation levels.

3.1 Application of fourier transformation

The main aim is to use collateral data (or meta data) based on information from literature, (Yu et al., 2008) to refine our understanding of the complex epigenetic system. The focus here is to investigate the human genome for multiple patterns of specific dinucleotides (AA, TT, AT) and (*CG* - discussed here), that play a major role in epigenetics. As stated before, recurrent evidence, (Glass et al., 2004) suggests that distribution of specific dinucleotides control events like DNA methylation and chromatin remodeling. The methylating enzymes (DNMT) help to monitor the location and level of DNA methylation, in all types of cells based on these distributions. Hence among the available methods in time-series analyses, Fourier Transformation was chosen to study the frequency domain of specific components in spatially (or sequentially), varying DNA sequences.

Input data or DNA sequences obtained using Map viewer, NCBI database (`www.ncbi.nlm.nih.gov`) and UCSC genome browser (`http://genome.ucsc.edu`) were classified and tabulated into three sets namely - (i)19 Genes, (ii) non-coding regions near the genes and, (iii) All CpG islands in chromosome 21, for Fourier analysis. Details of specific genes, chosen due to their association with disease conditions, are given in Table 1.

Figure 1 shows how the CG patterns are screened for auto correlation, (associated with epigenetic mechanisms). Following screening, the amplitude of Fourier Wave Function for contributing periodicities was derived for the 19 genes, corresponding non coding regions and all CpG islands present in chromosome 21, (using equation 1).

3.2 Results on fourier methods

Fourier analysis of dinucleotide patterns in human DNA sequences, seeks to determine significant DM levels associated with these features. In particular, CG patterns are of interest, as this dinucleotide is known to be involved in DNA methylation. Figure 2 represents average

S.No.	Genes	Diseases associated with Genes
1.	PRSS7	Enterokinase Deficiency
2.	IFNGR2	Arthritis Lupus Erythematosus
3.	KCNE1	Jervell and Lange–Nielsen syndrome type 2 (JLNS2)
4.	MRAP	Glucocorticoid Deficiency type 2 (GCCD2)
5.	IFNAR2	Myeloid Leukemia, Hepatocellular Carcinoma, Behcet Syndrome, lung and bladder cancer
6.	SOD1	Amyotrophic Lateral Sclerosis type 1 (ALS1)
7.	KCNE2	Atrial fibrillation familial type 4 (ATFB4)
8.	ITGB2	Leukocyte Adhesion deficiency type I (LAD1)
9.	CBS	Atherosclerosis, Atherosclerosis, Coronary, Breast cancer and cystathionine beta-synthase deficiency
10.	FTCD	Glutamate Formiminotransferase Deficiency (GLUFORDE)
11.	PFKL	Mediterranean Myoclonus
12.	RUNX1	Asthma, Myeloblastic Leukemias
13.	COL6A1	Bethlem myopathy (BM)
14.	COL6A2	Bethlem myopathy (BM), Ullrich Congenital Muscular Dystrophy (UCMD), Autosomal Recessive Myosclerosis
15.	PCNT2	Microcephalic Osteodysplastic Primordial Dwarfism type 2 (MOPD2)
16.	CSTB	Neurodegenerative Disorder
17.	LIPI	Dyslipidemia
18.	TMPRSS3	Deafness and Nonsyndromic
19.	APP	Alzheimer's Disease, Dementia, Attention Deficit and Oppositional Defiant disorder

These gene sequences were used in Fourier Analyses.

Table 1. Dataset containing Genes and Diseases associated with them

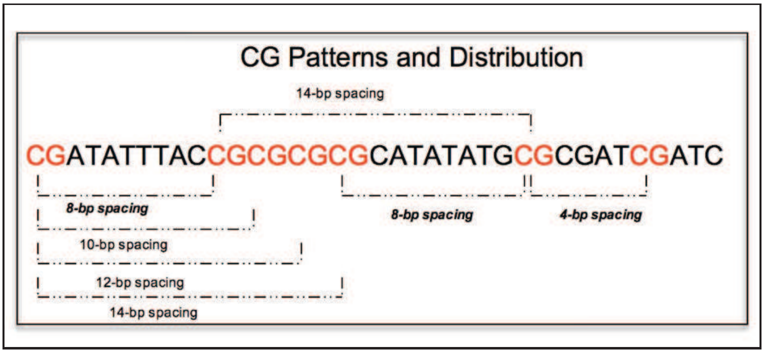


Fig. 1. Distribution of CG in Human DNA sequences.

amplitudes of the power spectrum for all values of CG periodicities possible. Genes/coding regions show an apparent peak at 3bp, which might be expected due to the codon bias in translating to amino acids, (Hosid et al., 2004). CpG islands, (throughout chromosome 21), also contribute to the peak at a periodicity of 3bp since these are present near the promoter

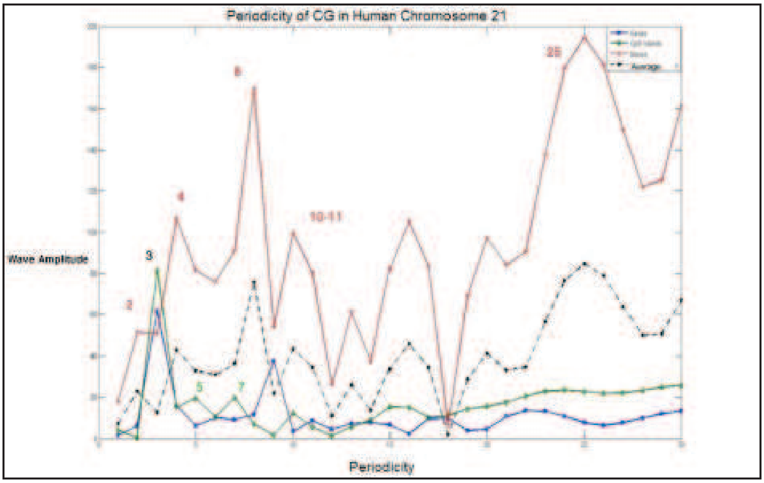


Fig. 2. Fourier analysis (Periodicity Vs Average Wave Amplitude) of global periodicities of CG dinucleotides in 19 Genes (blue line), non-coding near them (red line) and all CpG Islands (green line) in chromosome 21. The average of the 3 region levels is shown as a dotted line.

regions [7]. A 7bp spacing is also observed, probably due to repeats containing CG, in an island located near methylated regions, (Glass et al., 2007). The placement of CG after 3bp, in genes and even more densely clustered in CpG islands prevents the DNMT complex from naturally methylating those regions, (Glass et al., 2004). Hence spacing repeats of CG dinucleotides, can be used to confirm a CpG island, in addition to the dinucleotide based criteria in any input sequence, (Takai & Jones, 2002). One of the more prominent and interesting features can be noted in the non-coding regions, which display unexplored patterns (between 24 and 26bp). Research indicates that 8bp, and also 4bp intervals, (preferred by satellite/short repeats), (Glass et al., 2004), in this dinucleotide, attract DNA methylation complexes. In fact, genes that are silenced in germ cells by the *De novo* methylation mechanism, have these distributions near their promoters. Another peak, observed in Figure 2, between 10 to 11bp periodicity has been confirmed to support genomic structural condensation, (Glass et al., 2004). Other peaks, at periodicity of 15 and 20bp, are less persistent and are possibly due to noise in relation to dense repeat regions in chromosome 21.

The hitherto unexplored periodicity of an interval of length 24 to 26bp, in the non-coding region is less readily explained, but may be connected to DNA methylating mechanisms. A major clue, indicated in (Li. et al., 2010), is the appearance of several million repetitive 25-mers in the human genome. Although not uniform throughout the chromosome 21, this occurrence is known to be high, on average in the human genome. Furthermore in a recent paper, (Yin & Lin, 2007), the authors explain that piRNA or Piwi protein associated iRNA[8], which is significantly involved in cellular processes and propagation of *de novo* DNA methylation is usually of length 24 to 26 nucleotides, (Raghavan et al., 2011). This

[7] Promoters are blocks of DNA sequences that control expression for a set of Genes

[8] iRNA is an unusual type of single stranded RNA derived from DNA which help in blocking genomic information for protein production.

new evidence is only a part of the story of human DNA sequence analyses, especially with respect to differential gene expression, as controlled by epigenetics. The average plot as a test of confirmation, represented by dotted line in Figure 1, appears to retains the feature of major peaks at 8, 24, 25 and 26bp for all 22 chromosomes, which could be proposed as standard "marker patterns" of the human genome. Thus FT methods helped to identify possible CG distributions both previously reported and unexplored and to furnish supportive evidences on their corresponding biological significance. Following the initial data analysis, the sequences were investigated for possible histone modifications using our novel stochastic tool based on fixed initial DNA methylation levels.

3.3 Conceptualization of Epigenetic Micromodel – (EpiGMP)

The initial attempt to mimic the biological epigenetic structure is illustrated in reference, (Raghavan et al., 2010) which shows a simplified construction of our model. The status of epigenetic profile in the model is defined in terms of the corresponding DNA Methylation and associated Histone Modifications and model execution portrays the evolving interactions or interdependencies of the epigenetic elements. This section explains how histones were encoded and chosen for defined levels of DM. Information, (Kouzarides, 2007), on the number and type of amino acids for each histone type provides inputs to the model before the simulation. Table 2 gives the details of the number of amino acids, their positions, the

S.No.	H. Type	Amino acid No./String size	Amino Acid & Position	Modification	No. of States
1.	H1	zero	-	-	-
2.	H2A	Four	S1-R3-K5-K9	Ph-Met-Ace-Ace	16
3.	H2B	Ten	K5-S10-K11-K12 S14-K15-K16 K20-K23-K24	Ace/Met-Ph-Ace-Ace Ph-Ace-Ace Ace-Met-Ace	1536
4.	*H3	Six	R2-T3-K4 R8-K9-S10 T11-K14-R17 K18-T22-K23 R26-K27-S28 T32-K36-K37	Met-Ph-Met Met-Ace/Met-Ph Ph-Ace/Met-Met Ace/Met-Ph-Ace/Met Met-Ace/Met-Ph Ph-Ace/Met-Met	6300
5.	H4	Five	S1-R3-K5-K8-K12	Ph-Met-Ace-Ace-Ace/Met	48

Details of specific amino acids and their corresponding modifications in all histone types.
* - H3 has a special type of representation based on amino acid type and the corresponding modification. K - Lysine, S - Serine, T - Threonine, R - Arginine, Ace - Acetylation, Met - Methylation, Ph - Phosphorylation, citepThomas

Table 2. Amino Acid Positions and Modifications

corresponding modification types and the possible number of histone states generated, (Allis et al., 2007; Cedar & Bergman, 2009; Jenuwein & Allis, 2001; Kouzarides, 2007; Turner, 2001). These data are stored in the model as possible combinations of histone modifications that exist in the real epigenetic system. The modifications for each amino acid are assigned a value between 0 and 3 (acetyl -1, methyl -2 phosphate -3 and no modification - 0), which can generate libraries of strings with varying length based on histone type. These numerical strings represent histone modification state in a precise and encoded form. In the previous and current model versions, each string is considered as a node that can be visited during simulation based on a Markov chain - transition probability. A large number of strings exist for each histone type to be sampled due to the fact that each histone has many amino acid modifications, (Raghavan et al., 2010). For example, in case of H2A, a histone *state* or node whose string length is 4 here would be "3011". In this node, the Serine amino acid is phosphorylated and Lysine 5 and 9 are acetylated. A time-step or *Iteration* of the model

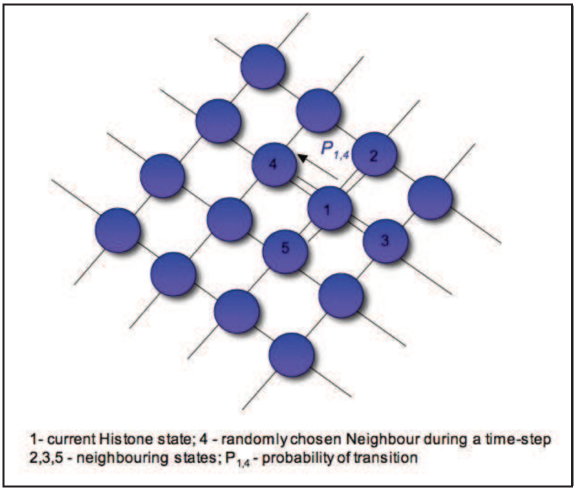


Fig. 3. The movement between active nodes or histone modifications in our model. Based on a random sampling, system shifts to node 4 from 1, based on an appropriate probability of transition. For example, if in case of H2A histone type, state 1 = "0000" and 4="3000", (Raghavan et al., 2010).

corresponds to moving between possible nodes, (i.e. if system chose to modify an amino acid) or remaining in the same node. Consequently, only one change or modification is made at each iteration when the model randomly samples between the possible histone states, based on probability of shift, (as shown in Figure 3). The potential shift to a "neighbouring state" from the current histone state is calculated during every iteration of the model. Computational graphs[9] or tables, of varying sizes based on the type of histone, are used in the system to store occurrence of dynamic modifications. These networks of graphs represent the level of modifications in all histone types and are used to calculate system outputs over several iterations. Our model can also handle multiple additions of the same modification in an amino acid (Mono/di/tri acetylation, methylation or phosphorylation, (Kouzarides, 2007)). Although this is invisible to the user, it is taken into account during calculation of global modification levels in each nucleosome. Hence for individual histone type, the modifications

[9] This is the application of graph theories which refers to use of appropriate data structures to store data whenever necessary.

are updated at each iteration, based on the influence of the DNA methylation values and output values of gene expression levels are calculated as depicted in Figure 4 and in reference, (Raghavan et al., 2010).

3.3.1 Epigenetic interdependency

A simple yet strong and well defined inter-dependency exists between histone evolution, transcription rate and level of DNA methylation inside each computational Block (or object, (Raghavan et al., 2010)). There are 3 main interactions in our model. The main dependency

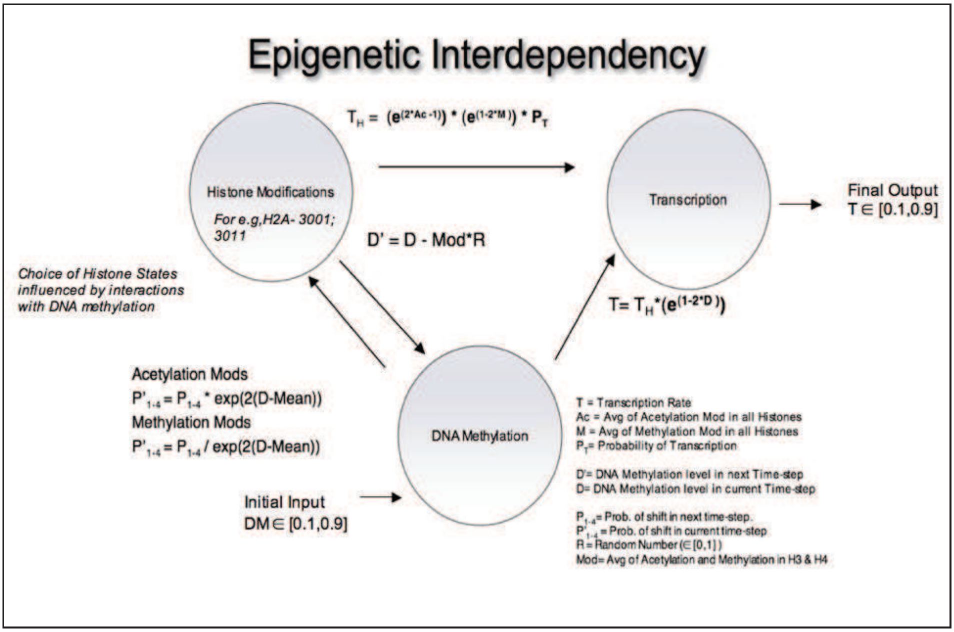


Fig. 4. Interactions between Epigenetic Elements in the Complex System. DM, associated with CG patterns in the DNA sequences and HM alter over each time step. *Transcription,* the output based on both parameters is calculated at regular intervals.

is mutually between Histone modifications and DNA methylation. Here the transition probability of histone states is altered by DNA methylation values, through use of exponential equations hence allowing the system to choose modifications preferentially. This crucial step is based on cumulative information extracted from laboratory experiments, which mention that specific patterns of modifications are explicitly preferred to other types during different levels of DNA methylation. Here, probabilities of shift, provide a window of control to introduce stress to the system so as to see how the output parameters fluctuate over several time-steps. The system is perturbed or subjected to stress through random initial probabilities for histone evolution, (or Monte Carlo based simulation) over different independent trials and subsequently system behaviour can be observed for changes in HM and DM based on their interactions.

Conversely, DM values are recalculated, *conditionally*, from average protein modification levels. This conditional step in DM calculation, has been implemented since literature states that DNA methylation levels are usually stable and less perturbed over several generations. The total output is expressed as "Transcription" which is calculated based on methylation levels in sequences and corresponding histone modifications. Details on the mathematical interdependency of the variables in the model are depicted clearly in Figure 4, (Raghavan et al., 2010). Results obtained from repeated simulation attempts are explained in the next section.

3.3.2 Simulation of combined model

The model consisting of DNA sequences and CG patterns together with histone states is executed to observe evolution of Histone modifications associated with DM in sequences similar to the real system. The steps given below explain the simulation process. The "Blocks" referred from here, are the computational representation of gene or island blocks of sequences within the EpiGMP model framework.

1. *Read and Store Inputs*
   (a) Histone Data -The possible combinations of Histone modifications as described in, (Raghavan et al., 2010) – states and transition probabilities.
   (b) DNA sequences with information on CG distribution throughout sequences are stored as well
   (c) User Selected Values are provided –
      i. Default Parameters: Maximum number of iterations(or time-steps), time-intervals and DNA methylation per a Block in a specific time-step.
      ii. Optional Parameters: preferred histone states in one or more blocks, set by the user (location during a time interval)
2. *Create Objects*
   (a) In one Block – Nucleosomes (number based on DNA sequence length) are created. Each nucleosome object, is assigned nine histone types (default) and 3 modification tables/graphs for each histone.
3. *Simulate*
   (a) Allow Markov Shifts among possible histone states for choice of solution.
   (b) For specific time-intervals, calculate DNA methylation if needed and output parameters: Transcription (based on interdependencies as in Figure 4).
   (c) Continue process till maximum number of iterations reached (for example 10,000 time steps).
4. *Store Outputs*
   (a) Results for the specified time interval, inside each Block –
      i. Transcription rate
      ii. DM value (assumed to be methylation of each CG dinucleotide)
      iii. Count of possible histone node visited per nucleosome

3.4 Model assumptions

As the major focus is on HM and DM progression, a few simplified assumptions were made to test the EpiGMP model reliability.

1. The model currently handles only three modifications i.e. Acetylation, Methylation and Phosphorylation as their biological role is known, (Kouzarides, 2007). More types of modifications can be included, given empirical or theoretical evidence on their significant contributions. (e.g. Role of Ubiquitination in H2B amino acids.)
2. One type of CG distribution, based on results from Fourier transformation method, i.e. CpG islands and gene blocks as shown in Table 1 are tested for prediction of possible histone modification under varying levels of DM.
3. H2A, H2B and H4 are encoded in a similiar fashion as explained above. However, H3 histone type has a large number of modifiable amino acids that can generate millions of possible histone states. Hence, to handle the large dataset, a special representation mode that could compress the possible histone states/nodes was developed. Methods to encode this histone type has been discussed in detail in, (Raghavan et al., 2010).
4. Independent simulation was carried out with three initial random transition probabilities. These values are generated by a system defined function (based on a pseudo random number generator - Mersenne Twister, which is robust, has a large range of period and a high order of dimensional equidistribution, (Matsumoto & Nishimura, 1998)). Hence the results obtained and discussed are the average of the three independent simulation trials.

This is a more advanced model in comparison to the one developed in (Raghavan et al., 2010), which considers both analysis of CG dinucleotide distributions and choice of histone modifications over the chosen sequences. The aim here was to observe histone evolution with DM associated sequence patterns in a manner similiar to real system and results thus obtained from this study are discussed in the next section.

4. Results and discussion

In order to investigate the system behaviour, 19 specific genes, and all CpG islands present in chromosome 21, were chosen. The datasets were preferred since they contain the maximum number of CG dinucleotides with 3bp intervals. These base pairs with specific distributions (usually associated with differentially expressed genes and promoters, (Allis et al., 2007)) were assigned DNA methylation values, based on equations shown in Figure 4. Outputs namely, Histone states, progress in transcription rate and DNA methylation, for the whole dataset were recorded every 1,000 time-steps (total number of time steps being 10,000). Although the system can trace and report evolution of all 4 types of histone, we discuss here only 2 types namely H4 and H2A. The following Figures 5 and 6 show the expected values of each histone node being chosen during several iterations over the 3 independent simulation trials.

The DNA methylation was set to a range of values, $\in[0.1, 1.0]$, for the 3 simulation runs (results not shown here). For initial values, (<0.2) of DM, the systems preferred least methylation modifications and inversely more acetylation changes. But for more sets of initial methylation values in the range [0.3, 0.6], and those (>0.75), methylation was apparently chosen repeatedly among other histone modifications. This was due to evolution of DM values to a closed range of [0.95, 1.0] over a time period of (> 10,000) iterations. Hence to observe histone evolution we discuss in detail two sets of results observed under (i) Low DM (<0.15 or 15%), and (ii) High DM (>0.85 or 85%). These simulations demonstrate effective emulation of the biological process of transcription of genes (e.g. Onco-genes expression) for low DNA methylation levels and reverse case of high DNA methylation and gene suppression (e.g. silencing of tumor suppressor/control genes). Figure 5 contrasts the different modifications observed

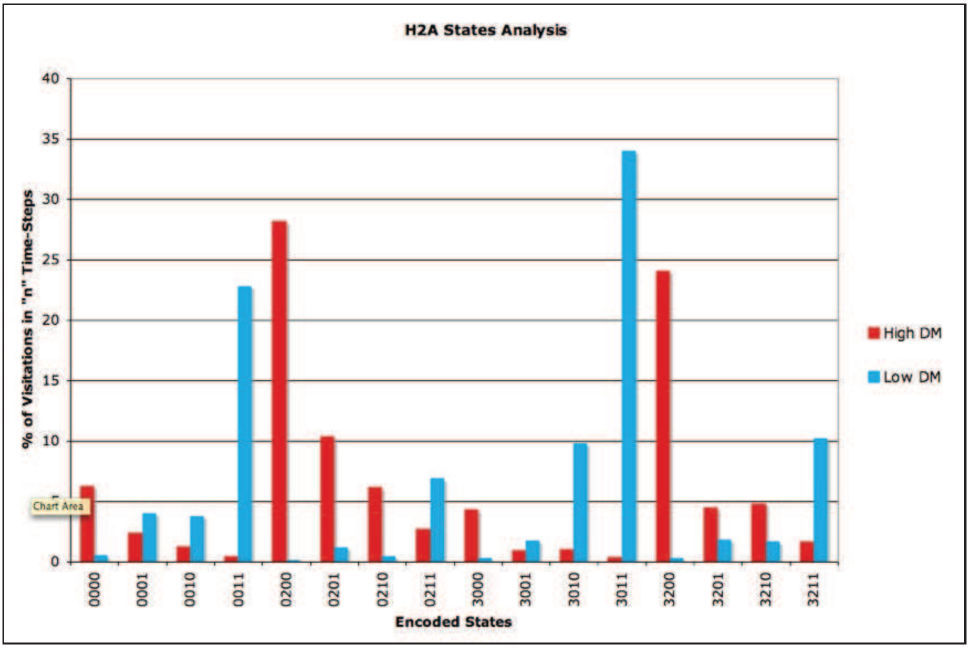


Fig. 5. A Comparison between the average (over 3 Simulation runs) preferences of H2A states for high (red) and low(blue) DNA Methylation Levels.

in H2A during high and low methylation conditions averaged over 3 simulation runs in all nucleosomes. During high methylation condition (DM level > 85%), selective states such as the 5^{th} and 13^{th} were most preferred i.e Arginine was methylated in H2A most frequently. Evidence, (Eckert et al., 2008) indicates that specific cell types, do not contain this modification and hence develop into tumorous cells, (this is an explicit evidence of down regulation of methylation modification leading to tumor growth). Under lower DM conditions (< 15%), the 4^{th} and 12^{th} states were most visited implying high priority to Lysine 5 and 9 modifications. Acetylation of Lysine 5 or (K5) is notably found more during gene expression while that of K9, is an unexplored modification, (Cuddapah et al., 2009; Wyrick & Parra, 2008). This hitherto unreported acetylation in H2A, could be a potential modification that supports gene expression. Figure 6 shows the preferences of H4 states for high and low DNA Methylation levels. Under low DM levels (initially set by the user), acetylated amino acids states, such as the 11^{th}, 35^{th} and 47^{th} predominated i.e. states containing acetylated amino acids such as K5, K8 and K12 (see Table 2) were highly visited. Even when the probability assigned to the three preferred states was lowered for a test set, the system preferred the other two states

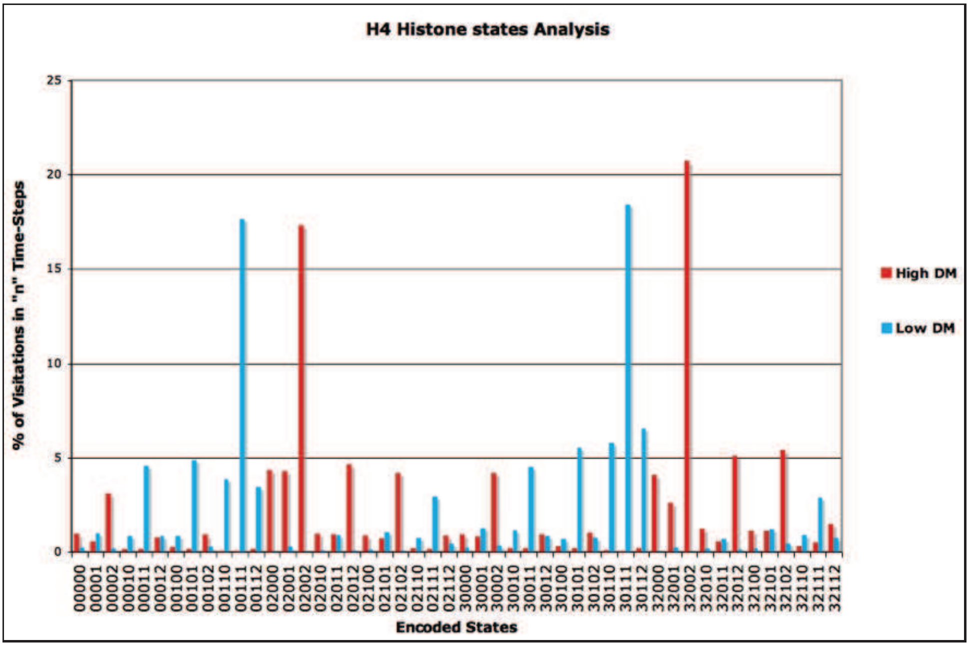


Fig. 6. A Comparison between the average (over 3 Simulation runs) preferences of H4 states for high (red) and low (blue) DNA Methylation Levels.

containing lysine acetylation. Such consistent results demonstrate the ability of our model to reproduce the presence of the modifications mentioned above, during transcription, (as reported, (Taplick, 1998; Zhang et al., 2007) in particular, during expression of oncogenes). For higher levels of DNA methylation (>0.85, Figure 6), the preference is more towards choosing methylated histone states leading to reduced transcription rate. During this high methylation condition, states such as the 15^{th}, 39^{th} and 45^{th} i.e. methylation of K12 was predominantly high. Such strong evidence, (removal of acetylation and adding methylation to amino acids) of modification to a crucial lysine position in H4, is a potential indicator of transcription repression and initiation of DNA methylation. Similiar to the observation in H2B (as recorded in literature, (Zhang et al., 2003)), there is appearance of serine phosphorylation (states 39 and 35 in Figure 6) during both conditions of DM values, which show the importance of this specific modification during expression or otherwise. This suggests that the modification could be present from the time that the H4 histone complex was formed, (Barber et al., 2004) and aid in structural condensation.

Hence a stochastic model of this type can successfully simulate simple concepts to show the possible molecular modifications that appear during different genetic events. The DM fluctuation over specific time-intervals is associated with specific CG dinucleotides in the sequences. In this example, effect of DM and its influence on histone modifications have been effectively illustrated. Futhermore, the same model can be used to study other CG distributions such as 7bp spacing in CpG islands, which can be validated against information on disease associated genes.

5. Conclusion and future directions

In this chapter, the background to epigenetics, their association with diseases and the developments of computational methods and modelling approaches to understand the complexity in this field have been discussed. Significance of growth of experimental data in recent years, which enables detection of DNA methylation influence in disease onset has also been considered. Early attempts at computational methods and models dealing with (i) association of DNA sequences and DM, and (ii) Interdependencies between DM and HM have been explained in detail. Further, we propose approaches to analyse the two elements such as DNA sequence patterns and HM evolution and their influence over DNA methylation mechanisms. Finally, evaluation of success achieved through such computational attempts is illustrated briefly in our results section.

The application of Fourier techniques helped to understand how the sequence patterns appear within the genome and also postulate their control over DM. The results consist of a range of distributions, which are analysed in relation to possible biological significance. The broad spectrum thus obtained, can be attributed to the self-adapting and dynamic nature of the human genome exhibited through events such as self mutations (mC to T, (Doerfler & Böhm, 2006)) or reassignment of DNA methylation patterns across different cells. This ability of cells to dynamically adapt to environmental stimulus by introducing molecular modifications or positive mutations, (which changes nucleotide distributions), is also referred to as "Phenotypic Plasticity". Based on such analyses of the human DNA sequences, further investigations of dynamic histone protein modifications were predicted using novel stochastic modelling techniques.

The EpiGMP model, based on this stochastic approach, has reported histone modifications that were previously recorded and also unexplored modifications and compared them with data recorded through laboratory experiments. For example, the effect of H2A modifications such as Arginine methylation, are not as explicit and strong as H4 but their scattered presence in specific cells/cancer conditions indicates their contribution in the big picture. Hence, based on comparison with experimental and the model results, we conclude that histone modifications while not always consistent do have a role in controlling gene expression and chromosome condensation in human genome.

DNA methylation controls the direction of histone evolution, i.e. the states visited for high levels of DM are not visited for low levels and vice versa. This robust result, obtained for three simulation trials, is a good indication of the reliability of EpiGMP model. This consistency has helped to cluster and predict characteristic histone modifications under defined DNA methylation levels, thus efficiently emulating the real system to an accurate level. The idea behind designing a comprehensive model to mimic epigenetic mechanisms is to address and utilize all of the distributed data available in literature. A generic model, which can simulate conditions of any epigenetically associated disease and report results, is the ideal target. As mentioned in the background section, basic quantitative analyses have reinforced the presence of *apriori* patterns and hence this has given rise to a vital need to design a predictive model with a common framework that can be tested for most conditions. The main advantages of our approach lie in modelling (for all histone types simultaneously) cumulative information such as increased acetylation modifications which occur during gene expression and more methylation during suppression. A further advantage is the expandable layout, which can be developed to accommodate more data in future (incorporating more modifications and multiple sequence patterns).

5.1 Parallelization of EpiGMP model

Parallel computing is an approach, which carries out calculations simultaneously or in a parallel manner using many computational resources at the same time. It is extensively used when there is a high complexity of computation or the data are very large. In our case, the current model definitely requires parallelization, because the random algorithm has to compute outputs from a large sample space, for long iterations or time-steps and most importantly to study several molecular events at genome level. Simulation of the model when applied to objects of size of a chromosome (for more than 1 million time steps) would require heavy computational resources. As a consequence, a parallel and serial version of the model have been developed simultaneously, which is discussed in detail, (Raghavan & Ruskin., 2011; Raghavan et al., 2010).

The field of epigenetics is growing rapidly with important findings being reported on a regular basis. The complex epigenetic layer in humans also houses secondary events through which control is exercised within the cell. For example, chromatin dynamics, which rely on molecular interactions (DNA molecules and proteins such as polycomb), play a major role in long term silencing of genes. Our current work involves, applying this stochastic framework to real gene networks extracted from epigenetic databases such as StatEpigen, (`http://statepigen.sci-sym.dcu.ie/`) in order to predict cancer from simple molecular interactions. To improve realism further, future models must account for secondary effects such as chromatin remodeling, and also role of external proteins such as methyl binding proteins, transcription binding proteins, polycomb amongst others, (Allis et al., 2007) for cellular events. The final goal is to build integrated/hybrid models, combining agent-based and network approaches across several scales, which can be applied to precisely predict epigenetic events based on multiple factors. This “bottom-up” approach facilitates low-level information processing between different molecules so as to understand how the phenotype or physical appearance of an organism evolves at higher level especially under abnormal conditions.

The Fourier analysis on DNA sequences was performed using Matlab software and the source code is available on request. The serial version of EpiGMP model has been developed mainly using C++ language, while routines from OpenMP and MPI libraries were included for the parallel version.

6. Acknowledgements

We gratefully acknowledge financial support from Science Foundation Ireland, project 07/RFP/CMSF724, in the early stages of this work and, subsequently, Complexity-Net /IRCSET pilot award. We thank ICHEC, (Irish High End Computing Centre) for providing access to major computational facilities, required for background work.

7. References

A'Hearn, M. F., Ahern, F. J. & Zipoy, D. M. (1974). Polarization Fourier spectrometer for astronomy, *Applied Optics* 13(5): 1147–1157.

Allis, C. D., Jenuwein, T., Reinberg, D. & Caparros, M. L. (2007). *Epigenetics*, Cold Spring Harbor Press.

Barber, C. M., Turner, F. B., Wang, Y., Hagstrom, K., Taverna, S. D., Mollah, S., Ueberheide, B., Meyer, B. J., Hunt, D. F., Cheung, P. & Allis, C. D. (2004). The enhancement of histone H4 and H2A serine 1 phosphorylation during mitosis and s-phase is evolutionarily conserved, *Chromosoma* 112(7): 360–371.

Baylin, S. B. & Ohm, J. E. (2006). Epigenetic gene silencing in cancer – a mechanism for early oncogenic pathway addiction, *Nature Review Cancer* 6(2): 107–116.
URL: *http://dx.doi.org/10.1038/nrc1799*

Bock, C., Walter, J., Paulsen, M. & Lengauer, T. (2007). CpG island mapping by epigenome prediction, *PLoS Computational Biology* 3(6): e110.

Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros., L. A., Lee, T. I., Levine, S. S., Wernig, M., Tajonar, A., Ray, M. K., Bell, G. W., Otte, A. P., Miguel Vidal, a. D. K. G., Young, R. A. & Jaenisch, R. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells, *Nature* 441(7091): 349–353.

Cedar, H. & Bergman, Y. (2009). Linking DNA methylation and histone modification: Patterns and paradigms, *Nature Review Genetics* 10(5): 295–304.

Chahwan, R., Wontakal, S. N. & Roa, S. (2011). The multidimensional nature of epigenetic information and its role in disease, *Discovery Medicine* 11(58): 233–243.

Chamberlain, S. J. & Lalandea, M. (2010). Neurodevelopmental disorders involving genomic imprinting at human chromosome 15q11–q13, *Neurobiology of Disease* 39(1): 13–20.

Chi, P., Allis, C. D. & Wang, G. G. (2010). Covalent histone modifications – miswritten, misinterpreted and mis-erased in human cancers, *Nature Reviews Cancer* 10(7): 457–469.
URL: *http://dx.doi.org/10.1038/nrc2876*

Clay, O., Schaffner, W. & Matsuo, K. (1995). Periodicity of eight nucleotides in purine distribution around human genomic CpG dinucleotides, *Somatic Cell and Molecular Genetics* 21(2): 91–98.

Collas, P. (2010). The current state of chromatin immunoprecipitation, *Molecular Biotechnology* 45(1): 87–100.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. & Walters, L. (1998). New Goals for the U.S. Human Genome Project: 1998–2003, *Science* 282(5389): 682–689.

Conlon, T., Ruskin, H. J. & Crane, M. (2009). Seizure characterization using frequency-dependent multivariate dynamics, *Computers in Biology and Medicine* 39(9): 760–767.

Cowan, R. (1991). Expected Frequencies of DNA Patterns using Whittle's Formula, *Journal of Applied Probability* 28(4): 886–892.

Cowley, D. E. & Atchley, W. R. (1992). Quantitative genetic models for development, epigenetic selection, and phenotypic evolution, *Evolution* 46(2): 495–518.

Cuddapah, S., Jothi, R., Schones, D. E., Roh, T., Cui, K. & Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains, *Genome Research* 19(1): 24–32.

Das, R., Dimitrova, N., Xuan, Z., Rollins, R. A., Haghighi, F., Edwards, J. R., Ju, J., Bestor, T. H. & Zhang, M. Q. (2006). Computational prediction of methylation status in human genomic sequences, *Proceedings of the National Academy of Sciences* 103(28): 10713–10716.

Doerfler, W. & Böhm, P. (2006). *DNA Methylation: Basics Mechanisms*, first edn, Springer.

Doerfler, W., Toth, M., Kochaneka, S., Achtena, S., Freisem-Rabiena, U., Behn-Krappaa, A. & Orenda, G. (1990). Eukaryotic DNA methylation – facts and problems, *Febs Letter* 286(2): 329–333.

Eckert, D., Biermann, K., Nettersheim, D., Gillis, A., Steger, K., Jack, H., Muller, A., Looijenga, L. & Schorle, H. (2008). Expression of BLIMP1/PRMT5 and concurrent histone H2A/H4 arginine 3 dimethylation in fetal germ cells, CIS/IGCNU and germ cell tumors, *BMC Developmental Biology* 8: 106.

Ehrlich., M., Sanchez, C., Shao, C., Nishiyama, R., Kehrl, J., Kuick, R., Kubota, T. & Hanash, S. (2008). Icf, an immunodeficiency syndrome: DNA methyltransferase 3b involvement, chromosome anomalies, and gene dysregulation, *Autoimmunity* 41(4): 253–271.

Epps, J. (2009). A hybrid technique for the periodicity characterization of genomic sequence data, *EURASIP Journal on Bioinformatics and Systems Biology* 2009.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps, *Nature Reviews Genetics* 8(4): 286–298.

França, L., Carrilho, E. & Kist., T. B. (2002). A review of DNA sequencing techniques., *Quarterly Reviews of Biophysics* 35(2): 169–200.

Fullgrabe, J., Kavanagh, E. & Joseph, B. (2011). Histone Onco – Modifications, *Oncogene* 30(31): 3391–3403.
URL: *http://dx.doi.org/10.1038/onc.2011.121*

Gao, F. & Zhang, C.-T. (2006). GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences, *Nucleic Acids Research* 34(2): 686–691.

Gertz, J., Varley, K. E., Reddy, T. E., Bowling, K. M. & Pauli, F. (2011). Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation, *PLoS Genetics* 7(8): e1002228.

Glass, J. L., Fazzari, M. L., Ferguson-Smith, A. C. & Greally, J. M. (2004). CG di-nucleotide periodicities recognized by the dnmt-3a-dnmt-3l complex are distinctive at retro-elements and imprinted domains, *Mammalian Genome* 20(9-10): 633–643.

Glass, J. L., Thompson, R. F., Khulan, B., Figueroa, M. E., Olivier, E. N., Oakley, E. J., Zant, G. V., Bouhassira, E. E., Melnick, A., Golden, A., Fazzari, M. J. & Greally, J. M. (2007). CG dinucleotide clustering is a species-specific property of the genome, *Nucleic Acid Research* 35(20): 6798–6807.

Goodman, J. W. (2005). *Introduction to Fourier Optics*, third edn, Roberts and Company.

Herzel, H., Weiss, O. & Trifonov, E. N. (1999). 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding., *Bioinformatics* 15(3): 187–193.

Hosid, S., Trifonov, E. N. & Bolshoy, A. (2004). Sequence periodicity of Escherichia coli is concentrated in intergenic regions, *BMC Molecular Biology* 5(1): 14.
URL: *http://www.biomedcentral.com/1471-2199/5/14*

Ito, T. (2007). Role of histone modification in chromatin dynamics, *Journal of Biochemistry* 141(5): 609–614.

Jenuwein, T. & Allis, C. D. (2001). Translating the histone code, *Science* 293(5532): 1074–1080.

Jung, I. & Kim, D. (2009). Regulatory patterns of histone modifications to control the DNA methylation status at CpG islands, *IBC* 1(4): 1–7.

Kaiser, G. (1994). *A Friendly Guide to Wavelets*, sixth edn, Birkhäuser.

Karlić, R., Chung, H., Lasserre, J., Vlahoviček, K. & Vingron, M. (2010). Histone modification levels are predictive for gene expression, *PNAS* 107(7): 2926–2931.

Kouzarides, T. (2007). Chromatin modifications and their function, *Cell* 128(4): 693–705.

Kwon, D., Vannucci, M., Song, J. J., Jeong, J. & Pfeiffer, R. M. (2008). A novel wavelet-based thresholding method for the pre-processing of mass spectrometry data that accounts for heterogeneous noise, *Proteomics* 8(15): 3019–3029.

Li., R., Zhu, H. & Ruan, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing, *Nucleic Acid Research* 20(2): 265–272.

Matsumoto, M. & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Transactions on Modeling and Computer Simulation* 8(1): 3–30.

Meng, C. F., Zhu, X. J., Peng, G. & Dai, D. (2009). Promoter histone H3 lysine 9 di-methylation is associated with DNA methylation and aberrant expression of p16 in gastric cancer cells, *Oncology Report* 22(5): 1221–1227.

Morrison, N. (1994). *Introduction to Fourier Analysis*, Wiley-Interscience.

Murrell, A., Rakyan, V. K. & Beck, S. (2005). From genome to epigenome, *Human Molecular Genetics* 14(1): 3–10.

Raghavan, K. & Ruskin., H. J. (2011). Computational epigenetic micromodel - framework for parallel implementation and information flow., *Proceedings of the Eighth International Conference on Complex Systems*, Vol. 8, NECSI Knowledge Press, pp. 340–353.

Raghavan, K., Ruskin, H. J. & Perrin, D. (2011). Computational analysis of epigenetic information in human DNA sequences, *Proceedings of the International Conference on Bioscience, Biochemistry and Bioinformatics 2011*, Vol. 5, International Proceedings of Chemical, Biological and Environmental Engineering, pp. 383–387.

Raghavan, K., Ruskin, H. J., Perrin, D., Burns, J. & Goasmat, F. (2010). Computational micromodel for epigenetic mechanisms, *PLoS One* 5(11): e14031.

Riggs, A. D. & Xiong, Z. (2004). Methylation and epigenetic fidelity, *PNAS* 101(1): 4–5.

Salz, J. & Weinstein, S. B. (1969). Fourier transform communication system, *Proceedings of the first ACM symposium on Problems in the optimization of data communications systems*, ACM, pp. 99–128.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G. & Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome, *Cell* 132(5): 887–898.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S0092867408002705*

Segal, E. & Widom, J. (2009). What controls nucleosome positions, *Trends in Genetics* 25(8): 335–343.

Silverman, B. & Linskera, R. (1986). A measure of DNA periodicity., *Journal of Theoretical Biology* 118(7): 295–300.

Strachan, T. & Read, A. P. (1999). *Human Molecular Genetics*, 2 edn, New York: Wiley-Liss.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M. P., Walker, J. R. & Hogenesch, J. B. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes, *Proceedings of* *the National Academy of Sciences of the United States of America* 101(16): 6062–6067. URL: *http://www.pnas.org/content/101/16/6062.abstract*

Su, J., Zhang, Y., Lv, J., Liu, H., Tang, X., Wang, F., Qi, Y., Feng, Y. & Li, X. (2010). Cpg_mi: a novel approach for identifying functional CpG islands in mammalian genomes, *Nucleic Acids Research* 38(1): e6.

Sun, J. M., Chen, H. Y., Espino, P. S. & Davie, J. R. (2007). Phosphorylated serine 28 of histone h3 is associated with destabilized nucleosomes in transcribed chromatin, *Nucleic Acids Research* 35(19): 6640–6647.

Sundararajan, N., Mao, D., Chan, S., Koo, T.-W., Su, X., Sun, L., Zhang, J., Sung, K.-b., Yamakawa, M., Gafken, P. R., Randolph, T., McLerran, D., Feng, Z., Berlin, A. A. & Roth, M. B. (2006). Ultrasensitive detection and characterization of posttranslational modifications using surface-enhanced raman spectroscopy, *Analytical Chemistry* 78(11): 3543–3550.

Takai, D. & Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22, *PNAS* 99(6): 3740–3745.

Taplick, J. (1998). Histone H4 acetylation during interleukin-2 stimulation of mouse t cells, *FEBS Letters* 436(3): 349–352.

Tsonis, A. A., Kumar, P., Elsner, J. B. & Tsonis, P. A. (1996). Wavelet analysis of DNA sequences, *Physical Review E* 53(2): 1828–1834.

Turner, B. M. (2001). *Chromatin and Gene Regulation – Mechanisms in Epigenetics*, 2nd edition edn, BlackWell Science Ltd.

Ushijima, T., Watanabe, N., Okochi, E., Kaneda, A., Sugimura, T. & Miyamoto, K. (2003). Fidelity of the methylation pattern, its variation in the genome, *Genome Research* 13(5): 868–874.

Waddington, C. H. (1942). The epigenotype, *Endeavour* 1: 18–20.

Whittle, P. (1955). Some distribution and moment formulae for the markov chain, *Journal of the Royal Statistical Society Series B (Methodological)* 17(2): 235–242.

Wyrick, J. J. & Parra, M. A. (2008). The role of histone H2A and H2B post-translational modifications in transcription: A genomic perspective, *Biochimica et Biophysica Acta* 1789(1): 37–44.

Yasuhara, J. C., DeCrease, C. H. & Wakimoto, B. T. (2005). Evolution of heterochromatic genes of drosophila, *PNAS* 102(31): 10958–10963.

Yin, H. & Lin, H. (2007). An epigenetic activation role of piwi and a piwi associated piRNA in drosophila melanogaster, *Nature* 450(7167): 304–308.

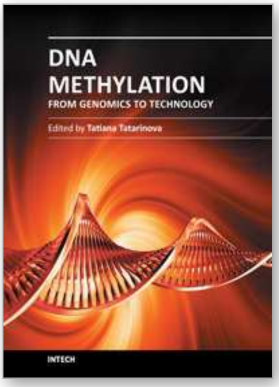
Yu, H., Zhu, S., Zhou, B., Xue, H. & Han, J. (2008). Inferring causal relationships among different histone modifications and gene expression, *Genome Research* 18(8): 1314–1324.

Zhang, L., Eugeni, E. E., Parthun, M. R. & Freitas, M. A. (2003). Identification of novel histone post-translational modifications by peptide mass fingerprinting, *Chromosoma* 112(2): 77–86.

Zhang, L., Su, X., Liu, S., Knapp, A. R., Parthun, M. R., Marcucci, G. & Freitas, M. A. (2007). Histone H4 n-terminal acetylation in kasumi-1 cells treated with depsipeptide determined by acetic acid. urea polyacrylamide gel electrophoresis, amino acid coded mass tagging, and mass spectrometry, *Journal of Proteome Research* 6(q): 81–88.

Zhao, J., Yang, X. W., Li, J. P. & Tang, Y. Y. (2001). DNA sequences classification based on Wavelet package analyses, *WAA '01: Proceedings of the Second International Conference on Wavelet Analysis and Its Applications*, Springer-Verlag.

Zheng, C. & Hayes, J. J. (2003). Structure and interactions of core histone tail domains, *Bioplymers* 68(4): 539–546.



DNA Methylation - From Genomics to Technology
Edited by Dr. Tatiana Tatarinova

ISBN 978-953-51-0320-2
Hard cover, 400 pages
Publisher InTech
Published online 16, March, 2012
Published in print edition March, 2012

Epigenetics is one of the most exciting and rapidly developing areas of modern genetics with applications in many disciplines from medicine to agriculture. The most common form of epigenetic modification is DNA methylation, which plays a key role in fundamental developmental processes such as embryogenesis and also in the response of organisms to a wide range of environmental stimuli. Indeed, epigenetics is increasing regarded as one of the major mechanisms used by animals and plants to modulate their genome and its expression to adapt to a wide range of environmental factors. This book brings together a group of experts at the cutting edge of research into DNA methylation and highlights recent advances in methodology and knowledge of underlying mechanisms of this most important of genetic processes. The reader will gain an understanding of the impact, significance and recent advances within the field of epigenetics with a focus on DNA methylation.